Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 201646, 11 pages
http://dx.doi.org/10.1155/2015/201646

Research Article

Modeling of Energy Demand of a High-Tech Greenhouse in Warm Climate Based on Bayesian Networks

César Hernández, José del Sagrado, Francisco Rodríguez, José Carlos Moreno, and Jorge Antonio Sánchez

University of Almería, Agrifood Campus of International Excellence (CeiA3), CIESOL Research Center on Solar Energy, Informatics Department, Carretera Sacramento s/n, 04120 Almería, Spain

Correspondence should be addressed to César Hernández; chdezh@ual.es

Received 19 January 2015; Accepted 25 March 2015

Academic Editor: Hang Xu

This work analyzes energy demand in a High-Tech greenhouse and its characterization, with the objective of building and evaluating classification models based on Bayesian networks. The utility of these models resides in their capacity of perceiving relations among variables in the greenhouse by identifying probabilistic dependences between them and their ability to make predictions without the need of observing all the variables present in the model. In this way they provide a useful tool for an energetic control system design. In this paper the acquisition data system used in order to collect the dataset studied is described. The energy demand distribution is analyzed and different discretization techniques are applied to reduce its dimensionality, paying particular attention to their impact on the classification model's performance. A comparison between the different classification models applied is performed.

1. Introduction

Greenhouses are production systems characterized by the intensive and efficient use of primary resources [1]. This involves the implementation of specific cultural techniques in climate-controlled environments created in light structures with transparent covers, the majority of which are plastic films. Despite the mechanical factors that are characteristic of buildings exposed to the weather, the design of these structures is based on the need for an optimum energy balance that provides the internal climatic conditions required for crop development [2]. This crop growth is primarily determined by climate and the amount of water and fertilizers applied through irrigation. Therefore, greenhouses are ideal for farming because they allow one to optimize these physical parameters via the photosynthesis process in order to enhance biomass production. This optimization requires energy consumption, depending on the crop's physiological requirements and, additionally, depending on the production patterns adopted for yield quantity and timing [3].

The warm climate, which exists in various locations around the world, is characterized by low annual precipitation and consequently high solar radiation availability [4]. This solar radiation availability for photosynthesis, together with warm summers and mild temperatures in winter, means greenhouses in these regions perform well, in general terms, provided that crop water requirements are met by local underground resources and by using optimized irrigation systems. In these zones, low thermal loads are required during the winter to reach optimal temperatures inside the greenhouse; this is thanks to the confinement effect, resulting from the decrease in air exchange with the outside environment, and to the low transparency to far infrared radiation (emitted by the crop, the soil, and the inner greenhouse elements) and high transparency to sunlight [5]. During the summer, especially in the Mediterranean and tropical areas where the inside temperature can exceed the recommended maximum threshold levels, indoor conditions can be controlled by passive means such as the use of natural ventilation [6]. Nonetheless, these averaged climatic conditions and strategies are the result of a dynamic interaction between the greenhouse structures and the varying daily values of outdoor temperature, solar irradiance, relative humidity, and wind velocity. In spite of the overall good performance shown above, it is possible that in a relevant number of hours over the year, active measures are required to maintain crop growth and increase greenhouse productivity [7–9]. In addition to this, the passive and eventually active climate control strategies must address both demands: heating and cooling, introducing an added complexity factor to the system's selection and operation [10]. In addition, a relevant quantity of energy in the form of electricity must also be considered for climate control in greenhouses [11]. Besides the eventual use of electrically driven heat pumps or chillers, the elements required for heat distribution inside the greenhouse are water and air pumps, as well as fans and motors for roof-windows. Also, irrigation pumps are fed by electricity and, as in the case of heat pumps, the corresponding energy load presents a climate dependent pattern because of the relationship between the evapotranspiration process and the inside and outside variables [4].

In this line, the energy demands can be identified as follows: related to indoor climate control (ventilation, heating system, humidification, dehumidification, . . .), related to irrigation, related to fertilization, related to CO_2 enrichment, and associated with the use of computers and devices for measuring and control, for lighting, and for the systems supporting harvest and postharvest tasks. The first two demands must be considered basic given that they are related to the physiological needs of the crops. Both demands are determined by an elementary process of mass and heat transfer involving the plants, the air inside, the cladding surfaces, and the meteorological variables. As a result of the corresponding balances, a certain amount of energy must be added or extracted to fulfil the established temperature or humidity set points [1]. Consequently, the electricity demand is directly related to climate [12]. A precise knowledge of this amount, either on an instantaneous or on an accumulated basis, is key to greenhouse energy demand prediction [2]. Accordingly, in order to formally characterize the energy demand of all the available elements within a High-Tech greenhouse, a model based on Bayesian networks is proposed.

Bayesian networks (BNs) are widely accepted as a reasoning technique for making predictions in many different research fields [13–15]. BNs have many applications; for example, they have been used in context of estimation [16, 17], control systems [13], and prediction [18]. In the context of control systems applied in greenhouses they have been proven to be useful for the temperature control problem [19]. This paper proposes an energy demand prediction method based on a BN classifier. The greenhouse energy demand is considered as the target class and the BN classifier is learnt from an historical dataset of observations of the identified contributing factors together with its associated energy demand [13]. Prediction can be performed by applying Bayes rule to compute the probability of target class given the particular instance of contributing factors, and the class label with highest posterior probability is the predictive result [18]. Besides, the learned BN classifier models show complex relationships among contributing factors.

Table 1: Sensory system installed in the greenhouse.

Variable	Model
	Inside of the greenhouse
Temperature and humidity	HMP45a, Vaisala, Finland
Global radiation	MRG-1P, ITC, Almeria, Spain
CO_2 concentration	UA-06, PRIVA B.V., De Lier, Holland
	Outside of the greenhouse
Temperature and humidity	HMP45a, Vaisala, Finland
Radiation PAR	PAR Lite
Global radiation	MRG-1P, ITC, Almeria, Spain
Wind speed	Model 12102, R. M. Young Company, Traverse City, Michigan, USA
Wind direction	D-034B-CA, Delta-T Devices Ltd., Cambridge, United Kingdom
CO_2 concentration	ZFP–DZ, Siemens, Munich, Germany
Rain	WS 10 R, JUNG Electro Iberica S.A., Barcelona, Spain
Energy demand	SINEAX M 561 with 1 and 2, respectively, 3 analog outputs

This paper presents promising ideas and results about the combination of Bayesian networks and energy demands prediction applied in greenhouses. The main objective is to obtain and validate a model, based on Bayesian networks, of energy demand in the greenhouse. The paper is organized as follows: in Section 2 the subject of research from this investigation and the method to be used for the study are presented. Then, the analysis of electric power and the obtained Bayesian network system are discussed in Section 3 including simulation results. Finally, conclusions and future work are given in Section 4.

