Bayesian Network for Predicting Energy Consumption in Schools in Florianópolis – Brazil

Matheus Soares Geraldi¹, Mateus Vinicius Bavaresco¹, Enedir Ghisi¹
¹Federal University of Santa Catarina, Florianópolis, Brazil

Abstract

It is important to study innovative approaches that consider real-world data to predict energy consumption, especially in existing buildings. This paper presents a data-driven model to predict energy consumption using Bayesian Networks. Monthly energy bills over three years were obtained from 90 public schools in Florianópolis, southern Brazil. Information such as floor-plan area, number of students, type of education, number of floors and occurrence of events were gathered for each building. The network output indicator was assessed using Energy Use Intensity based on floor-plan area or number of students. Three types of discretization methods and three network structures were tested, generating eighteen networks. A performance analysis comparing predicted as well as real Energy Use Intensity determined the Normalized Root Mean Square Error for each network and pointed out Equal Width Discretization as the best method and Naïve-Bayes as the most advantageous structure type. The discretization method had a high impact on the network performance. In addition, the Energy Use Intensity based on floor-plan area was more reliable than that based on the number of students.

Introduction

Predicting energy consumption is important for public buildings once it yields better resource management and improves optimization and retrofits. In Brazil, buildings were responsible for 43% of the total energy demand in 2017 (Brazil, 2017).

For new buildings, the energy consumption is commonly predicted using computer simulation or simplified models. However, predicting energy consumption for actual buildings lacks dependable methods along with challenges to be addressed, such as the non-availability of measured data and the unwillingness to share existing data (Borgstein and Lamberts, 2014). Thus, a statistical approach can be helpful to overcome those difficulties and obtain a representative result.

Computer simulations and statistical approaches are complementary techniques. Building simulation using software (e.g., EnergyPlus) is useful for forward modelling at the building design phase and data-driven modelling is useful for retrofits, building performance analysis and energy purchasing in smart grids (Kontokosta and Tull, 2017). Moreover, authors have emphasized the gap between simulation results and actual performance (Jones et al., 2015; Khoury et al., 2017; Menezes et al., 2012), and this fact leads researchers to look for new ways to predict energy consumption by calibrating forward models or using data-driven modelling (Tardioli et al., 2015; Wei et al., 2018). An evident need to forecasting energy consumption based on real data is expressed by recent studies, especially for existing buildings (Hamilton et al., 2015, 2016; Huebner et al., 2015; Huebner et al., 2016; Stuepels et al., 2013).

Statistical approaches can be used to create data-driven models using information from the building stock. Usually, data-driven approaches are used to calculate baselines for measurement and verification proceedings (e.g., retrofits) (Burman et al., 2014; IPMVP, 2001). Human variables can also be taken into account to improve the accuracy of such data-driven models by considering user’s behaviour (Liang et al., 2016).

Wei et al. (2018) categorized data-driven approaches in two major classes: to predict building energy consumption (using artificial neural networks, support vector machines, statistical regression, decision tree and genetic algorithm); or to classify building energy consumption (using k-mean clustering, self-organizing map and hierarchy clustering).

For example, using a data-driven approach, Lindelöf et al. (2018) developed a model to reduce the cost of retrofitting by estimating the baseline period through a Bayesian verification (no specific monitoring period was needed). In addition, the result was expressed by probability density function, which gives a confidence interval and not a static value. This confidence interval, rather than a blind result, supports the stakeholder to decide about the most suitable retrofit.

The Bayes theorem introduces the conditional probability, which is a powerful approach to face complex and interactive problems, such as energy consumption, due to its capacity to express results in a degree of uncertainty (Borunda et al., 2016). Bayesian Networks are acyclic graphical models used to assess inferences regarding the relationship between input and output variables, represented as nodes. This technique has been used for many purposes, such as calibrating simulation models (Heo et al., 2011; Heo et al., 2012), defining archetypes for stock modelling (Menberg et al.,...
To forecast energy consumption, Bassamzadeh and Ghanem (2017) proposed a Bayesian Network to predict the demand for purchasing energy in Smart Grids based on high resolution data (5 min). The aim was to investigate the dependent relationships between contributing variables and to build a model to help managers to buy the right amount of energy using high-granularity data. The main variables used were past demand, outside temperature, weekday or weekend day and price of energy. O’Neill and O’Neill (2016) proposed a Bayesian Network for forecasting hot water energy consumption in an office building. The model was based on an hourly dataset of outside temperature and energy demand collected during the cold season. The Network was very dependent on the discretization process and the uncertainty had increased when predicting consumption in other seasons.