2. Materials and Methods

2.1. Greenhouse Environment and Acquisition Data System. The system subject of research in this paper is a greenhouse multispan "Parral type," with a surface area of 877 m^2, oriented to the N-S direction. The facilities are placed on the experimental station of Cajamar Foundation, "*Las Palmerillas*," at the municipal term of "El Ejido," in Almeria, at the S-E of Spain [20]. The greenhouse has lateral and cenital ventilations powered through motors AC independent, aerothermo, heating system by hot water pipes fed with biomass, an enriched system of CO_2 created by the biomass burning, shade nets, and systems of feeding for water and nutrients. It is equipped with a measure equipment of 52 variables and is designed to develop identification tests and to implement strategies of climate control, fertigation, and electric power (Figure 1). In Table 1 the variables measured and the model of the sensors used are detailed [21].

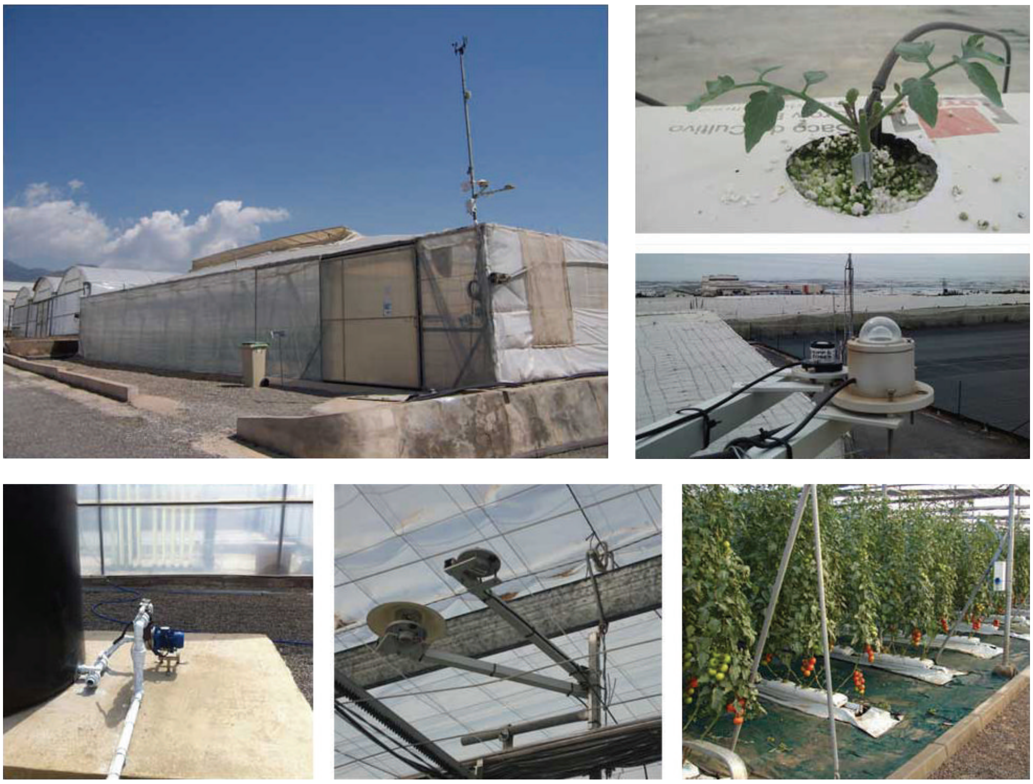


Figure 1: Greenhouse and acquisition data system.

2.2. Bayesian Networks for Energy Demand Characterization. The advantages of using BNs [22] mainly are their ability to reason under uncertainty, along with the combination of a graphical representation based on sound mathematics (i.e., probability theory). The graphical representation of a BN provides a representation of the relations between variables in the model, which is accessible and can be easily interpreted. The strength of these relations is measured using conditional probability distributions. Thus, a BN provides a white box model with the ability of propagating (applying probabilistic reasoning) the impact that the available information (i.e., evidence) has on the other model variables modifying their states. Formally, a BN can be defined as follows.

Definition 1. A Bayesian network $\mathbf{B} = (G, \mathbf{P})$ for a set of variables $\mathbf{V} = X_1, X_2, \ldots, X_n$ consists of the following:

(i) a direct acyclic graph $G = (\mathbf{V}, \mathbf{L})$, where each node represents one of the variables in $\mathbf{V}$ and a set $\mathbf{L}$ of directed links among them;

(ii) a set $\mathbf{P}$ of conditional probability distributions, $P(X_i \mid pa(X_i))$, of each variable X_i given their parents set $pa(X_i)$ in the directed acyclic graph G.

From the set $\mathbf{P}$ of conditional probability distributions, the joint probability distribution can be obtained by applying the chain's rule:

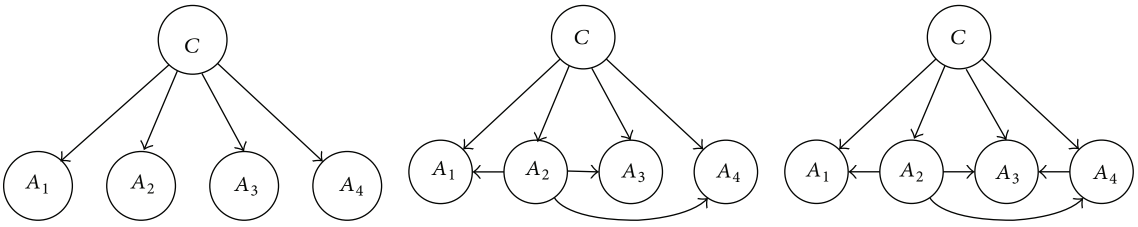
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa(X_i)). \qquad (1)$$

Classifiers Based on Bayesian Networks. A Bayesian classifier can be considered as a particular case of a Bayesian network, where there is a variable that accomplishes the class role and the other variables are considered attributes. The network's structure depends basically on the kind of classifier [22]. Once the structure is learned, the conditional probability distributions $P(X_i \mid pa(X_i))$ are estimated from data applying Laplace's rule of succession as

$$P(X_i \mid pa(X_i)) = \frac{n(x_i, pa(x_i)) + 1}{n(pa(x_i)) + |X_i|}, \qquad (2)$$

where $n()$ is the number of rows in the data containing a given observation x_i for X_i and $pa(x_i)$ for its parents and $|X_i|$ is the number of possible values of X_i.

The classification process consists in identifying to which category c_i of a set of categories $\mathbf{C} = \{c_1, c_2, \ldots, c_m\}$ a new object $\mathbf{o} = (a_1, a_2, \ldots, a_n)$, characterized by individual observations of a set of features (or attributes) $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$, belongs. In the probabilistic case, this task can be performed in an accurate way by applying Bayes' theorem. However, working with a joint probability distribution is often unmanageable and it is appealed to simple models based on factorization of the mentioned distribution. Due to this fact and in order to study greenhouse energy demand, three paradigms based on Bayesian networks will be applied: Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN), and K-Dependence Bayesian (KDB). These models are characterized by the following.



(a) Naïve Bayes (b) Tree Augmented Naïve Bayes (c) K-Dependence Bayesian

Figure 2: Network structures examples with four attributes NB, TAN, and KDB2.

(i) *Naïve Bayes.* On this model, the conditional probability of the class is estimated, assuming that all the attributes are conditionally independent once the value of the class is given (see Figure 2(a)). This assumption translates into the following factorization:

$$\forall c_j \in C, \quad p\left(\mathbf{o} \mid c_j\right) = \prod_{i=1}^{n} p\left(a_i \mid c_j\right). \tag{3}$$

Then, when classifying, the maximum a posteriori (MAP) hypothesis is used:

$$\begin{aligned} c_{\mathrm{MAP}} &= \arg\max_{c_j \in C} p\left(c_j \mid \mathbf{o}\right) \\ &= \arg\max_{c_j \in C} \left(p\left(c_j\right) \prod_{i=1}^{n} p\left(a_i \mid c_i\right) \right). \end{aligned} \tag{4}$$

(ii) *Tree Augmented Naïve Bayes (TAN).* Tree Augmented Naïve Bayes [23] relaxes the conditional independence assumption of NB. It does so through an expansion of NB graphical structure (see Figure 2(b)), in which each attribute has as parents set the class variable and, at most, another attribute variable. This is achieved by learning a maximum weight spanning tree [24] based on conditional mutual information between two attributes given the class value. Thus, the structure returned corresponds to a NB augmented with this tree [23]. In this way TAN achieves a compromise between complexity and precision.

(iii) *K-Dependence Bayesian.* This model is based on the idea of *k-dependence* estimators, introduced in [25]. According to this, the probability of each attribute is conditioned by the class and as most other k attributes (see Figure 2(c)). This structure can be learned performing a heuristic search applying $K2$ algorithm [26]. Modifying the value of k (i.e., the maximum number of parents that an attribute can have) more dense graphical structures can be obtained. The advantage of KDB with respect to TAN is its flexibility. As TAN sets to one the maximum number of attributes that a variable can have as a parent, in addition to the class, the dependences that can be modelled between each attributes group are restricted.

Table 2: Maximum energy demand from the main systems.

Systems	Energy demand (KW)
Motors AC (ventilation)	0.38
Pump for heating	1.86
Blower (enrich system of CO_2)	1.8

3. Results and Discussion

3.1. Analysis of the Distribution of Energy Demand in the Greenhouse. This analysis aims to study the distribution of energy demand in the greenhouse, collected in the dataset, to apply different discretization techniques and to observe their impact on the characterization of the distribution of energy demand when it is modeled by means of classifiers based in Bayesian networks. To analyze the energy demand in the system, data measured at intervals of 1 minute have been used, taken in the period between the months of October 2013 and May 2014. This set has 330, 232 instances. The maximum power consumption from the main integrated systems is described in Table 2. Table 3 lists the variables used in the initial dataset, units, and associated measurement range. In order to quantitatively describe the set of observations about energy demand in the greenhouse, a five-number summary is used: sample minimum (0 KW), first quartile (0.056 KW), median (0.066 KW), third quartile (0.111 KW), and sample maximum (4.3 KW). The mean energy demand is 0.241 KW with a typical deviation of 0.583 KW. In this case, the energy demand distribution is asymmetric and has a positive skewness as the histogram (see Figure 3(a)), computed from the observations in the dataset, shows.

For the system under study the sun is the main energy source. In consequence we wonder if there is a significant energy demand difference between diurnal and nocturnal periods. According to the criteria established by the Andalusian Energy Agency [27], the diurnal period ranges from sunrise (7:00 h) to sunset (20:00 h), whilst the rest of time slots in a 24-hour period (i.e., from 00:00 h to 6:59 h and from 20:01 h to 23:59 h) correspond to the nocturnal period.

The dataset under study contains 192, 826 observations that correspond to the diurnal period. The distribution of energy demand during this period is described by the following five-number summary plus mean and standard deviation: sample minimum (0 KW), first quartile (0.051 KW),

Table 3: Variables included at the dataset.

Variable	Range of variation	Unit
Day	$\{1, 2, \ldots, 31\}$	Day
Month	$\{1, 2, \ldots, 12\}$	Month
Hour	$\{0, 1, \ldots, 23\}$	Hour
Minute	$\{0, 1, \ldots, 59\}$	min
Outside temperature	$[0, 50]$	°C
Outside relative humidity	$[0, 100]$	%
Outside wind speed	$[1, 24.2]$	m/s
Inside temperature	$[0, 60]$	°C
Inside relative humidity	$[0, 100]$	%
Inside global radiation	$[0, 672.4]$	W/m^2
Pump for heating	$\{0, 1\}$	Logic
Cenital ventilation	$[0, 100]$	%
Lateral ventilation	$[0, 100]$	%
Blower	$\{0, 1\}$	Logic
Energy demand	$[0, 4.3]$	KW

median (0.060 KW), third quartile (0.119 KW), sample maximum (4.3 KW), mean (0.221 KW), and standard deviation (0.535 KW). Figure 3(b) shows the histogram of the diurnal distribution of energy demand, which also presents a positive skew.

In the case of nocturnal period the dataset contains 137, 406 observations and the distribution of energy demand can be summarized as follows: sample minimum (0.031 KW), first quartile (0.061 KW), median (0.070 KW), third quartile (0.077 KW), sample maximum (4.2 KW), mean (0.268 KW), and standard deviation (0.643 KW). Its histogram (see Figure 3(c)) also presents a positive skew.

Table 4 provides a comparison of the sets of observations considered in the analysis of the energy demand in the greenhouse. It can be observed that there are some little differences between diurnal and nocturnal periods of energy demand. Therefore, from now on we do not make any distinction between diurnal and nocturnal periods. The basic electronic equipment (see Section 2.1) installed in the greenhouse introduces a latent electric power which transforms into noise in the data when we are trying to characterize the energy demand as a BN classifier. Thus, observations corresponding to an energy demand less than 0.111 KW (that stems from the basic electronic equipment) are removed from the initial dataset avoiding discontinuities inside a day (i.e., inside a given day cannot be discontinuities in the observations once a starting time point is set). Thus, 75, 581 observations of the initial set are kept. The energy demand distribution for observations greater than or equal to this latent energy demand threshold (see Figure 3(d)) has sample minimum (0.111 KW), first quartile (0.169 KW), median (0.269 KW), third quartile (1.45 KW), sample maximum (4.3 KW), mean (0.779 KW), and standard deviation (0.989 KW). The consequences of doing this filtering are as follows: (i) there is a substantial reduction in the number of observations being considered, (ii) the skewness of the distribution is preserved, and (iii) energy demand peaks become visible. All of them can be an aid during energy demand characterization needed to build the BN classifier.