However, both proposed frameworks required a high level of data resolution and lacked to predict consumption when the building was not monitored with smart meters. In developing countries as Brazil, buildings with monthly energy measurement (billing) are the majority and there is a need to develop models to consider this portion of the stock. In addition, despite the high availability of the Bayesian Network to predict energy demand in buildings, there is a lack of a tool for forecasting electric energy use intensity (EUI), since it has been used to estimate HVAC demands in cold locations. The objective of this work is to introduce a data-driven model using Bayesian Network to predict electric energy use intensity in public schools based on monthly data. The model was constructed using billed data from 90 state schools in Florianópolis, southern Brazil. The method presents the step-by-step construction of the Bayesian Network and registers how the nodes, data classifications and node connections were settled to obtain an optimal network performance.

Method
The study is comprised of three steps: dataset characterization and discretization process, Bayesian Network construction, and performance analysis. Figure 1 shows the method flow chart.

Dataset

EFD

EWD

PKID

Naive Bayes

Net Bayes

TAN

Naive Bayes

Net Bayes

TAN

Naive Bayes

Net Bayes

TAN

Naive Bayes

Net Bayes

TAN

Naive Bayes

Net Bayes

TAN

Comparison of performance indicators (real vs predict)

Optimal Bayesian Network

Figure 1: Method flow chart.

Dataset characterization
The first step was to collect and to characterize the dataset. Data from 118 schools were obtained and each school was characterized by including information such as monthly energy bills from 2016 to 2018, floor-plan area (in m²), number of students, type of education (basic, high or both). The presence or not of significant events was also identified for each monthly bill in order to indicate if the school had a situation, such as a party or science fair, which impacts on the energy consumption. The data collection process was based on the Energy Star® Portfolio Management worksheet for schools K-12 (EPA, 2016).

A first exploratory analysis cleaned the dataset aiming to exclude missing data, removing outliers and correcting inconsistent data. Some inconsistent data that should be avoided (such as hotwire power supply) could lead to an unreliable modelling. This process included the comparison of the average and the standard deviation of the annual electricity consumption. Schools with average annual consumption differences greater than 50% and schools with monthly peaks greater than two standard deviations of the annual consumption were excluded. The result of this step was the dataset ready to be used for discretizing.

The EUI was obtained for each school in two ways: rating the monthly consumption by floor-plan area and by number of students. Despite the relationship between the consumption and the building characteristics, a Bayesian Network was proposed because it is not possible to predict energy consumption using a directly linear model. Figure 2 shows electricity consumption versus floor-plan area and number of students and evidences that a linear model will be very unreliable.

DISCRETIZATION PROCESS
Values of an attribute are either discrete or continuous. The conditional probability of an attribute $X_i$ that will take a particular state $x_i$ when the value of the class $C$ that conditions $X_i$ is $c$ can be described as Equation 1.

$$P(X_i = x_i | C = c)$$  \hspace{1cm} (1)

Where $X_i$ is an attribute of the variable $X$; $x_i$ is a given state of the attribute; $C$ is a class of the attribute; and $c$ is an adopted value for class $C$.

Attribute value $c$ for the class $C$ is considered a discrete value with a finite range. When the class $C$ has a
continuous attribute range, it is necessary to group them into classes to calculate a probability function. Even for attributes that have a discretized but large amount of values, it is often advisable to group ranges of values into a smaller range for the purpose of estimating the probabilities. So, it is needed to discretize the continuous variables to allow the calculation of the probabilities.

Some of the variables gathered were continuous and needed discretization, but some variables were discrete. Table 1 shows a summary of the variables used to construct the Bayesian Network.