3.2. Discretization of the Distribution of Energy Demand in the Greenhouse.

The discretization process allows replacing the real distribution of data by a blend of uniform distributions. This process also reduces the data dimensionality, transforming the data rank into a subset of discrete values or subranks. Some reasons to discretize are as follows: restrictions on the type of variables (i.e., categories or nominal) that a classification algorithm can handle, the improvement of computation time by reducing the number of partitions to be evaluated, and noise reduction in the data by reducing the set of values.

Usually a distinction is made between supervised discretization methods (i.e., based on entropy [28]) and unsupervised methods (i.e., discretization by equal-width or equal-frequency [29]). This distinction is based on whether they employ or not class information when locating the intervals. Without any doubt when we face a particular problem (to its dataset), selection of the discretization method has a straight influence on further classification tasks (classification rate, area under a receiver operating characteristic curve, etc.).

In the case of the energy demand in the greenhouse, the variable which we want to discretize is also the class variable. For this reason, only not supervised discretization methods are suitable. Basically, these methods try to divide the rank of observed data in intervals (usually, the number of intervals ranges from 5 to 20).

3.2.1. Discretization by Intervals of Equal-Width.

This is the simplest method for data discretization [29], which splits the variable range in b intervals (where b is a parameter provided by the user). If the observed variable X takes values in the range between $x_{\min}$ and $x_{\max}$, the width of the intervals is computed as follows:

$$\delta = \frac{(x_{\max} - x_{\min})}{b}. \tag{5}$$

The intervals high extremes will be $x_{\min} + i \cdot \delta$, where $i = 1, 2, \ldots, b-1$. Usually, the number b of intervals is set between 5 and 10, although its optimal value depends on the size of the dataset, among other factors.

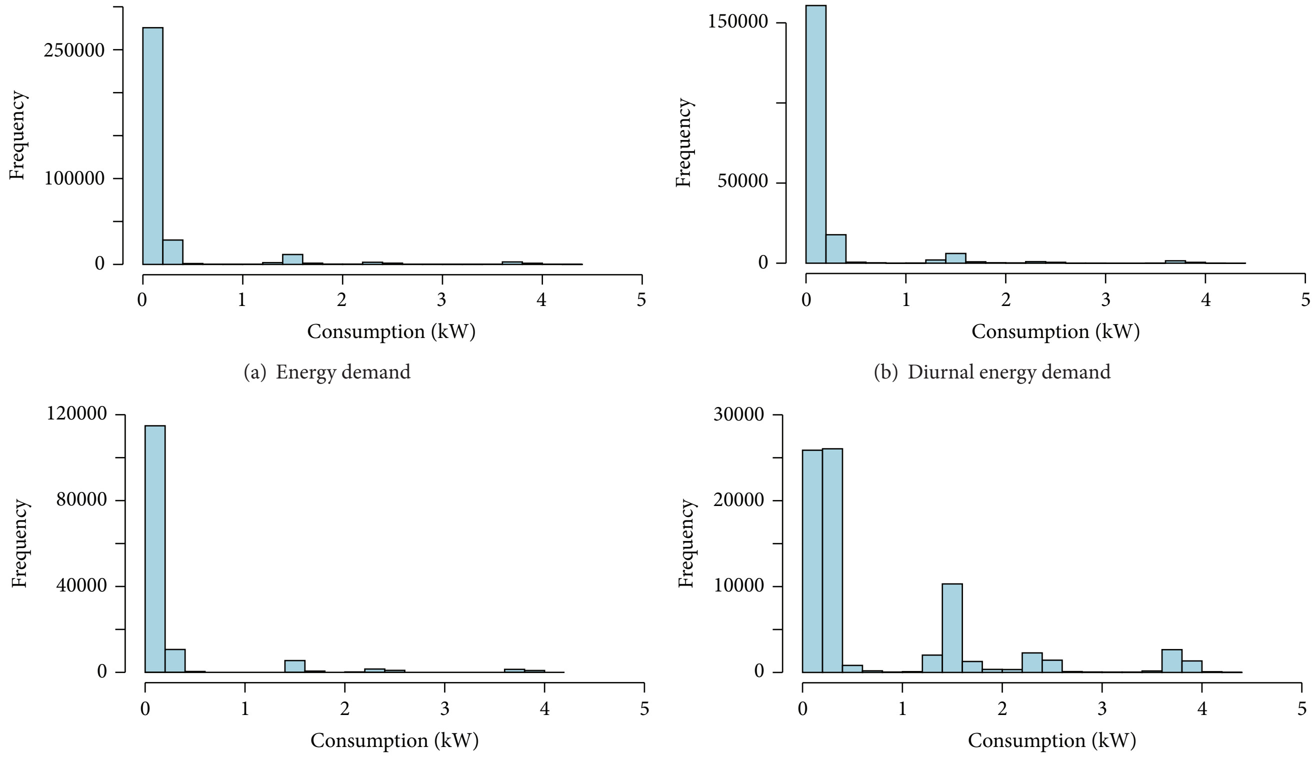
In order to discretize energy demand of the greenhouse by equal-width, the number of intervals b has been set equal to 4 trying to characterize the energy demand peaks shown in Figure 3(d). The intervals obtained in this way are bin1 $= [0.111, 1.15825]$, bin2 $= (1.15825, 2.205]$, bin3 $= (2.205, 3.253]$, and bin4 $= (3.253, 4.3]$. Figure 4(a) shows the histogram for the distribution of energy demand observations after its discretization in 4 equal-width intervals. Note that the “shape” of original distribution is roughly kept.

3.2.2. Discretization by Intervals of Equal-Frequency.

This unsupervised discretization technique [29] first sorts the values of the variable that is going to be discretized and

Table 4: Summary of the distributions of the observations of energy demand in the greenhouse.

Measures	Energy demand	Diurnal energy demand	Nocturnal energy demand
Min	0.000	0.000	0.031
First quartile	0.056	0.051	0.061
Median	0.066	0.060	0.070
Mean	0.241	0.221	0.268
Third quartile	0.111	0.119	0.077
Max	4.3	4.3	4.2
Standard deviation	0.583	0.535	0.643



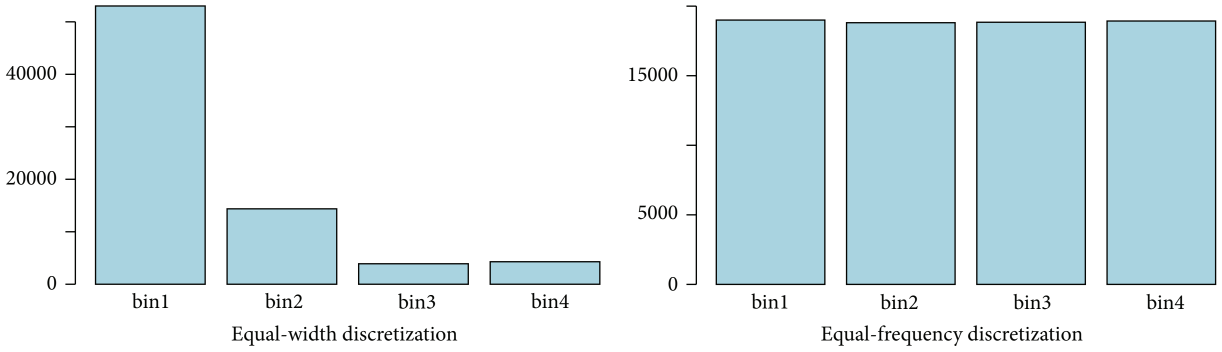
(a) Energy demand

(b) Diurnal energy demand

(c) Nocturnal energy demand

(d) Energy demand greater than 0.111 KW

Figure 3: Distribution of energy demand in the greenhouse during the period from October 2013 to May 2014.



(a) Equal-width

(b) Equal-frequency

Figure 4: Discretization of the energy demand distribution.

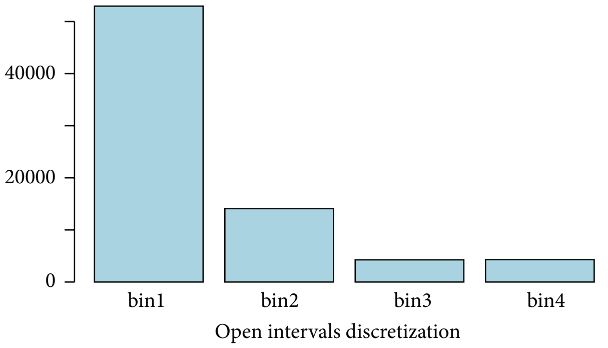


Figure 5: Open intervals discretization of the energy demand distribution.

then splits the values into b intervals, each one containing approximately the same number of observations of adjacent values. As each group of observations on the same value has to be placed into the same interval, it is not always possible that the intervals contain the same number of values. The obtained discretization will be more balanced between the different intervals.

Figure 4(b) shows the histogram obtained after performing the equal-frequency discretization of energy demand into 4 intervals: bin1 $= [0.111, 0.169]$, bin2 $= (0.169, 0.269]$, bin3 $= (0.269, 1.45]$, and bin4 $= (1.45, 4.3]$. In this case the original distribution is transformed into a uniform one.

3.2.3. Discretization by Open Intervals. In situations where same-size intervals are not suitable, different-size intervals or open intervals can be used instead. Intervals are chosen in a way that the central points coincide with real-observed data. Thus, clustering error is mitigated in subsequent mathematical analysis. However, interval boundaries do not have to coincide with observed data.

In order to characterize energy demand in the greenhouse by open intervals, we have chosen as central points the peaks of energy demand that shows the distribution in Figure 3(d) and the interval boundaries as a compromise between separability of observations and energy demand of the active systems in the greenhouse (see Table 2). With these assumptions the resulting intervals are bin1 $= [0.111, 0.9]$, bin2 $= (0.9, 2.01]$, bin3 $= (2.01, 3.25]$, and bin4 $= (3.25, 4.3]$. Figure 5 shows the histogram obtained. After applying discretization the "shape" of original distribution is roughly kept.

3.3. Classification of Energy Demand in the Greenhouse. The objective in this paper is to characterize the greenhouse's energy demand applying BN classifiers. The discretization of energy demand made in Section 3.2 serves as basis for setting the different "tags" that are assigned to the class variable "energy consumption." Therefore, depending on the applied discretization, the intervals identified as class labels are used. Notice that as was previously mentioned discretization has a direct impact on classification results. So the next challenge is to select, from the discretization techniques applied, the one that allows the best classification of energy demand with available dataset. This is achieved in a two-stage process. First, for each discretization of the class and for each BN classifier type (i.e., NB, TAN, and KDB2), a model is learned (see Section 2.2) using stratified tenfold cross validation and the dependence structure learned is validated ad hoc by two experts in the field. Second, the classification accuracy of the models is evaluated using the proportion of correctly classified instances. Table 5 shows the classification rates for the different discretization techniques and BN classifiers learned. KDB2 with open intervals discretization gives the best classification rate (97.77%), which is slightly better than provided by equal-width discretization (97.41%). Results suggest that dependences between attributes are important because the classification rate increases as more dependences are considered in the BN classifier and that if the discretization technique preserves the asymmetry of the original distribution of energy demand (i.e., equal-width and open intervals discretization), then better classification results are obtained. Misclassification of observations is due to the skewness of the original distribution and to the large number of observations for very low energy demand values. Discretization by open intervals corrects partially this problem by increasing the separability of observations with respect to equal-width discretization.

Table 5: Classification rate (%) for different discretization techniques and BN classifiers.

BN classifier	Equal-width	Equal-frequency	Open intervals
NB	88.62	53.88	88.97
TAN	96.81	68.64	97.04
KDB2	97.41	72.02	97.77

Table 6: Performance results (%) of KDB2 classifier.

Energy demand	TP rate	FP rate	Precision
$[0.111, 0.90]$	99.2	2.1	99.1
$(0.900, 2.01]$	95.1	0.9	95.9
$(2.010, 3.25]$	90.3	0.7	89.1
$(3.250, 4.30]$	96.0	0.3	95.8

Figure 6 shows the learned structure of KDB2 Bayesian classifier for initial dataset discretized by open intervals ($b = 4$). It is worth noticing that the temperature inside the greenhouse (*inside temperature*) influences the operation of the manipulable variables *pump heating* and *cenital ventilation* (which is also related to *lateral ventilation*), as one of the objectives in a greenhouse is to keep the temperature under control. In this sense, the *month* and *hour* influence outside temperature and radiation, as expected by domain experts.

Now, let us study the behaviour and use of the K-Dependence Bayesian classifier for predicting the energy demand in the greenhouse. The intervals used as class labels are those identified when the energy demand was discretized by open intervals. We are interested in knowing the effect that a dependences increase has on the classification of energy