Table 1: Summary of the variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable type</th>
<th>Node number</th>
<th>Node type</th>
<th>Nº of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUI</td>
<td>Continuous</td>
<td>1</td>
<td>Output</td>
<td>5</td>
</tr>
<tr>
<td>Area</td>
<td>Continuous</td>
<td>2</td>
<td>Input</td>
<td>3</td>
</tr>
<tr>
<td>Students</td>
<td>Continuous</td>
<td>3</td>
<td>Input</td>
<td>3</td>
</tr>
<tr>
<td>Floors</td>
<td>Discrete (One or Two)</td>
<td>4</td>
<td>Input</td>
<td>2</td>
</tr>
<tr>
<td>Education</td>
<td>Discrete (Class 1, 2 or 3)</td>
<td>5</td>
<td>Input</td>
<td>3</td>
</tr>
<tr>
<td>Month</td>
<td>Discrete (12 months)</td>
<td>6</td>
<td>Input</td>
<td>12</td>
</tr>
<tr>
<td>Event</td>
<td>Discrete (Yes or No)</td>
<td>7</td>
<td>Input</td>
<td>2</td>
</tr>
</tbody>
</table>

To categorize the continuous variables, three discretization methods were used aiming the best network performance. Three methods were tested because, as concluded by O’Neill and O’Neill (2016), the discretization criteria have a high impact on the network output. The discretization methods were chosen according to Yang and Webb (2002) due the nature and amount of data available in this work.

The first discretization approach was EWD (Equal Width Discretization). This method divides the number of observations into k intervals of equal width, where k corresponds to the number of classes. The interval width is given by Equation 2.

\[ D_{EWD} = (v_{\text{max}} - v_{\text{min}})/k \]  

Where:

- \( D_{EWD} \) is the intervals width;
- \( v_{\text{max}} \) is the maximum observed value;
- \( v_{\text{min}} \) is the minimum observed value; and
- \( k \) is the number of intervals.

The cut points start in \( v_{\text{min}} \) and continue by summing \( D_{EWD} \) until \( v_{\text{max}} \).

The second discretization approach was the EFD (Equal Frequency Discretization). This method divides the dataset into k intervals where each one contains approximately the same number of training cases (equal frequency). \( k \) is a predefined factor equal to the number of classes.

Both EWD and EFD methods possibly jeopardize attribute information since \( k \) is determined by an assumption and without considering the dataset properties. However, both methods are often used and work surprisingly well for Naïve-Bayes classifiers (Yang and Webb, 2002). To explore that limitation, the final discretization approach was the PKID (Proportional k-Interval Discretization) which discretizes the dataset into \( k \) intervals with \( s \) size. PKID adjusts discretization because it considers the relationship between interval size and number of intervals. The higher the number of intervals, the smaller the interval size. Consequently, the larger the interval size (the smaller the number of intervals), the lower the variance but the higher the bias. The opposite is true. PKID gives equal weight to discretization bias and variance decrease by setting both interval size and interval number equally and proportionally to the dataset size, as presented in Equation 3.

\[ s = t = (n/k)^{1/2} \]  

Where \( s \) is the interval size; \( k \) is the number of intervals; and \( n \) is the dataset size.

The result of this step was a table with all variables classified according to the three discretization criteria adopted.

**Bayesian Network construction**

The construction of the Bayesian Network was based on the Bayesian Theory and the calculation of conditional probabilities as shown in Equation 4.

\[ P(A|B) = P(B|A).P(A)/P(B) \]  

where \( P(A|B) \) is the probability \textit{a posteriori} of the event A, conditioned by the event B; \( P(B|A) \) is the probability \textit{a priori} of the event B that conditions an event A; \( P(A) \) is the probability \textit{a priori} of the event A; and \( P(B) \) is the probability \textit{a priori} of the event B.

The probabilities \textit{a priori} were based on their own variable distribution and the probabilities \textit{a posteriori} were calculated as a function of the conditioning event.

The nodes represent the attributes of the observations, i.e., the variables gathered in the dataset. Node classes represent the variable states.

The connections among the nodes are called directed arc of probability and express the likelihood that the arrow-headed node conditions the arrow-ended node. The structure of the network depends on the arrangement of those arcs among the nodes.