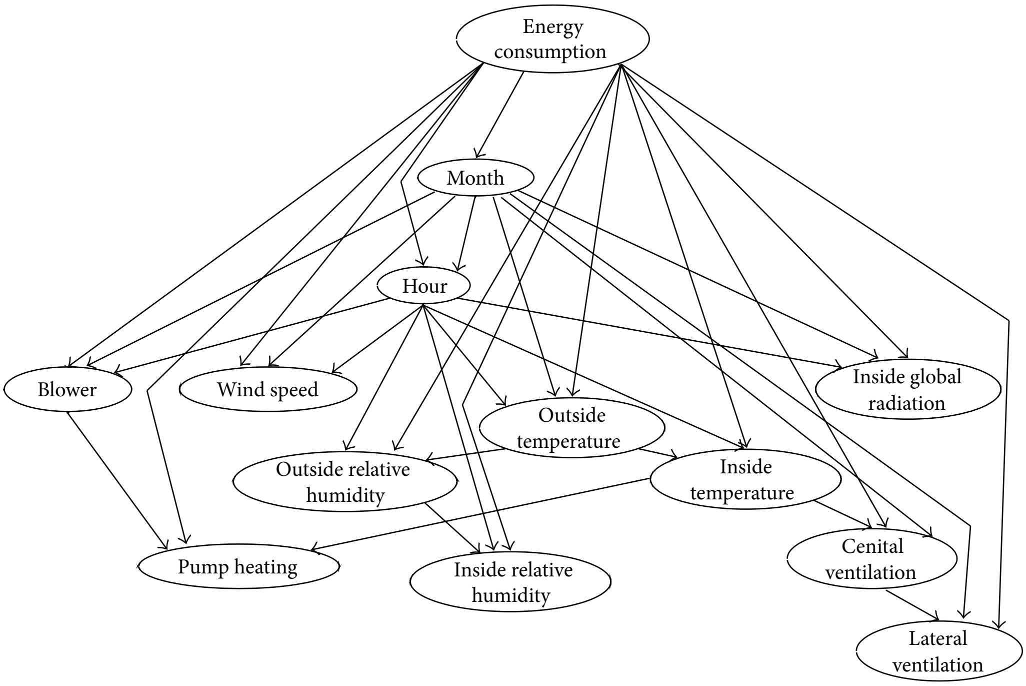


Figure 6: KDB2 structure learnt using open intervals ($b = 4$).

demand but keeping a balance between complexity and classification accuracy. Thus, we evaluate the classification accuracy of the model on each of the values of the class using three evaluation measures: *true positive* (TP) or *recall* rate, the proportion of observations which are correctly classified; *false positive* (FP) rate, the proportion of observations that were incorrectly classified in the class; and *precision*, the proportion of observations which truly belong to a given class among those classified in the class. Table 6 shows the values of these measures for the KDB2 classifier. KDB2 behaves bad in the energy demand interval $(2.01, 3.25]$, providing the worst TP rate and precision. Perhaps this is due to a more complex interaction between systems of the greenhouse in this energy demand range that is not learned by the model.

As example, to show how the KDB2 classifier works, the real energy demand in the greenhouse during December 2, 2014, discretized applying the open intervals identified in Section 3.2.3, is depicted at the top of Figure 7. Data from sensors are used as evidences and sent to the Bayesian network that propagates them in order to obtain the state of the class variable (i.e., the interval of energy demand) reaching the maximum a posteriori probability value. The predicted energy demand computed, in this way, by KDB2 classifier is depicted at the middle of Figure 7. An explanation for the differences between predicted and real energy demand around 9:00 and 19:00 is that they are due to the fact that, in the interval $(2.010, 3.25]$, where the errors reside, the KDB2 classifier returns the worst TP rate and precision. Also, they can be due to an overestimation induced from the use of less informative variables. This fact can be partially corrected making a redesign of the model by deleting those less informative variables. Doing this, we reduce the model's complexity and partially correct some overestimation (note that errors due to the discretization schema applied cannot be eliminated).

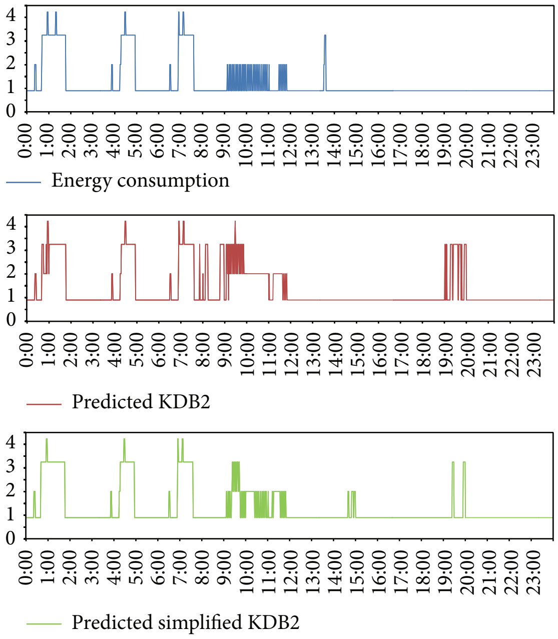


Figure 7: Real versus predicted energy demand on December 2, 2014.

Redesign of the Energy Demand Classification Model. The dependencies between variables and the complexity of the

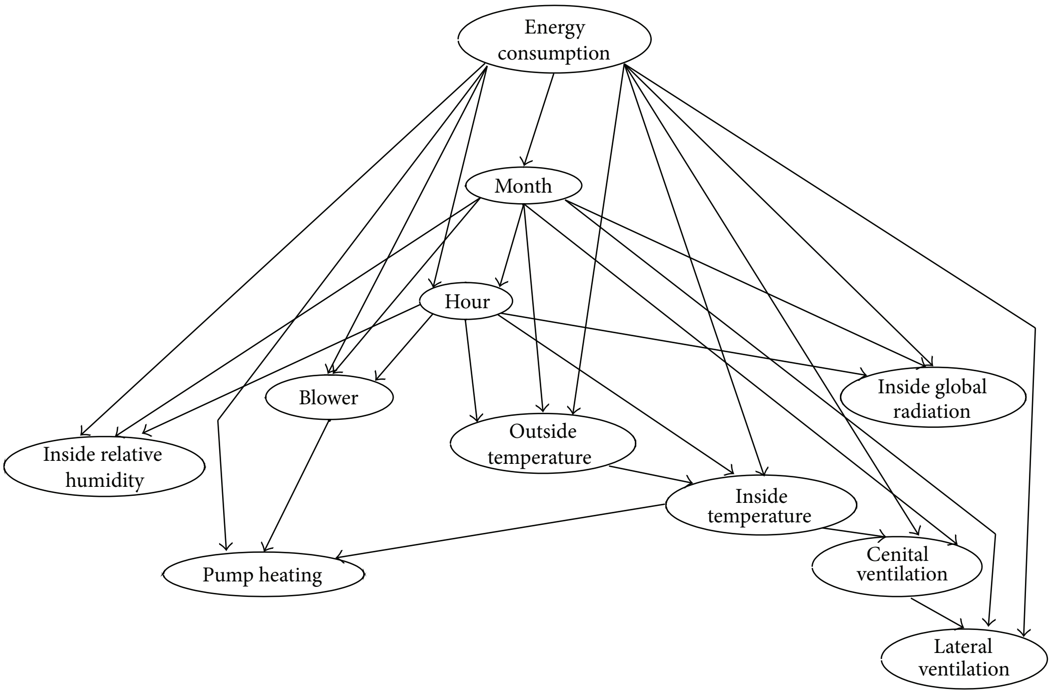


Figure 8: Structure of the simplified KDB2.

BN classifier have a direct relationship: the more the dependencies included in the model, the greater its complexity. The idea is to find a balance between complexity and classification accuracy, removing some of the initial variables. In order to decide which are the most relevant attributes, information gain ratio [30] has been used, which measures the relevance of an attribute with respect to the class. The process is as follows. First the attributes included in the KDB2 classifier are ranked using the information gain ratio (values near 1 are preferred); then those attributes with an information gain ration less than or equal to 0.01 are removed. Finally a new KDB2 classifier is learned on the selected attributes and the classification accuracy of the model is evaluated.

Table 7 shows the attributes ranked using their information gain ratio values. The attributes removed are *wind speed* and *outside relative humidity*, because they have an information gain ration less than 0.01. Figure 8 shows the structure of the new KDB2 classifier learned. This new model achieves a classification rate of 97.94%. Table 8 collects the measures of classification accuracy returned by the new KDB2, detailed by energy demand interval.

The simplification of the KDB2 model has implied an increase in the classification rate due to an increase in TP rate in all energy demand intervals except (2.01, 3.25]. Precision has also increased in all intervals except (3.25, 4.30], but here it is only slightly worse. The same applies to FP rate, but in this case there is a significant worsening at interval (3.25, 4.30]. We want to check if what these measures indicate is reflected in the behavior of the new simplified KDB2 classifier. Thus we draw on the real energy demand data in the greenhouse during December 2, 2014, discretized applying the open intervals identified in Section 3.2.3 and using the simplified KDB2 as a predictor, in the same way as in previous case. The energy demand predicted is depicted at the bottom of Figure 7. An improvement is observed, if the new predictions are compared with those of the initial KDB2. Although most errors remain in interval (2.01, 3.25].

Table 7: Variables sorted by decreasing value of information gain ratio.

Attribute	Information gain ratio
Blower	0.704
Pump heating	0.695
Cenital ventilation	0.097
Lateral ventilation	0.087
Outside temperature	0.065
Inside temperature	0.062
Month	0.060
Hour	0.047
Inside relative humidity	0.031
Inside global radiation	0.031
Wind speed	0.009
Outside relative humidity	0.007

Table 8: Performance results (%) of new KDB2 classifier.

Energy demand	TP rate	FP rate	Precision
[0.111, 0.90]	99.4	1.9	99.2
(0.90, 2.01]	95.2	0.8	96.3
(2.01, 3.25]	89.8	0.6	90.3
(3.25, 4.30]	97.5	1.6	95.5

4. Conclusions

The paper analyzes energy demand in a greenhouse and its characterization with the objective of building and evaluating classification models based on Bayesian networks: NB, TAN, and KDB. These models have the ability to extract from data, and depict graphically, relations among variables that can be easily understood. One advantage of a BN classifier is that it can provide an answer when receiving as input a partial set of observations of the variables included in the model. In future studies, the models learned can be used inside an energetic control system for the optimization of energy demand. The energy demand distribution in a greenhouse is analyzed and different discretization techniques are applied to reduce its dimensionality, paying particular attention to their impact in the classification model's performance. The classification accuracy of the models serves as basis for model selection. KDB2 has obtained the best evaluation and its behaviour as predictor and generalization capacity are tested on real energy demand data extracted from the greenhouse during a day that is not included in the training dataset. In order to find a balance in KDB2 between complexity and classification accuracy, some of the initial variables are removed using information gain ratio. The simplified KDB2 model shows better classification performance, being contrasted on the same real energy demand data. In both cases, misclassification errors suggest that interactions between greenhouse's systems deserve a deeper study from the point of view of energy demand characterization. The analysis of energy demand in the greenhouse, the discretization techniques applied, and the BN classifiers allow characterizing energy demand and make predictions based on observations with more than acceptable classification rates. The method used in this paper can be used in greenhouses of any geographical area. In future we plan to use BN classifiers as additional technique to control energy demand in the greenhouse. Furthermore, tests will be done in order to eliminate internal variables.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research has been funded by the "Controlcrop" Project, P10-TEP-6174, supported by the Andalusian Ministry of Economy, Innovation and Science (Andalusia, Spain), and the project DPI2010-21589-C05-04 supported by the Spanish Ministry of Science and Innovation.

References

[1] A. Ramírez-Arias, F. Rodríguez, J. L. Guzmán, and M. Berenguel, "Multiobjective hierarchical control architecture for greenhouse crop growth," *Automatica*, vol. 48, no. 3, pp. 490–498, 2012.

[2] J. J. Hanan, *Greenhouses: Advanced Technology for Protected Horticulture*, CRC Press, Boca Raton, Fla, USA, 1998.

[3] Y. Tüzel and C. Leonardi, "Protected cultivation in mediterranean region: trends and needs," *Journal of Ege University Faculty of Agriculture*, vol. 46, pp. 215–223, 2009.

[4] J. A. Sánchez, F. Rodríguez, J. L. Guzmán, and M. R. Arahal, "Virtual sensors for designing irrigation controllers in greenhouses," *Sensors*, vol. 12, no. 11, pp. 15244–15266, 2012.

[5] N. Castilla and E. Baeza, "Greenhouse site slection," in *Good Agricultural Practices for Greenhouse Vegetable Crops—Principles for Mediterranean Climate Areas*, FAO Plant Production and Protection Paper, pp. 21–34, 2013.

[6] J. K. Gruber, J. L. Guzmán, F. Rodríguez, C. Bordons, M. Berenguel, and J. A. Sánchez, "Nonlinear mpc based on a volterra series model for greenhouse temperature control using natural ventilation," *Control Engineering Practice*, vol. 19, no. 4, pp. 354–366, 2011.

[7] C. Von Zabeltitz, *Integrated Greenhouse Systems for Mild Climates: Climate Conditions, Design, Construction, Maintenance, Climate Control*, Springer, 2010.

[8] G. Giacomelli, N. Castilla, E. van Henten, D. Mears, and S. Sase, "Innovation in greenhouse engineering," in *Proceedings of the International Symposium on High Technology for Greenhouse System Management (Greensys '07)*, vol. 801, pp. 75–88, 2007.

[9] W. Baudoin and E. Baeza, "Good agricultural practices for greenhouse vegetable crops: principles for mediterranean climate areas," FAO Plant Production and Protection Paper, 2013.

[10] G. Zaragoza, M. Buchholz, P. Jochum, and J. Pérez-Parra, "Watergy project: towards a rational use of water in greenhouse agriculture and sustainable architecture," *Desalination*, vol. 211, no. 1–3, pp. 296–303, 2007.

[11] A. Baille, J. C. López, S. Bonachela, M. M. González-Real, and J. I. Montero, "Night energy balance in a heated low-cost plastic greenhouse," *Agricultural and Forest Meteorology*, vol. 137, no. 1-2, pp. 107–118, 2006.

[12] N. Castilla and J. Hernandez, "Greenhouse technological packages for high-quality crop production," *Acta Horticulturae*, vol. 761, pp. 285–297, 2007.

[13] S. H. Lee and I. H. Suh, "Bayesian network-based behavior control for skilligent robots," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '09)*, pp. 2910–2916, May 2009.

[14] K. Numata, S. Imoto, and S. Miyano, "A structure learning algorithm for inference of gene networks from microarray gene expression data using Bayesian networks," in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE '07)*, pp. 1280–1284, January 2007.