Those probabilities were assessed by calculating the frequency of occurrence for each variable state related to the variable state of the connected node using a frequency table. Table 2 shows the questions asked whose response values complete the frequency table for two generic nodes.

Table 2: Analytical composition of a generic frequency table of two nodes with two classes.

<table>
<thead>
<tr>
<th>Node A</th>
<th>Node B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>Class 1</td>
</tr>
<tr>
<td>Class 2</td>
<td>Class 2</td>
</tr>
</tbody>
</table>

| How many observations had Node A classified as Class 1 and Node B classified as Class 1? |
| How many observations had Node A classified as Class 1 and Node B classified as Class 2? |
Three structures of Bayesian Network were tested. The most basic one was Naïve Bayes, wherein the input nodes were settled by a direct arc to the output node. The Tree Augmented Naïve-Bayes type (TAN) was similar to Naïve Bayes, but arcs among input nodes were used. The Net Bayes type combines the arcs among input nodes and the use of latent nodes – nodes not directly connected to the output node. Usually, latent nodes are used to represent non-observational variables. Figure 3 presents a graphical example of those three types of structure.

![Graphical example of Bayesian Network structures](image)

**Figure 3: Differences among Bayesian Network structures.**

Therefore, three types of networks were developed each one combining three different discretization approaches for continuous variables, resulting in nine networks. To use the network, one must settle the class in all input nodes regarding to its school information according to the intervals defined in the discretization. The output node was the energy use intensity, and, despite this node was denoted by a discrete node, the output is a continuum value because the Bayesian Network can be used as a predictor by multiplying the output probabilities of each output class by their average value. The predicted energy use intensity can be estimated by means of Equation 5.

\[
EUI = \Sigma (p_i \times a_i)
\]

Where \( EUI \) is the predicted energy use intensity (output of the Bayesian network); \( p_i \) is the probability outputted for each class; and \( a_i \) is the average value of energy intensity for each interval class of the EUI node.

The standard deviation of the predicted energy use intensity was also calculated, using the same mathematical structure of Equation 5 but considering the standard deviation instead of the average value of each interval class of the EUI node. Considering a distribution of Student, the final predicted EUI value was shown with a confidence interval, i.e., predicted EUI ± or – confidence interval. In the first attempts, the total consumption was used as output node, but it was found that using EUI as output node the network achieved more reliable results.

Two types of EUI were used to test the network performance. The first EUI was based on the rate between energy consumption per month and floor-plan area, used by many authors to measure performance (Chung 2011). The second was based on energy consumption per month and number of students, as suggested by Dias Pereira et al. (2014) to assess performance of school buildings.

**Performance analysis**

The performance analysis measured the capacity of each network to predict reliable results. A dataset containing known-output values (real values) were inputted in the network and the outputs predicted by the network were compared to those already acknowledged. This process was performed using a bootstrap routine which resamples the dataset that was used to train the network.

A first analysis compared the predicted output values with the known-output values by verifying if the known-output values were fitted inside the confidence interval outputted by the network. If so, the network was considered able to predict monthly EUI, considering the confidence interval. This analysis compared the correct and the non-correct results grouping them in a bar chart. To determine the network that had the best performance, a numerical approach was used. To assess the performance, the NRMSE (Normalized Root Mean Square Error) was adopted as an indicator, which can be calculated according to Equation 6.

\[
NRMSE = 100 \times \left( \frac{\Sigma (y'_i - y_i)^2}{(y_{max} - y_{min})^2} \right)^{1/2}
\]

Where \( NRMSE \) is the Normalized Root Mean Square Error (indicator of performance); \( y'_i \) is the predicted EUI (network output) corresponding to the case \( i \); \( y_i \) is the real EUI corresponding to the case \( i \); \( t \) is the number of cases; \( y_{max} \) is the maximum value of real EUI; and \( y_{min} \) is the minimum value of real EUI.

In summary, the square difference of the predicted output energy use intensity and the real energy use intensity was computed for each case of the bootstrapping sample, divided by the number of cases in the sample and then the square root was taken. Then, the indicator was normalized to exclude the influence of the scale on the result. This indicator expresses a reliable network performance measure because it gives the idea of the potential global fitness (Hyndman and Koehler, 2006). Thus, the lower the NRMSE the stronger the network for predicting energy intensity.

The NRMSE was calculated using the bootstrapping for each network combination, i.e., three discretization methods times three structures equals nine networks. Since two EUI indicators were used as output node, the NRMSE was calculated for the networks that use EUI as a function of floor-plan area and as a function of number of students, totaling eighteen networks.
A final performance analysis used the best network selected from the previous analysis (smaller NRMSE) to predict the EUI over a year for a sample of schools (monthly). This result was compared to the real EUI to evidence the Bayesian Network behaviour in practice.

Results

The Bayesian Network

Eighteen Bayesian Networks were built. An example of a Naïve Bayes Network is shown in Figure 4. Each square represents a node, and their classes are displayed by bars with the probabilities a priori. Those probabilities a priori were calculated using the dataset to train the network. The trained network can be used to insert a case by setting the states on classes of each node according to the classes of the case analysed.

Then, the output node gives the predicted EUI by calculating the probabilities a posteriori multiplied by the average values of the classes of the output node. The NRMSE was calculated for each combination of discretization method and structure types used to build the networks. The networks were identified by means of an id number, followed by the discretization method and structure types used to build the networks.

The NRMSE was calculated for each combination of discretization method and structure types used to build the networks. The networks were identified by means of an id number, followed by the discretization method and structure types used to build the networks.

![Figure 4: Bayesian Network.](image)

A first analysis compared a resampled dataset with real output values with those predicted by the network. If the real value was within the confidence interval, the result was considered a correct prediction; otherwise, it was considered an incorrect prediction. Figure 7 shows the result of this analysis.

![Figure 5: Correct and incorrect predictions.](image)

The variation in results due to using different structures was very low. The major variation was due to the utilization of different discretization methods.

The discretization method that led to the best result — higher correct predictions — was the EWD. Both EFD and PKID methods led to higher inaccuracy predictions. This could be explained because both EFD and PKID methods considered low variance among bins, what does not represent the dataset.

In this analysis, it was not possible to define which output node was the best (EUI rated by floor-plan area or by number of students).

Performance analysis

To compare the representation capacity of the output node two indicators were used: EUI based on kWh/m².month and EUI based on kWh/student.month. Figures 6 shows the NRMSE for the networks 1 to 9, which used kWh/m².month, and Figure 7 shows the NRMSE for the networks 10 to 18 which used kWh/student.month. The NRMSE analysis assessed the numerical difference between the predicted and the real values. So, it was an analysis with more accuracy.

![Figure 6: NRMSE for energy use intensity based on floor-plan area (kWh/m².month).](image)

![Figure 7: NRMSE for energy use intensity based on number of students (kWh/student.month).](image)

Assessing the NRMSE, the best discretization method was the EWD. This is due to the nature of the continuum variables that are best categorized using a simple binning method that led to more representative bins. Indeed, the data showed that same-sized bins (as created by EFD and PKID methods) do not represent the actual data behaviour observed in the dataset. For example, for the variable EUI, the majority of the data (90%) ranged from 0.3 to 6.0 kWh/m².month, while 10% ranged from 6.0 to 10.0 kWh/m².month.
16.0 kWh/m².month. For the nodes “Area” and “Students”, the trend is more linear. This could be explained because school buildings in general have a very constant pattern of consumption despite a few outlier values as observed. In addition, PKID method cuts the dataset into a lot of bins, what might make the network performance decrease due to the increase of the bias. This method leads to a better performance for small datasets (Yang and Webb, 2002), so it did not fit well on this dataset.

The best structure method was the Naïve-Bayes which was also the simplest one used. This could be explained because there was no conditional relation among the variables used to describe the consumption. Despite the attempts to link the nodes and create statistical relations, a more effective statistical test must be performed to find real relationships among the input variables.

Comparing the output unit used, either relative to the floor-plan area or to the number of students, the best network performance (lower NRMSE) was observed using the EUI rated by floor-plan area.

This could be explained because the appliances used in schools in Brazil are usually HVAC for cooling, lights and computers for employees, as observed in situ. This makes the consumption somehow independent from the number of students and more dependent from the size of the school. For example, if the lights and HVAC were turned on in a classroom, they would consume very similar energy either for ten or thirty students.

In fact, if one takes the EUI rating by floor-plan area or number of students, it is possible to observe the differences, as shown in Figure 8.

Figure 8: Histogram of EUI rated in relation to (a) floor-plan area and (b) number of students.

The kurtosis of the histogram (a) was 0.9 and for the histogram (b) was 14.0, which indicates that the peak of the frequency-distribution curve of (b) was sharper than (a). In other words, the histogram of EUI rated by floor-plan area presented more similarity to the normal distribution if compared to EUI rated by number of students. Despite the impossibility to determine a cause-effect model as shown in Figure 2, the relationship between energy consumption and floor-plan area was very strong, since the observations were more equally distributed among the bins.

To analyse the network usage in practice, each one of the 90 schools were inserted in the network to predict their EUI throughout the year. The best performance network structure combination was used: Naïve-Bayes constructed with EWD that outputs EUI rated by floor-plan area (network #1). Figure 9 shows the real and predicted EUI for a single school as example.

Figure 9: Real vs predicted EUI for a single school throughout the year (real data over 2016).

In general, the predicted values overlapped on real values, but this case was chosen to illustrate some interesting observations, such as noted in other cases.

From Figure 9 one can observe that the predicted values were very similar to the real values despite some outlier points as in March, June and August. In March and August, there were indications of some peak consumptions, probably due to the educational schedule variation from one year to another (sometimes school year starts in February and sometimes in March, depending on holidays, and returns from winter break either in July or in August). Since the dataset used was composed of three years, an improvement in the network could be made by using larger dataset, contemplating several years.

In June, the energy usage was lower than the predicted also probably due to the schedule variation (holidays in June) or because most of the schools had events in June that are not common on that school.

In fact, the network was a more stable predictor, i.e, the predicting throughout the year followed a steady trending but not accomplished peaks or outlier values. This is also concluded by other authors (Borunda et al., 2016; Sokol et al., 2017), who recommended the Bayesian Network as a robust technique for classification, but not so robust to be used in regression analysis, as predicting energy consumption requires.

Another possible explanation for these inaccuracies can be found in a limitation of the network: the incapacity of differentiation year to year. The network always predicts for an undefined year and it is not possible to predict for 2019 or 2020 differently, for example. However, the purpose behind the Bayesian Network is retro-feeding, i.e., the possibility of inserting new information successively. Therefore, while more information is added to the network more possibilities to predict specific events can be built. Some examples of improvement are the addition of specific information as weather data (e.g., average monthly temperature), or
rearrangement of the network structure due to newly found relationships.

**Conclusion**

This paper presented a Bayesian Network constructed to predict energy consumption in schools in Brazil. The database of 90 schools was used. This study analysed the performance of eighteen networks built using different structures and discretization methods. This study addresses the noteworthy findings as follows:

(a) The type of discretization method expressively impacted on the network performance, and the EWD was the most suitable method for discretizing the continuum variables considered in the network construction of this study;
(b) EUI rated by floor-plan area led to better predictions instead of EUI rated by number of students due to the relationship between appliances and building sizes found in schools;
(c) The network with the best performance could lead to a proper accuracy, but it failed to predict peak values or to distinguish one year from another;
(d) Some limitations of this method could be drawn, such as the need of a greater dataset and the incapacity to predict peak or outlier values. However, if the limitations are solved, the Bayesian Network will be useful in contexts where there is limited information available, such as developing countries like Brazil.

The detailed reasons for the aforementioned findings will be investigated in future studies. For example, the improvement of the database with more characteristics and enlargement of the actual data. Furthermore, the use of Bayesian Networks for classifying energy performance instead of predicting energy consumption would be explored.

**Acknowledgement**

The authors would like to thank the Brazilian governmental agency CAPES – “Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” for the financial support that enabled the development of this research.

**References**

for Determining Energy Savings in New Construction - Volume III.