[15] M. Hunt, B. von Konsky, S. Venkatesh, and P. Petros, "Bayesian networks and decision trees in the diagnosis of female urinary incontinence," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 551–554, IEEE, July 2000.

[16] H. Handa and O. Katai, "Estimation of Bayesian network algorithm with GA searching for better network structure," in *Proceedings of the International Conference on Neural Networks and Signal Processing (ICNNSP '03)*, vol. 1, pp. 436–439, December 2003.

[17] S. R. Tinoco-Martínez, F. Calderon, C. Lara-Alvarez, and J. Carranza-Madrigal, "Una técnica bayesiana y de varianza mínima para segmentación del lumen arterial en imágenes de ultrasonido," *Revista Iberoamericana de Automática e Informática Industrial*, vol. 11, no. 3, pp. 337–347, 2014.

[18] M. Wang and J. Zhou, "A Bayesian network-based classifier for machining error prediction," in *Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM '14)*, pp. 841–844, IEEE, July 2014.

[19] J. del Sagrado, F. Rodríguez, M. Berenguel, and R. Mena, "Bayesian networks for greenhouse temperature control," in *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, Á. Herrero, B. Baruque, F. Klett et al., Eds., vol. 239 of *Advances in Intelligent Systems and Computing*, pp. 161–170, Springer International, 2014.

[20] F. Rodríguez, M. Berenguel, J. L. Guzmán, and A. Ramírez, *Modelling and Control for Greenhouse Crop Growth*, Springer, London, UK, 2015.

[21] J. A. Sánchez-Molina, J. V. Reinoso, F. G. Acién, F. Rodríguez, and J. C. López, "Development of a biomass-based system for nocturnal temperature and diurnal CO_2 concentration control in greenhouses," *Biomass and Bioenergy*, vol. 67, pp. 60–71, 2014.

[22] T. D. Nielsen and F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2009.

[23] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[24] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[25] M. Sahami, "Learning limited dependence bayesian classiérs," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96)*, vol. 96, pp. 335–338, Portland, Ore, USA, 1996.

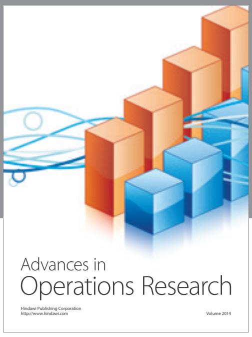
[26] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309–347, 1992.

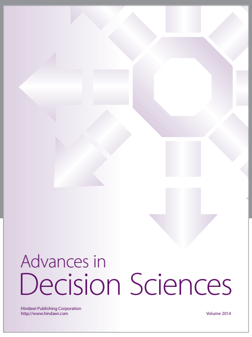
[27] Agencia Andaluza de la Energía, *Guía de Ahorro y Eficiencia Energetica en Municipios*, Agencia Andaluza de la Energía, Sevilla, Spain, 2011.

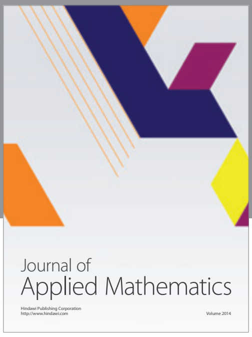
[28] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI '93)*, pp. 1022–1027, 1993.

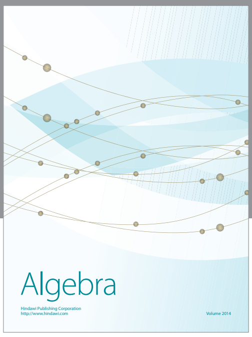
[29] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," in *Proceedings of the 12th International Conference on Machine Learning*, pp. 194–202, Tahoe City, Calif, USA, July 1995.

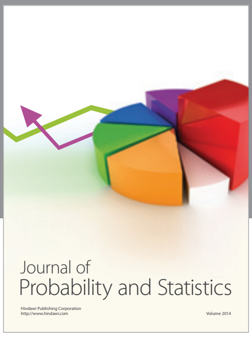
[30] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs decision trees in intrusion detection systems," in *Proceedings of the 2004 ACM Symposium on Applied Computing*, pp. 420–424, March 2004.


Advances in
Operations Research

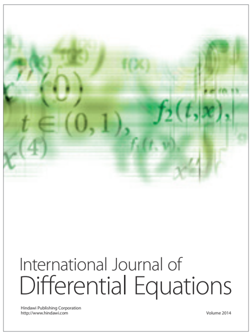

Advances in
Decision Sciences


Journal of
Applied Mathematics

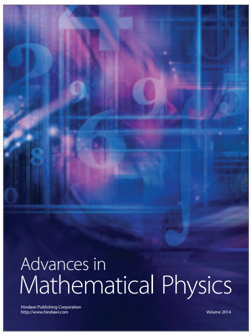

Algebra

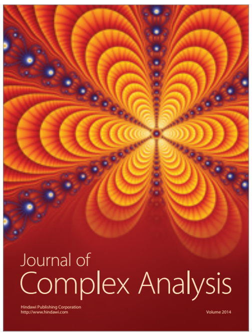

Journal of
Probability and Statistics

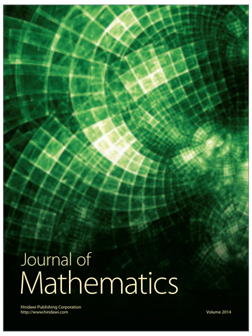
The Scientific
World Journal

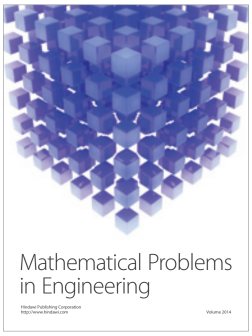

International Journal of
Differential Equations

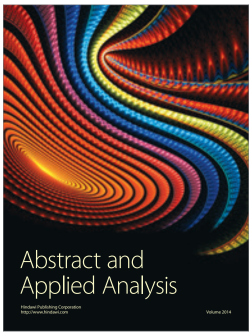
International Journal of
Combinatorics

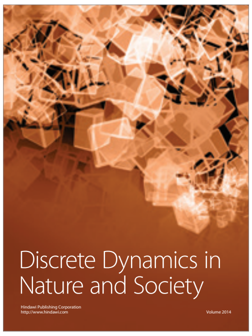

Advances in
Mathematical Physics

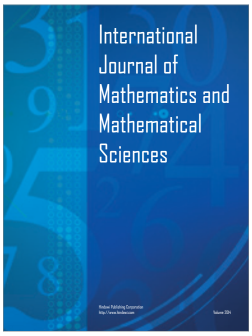

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

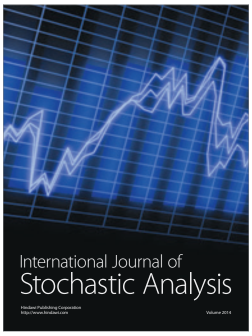

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization