Bayesian network feeding into a water sector demand and supply model.

Christiaan M. van der Walt, Nicolene Botha
CSIR, Modelling and Digital Science (Advanced Mathematical Modelling)

1. Introduction

Probabilistic graphical models (PGMs) combine graphical networks that define relationships between variables with probability theory. This allows for probabilistic reasoning and scenario analysis while paying respect to the defined relationships between variables. These models have been used with great success in numerous application areas due to their generality. PGMs can be used in a data driven fashion where the graphical structure (relationships between variables) and model parameters can be learned from data; or the model structure and parameters can be specified by a domain expert. A combination of expert knowledge and data can also be used in a single model - which makes PGMs extremely powerful for many real-world problems. Bayesian networks (BNs) are a specific class of PGMs that are directed, since causalities are modelled as directed graphs.

We investigate the use of Bayesian Networks on water demand data for South Africa. This model provides the typical input into predicting the stress on the water demand due to changes in populations.

In Section 2 we provide a description of the water demand data that was obtained and how the data was processed for the purposes of our experiments. In Section 3 we provide a BN and we show how this network can be trained from data. We also provide an example of how this network can be used to answer an interesting question by performing scenario analysis. In Section 4 instructions on how to load the trained model discussed in Section 3 are given. Section 5 lists initial technical requirements of the final model. In Section 6 conclude on our findings and propose directions for future work, including how the BN framework can be applied to perform integrated modelling of various sectors of the economy, where skills demand and supply forecasting is required.

2. Dataset description

Two public data sources of South African water data was identified, namely the National Water Services Knowledge System (WSKS) (https://www.dwa.gov.za/wsks/) and the National Integrated Water information System (NIWS) (http://niwis.dws.gov.za/). Both of these public data sources are made available by the South African Department of Water and Sanitation (DWA). The WSKS system contains data relating to water demand on municipal level (including demographic data), while the NIWS system contains data relating to water supply.

For this model the focus was on water demand modelling using the WSKS data. This dataset consists of datasets describing: access to basic services, demography, financials, water conservation and demand, water quality management and water boards. Each of these categories contained various datasets. For the purpose of this demand model, data from access to basic services, demography and water conservation and demand on municipal level, were used and summarised in Table 1.

Table 1: Summary of water demand data

Category	Datasets
Access to basic services	Households served with water
	People served with water
Demography	Household density
	Households rural/urban
	Population rural/urban
Water conservation and demand	Input litres per capita per day
	Billed litres per capita per day
	Billed metered consumption
	Billed m3 per household per month

The distribution of the values for each of the variables in Table 1 is displayed in Figure 1.

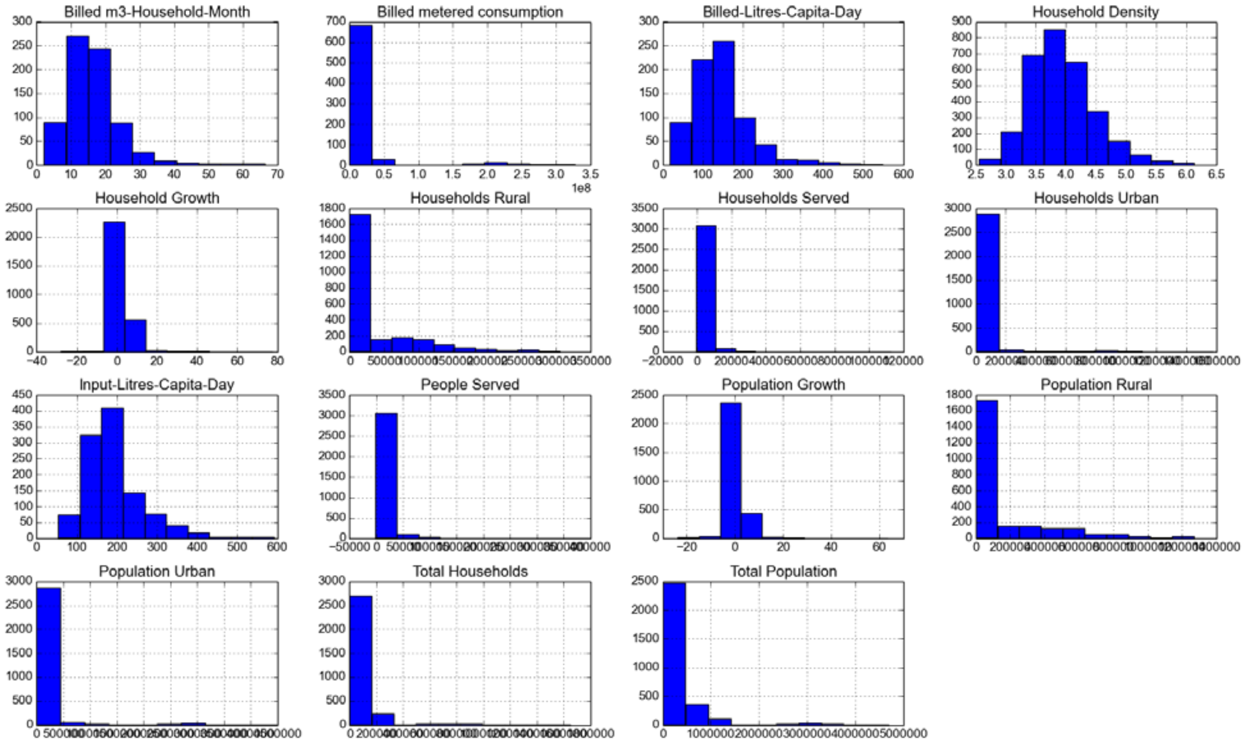


Figure 1. Distribution of variables in the dataset (equally spaced bins)

3. Example of Bayesian Network scenario analysis

A simple BN was constructed where it is assumed that the rural population and urban population across all municipalities have a direct causal influence on the input litres per capita per day of a municipality. Thus, ‘Population_Rural’ and ‘Population_Urban’ directly affect ‘Input_Litres_Capita_Day’. This network is illustrated in Figure 2. Each node in this network thus represents a variable and each arrow indicates the existence and direction of causality between variables. The monitor for each node provides the probability distribution of the variable that is represented by the node.

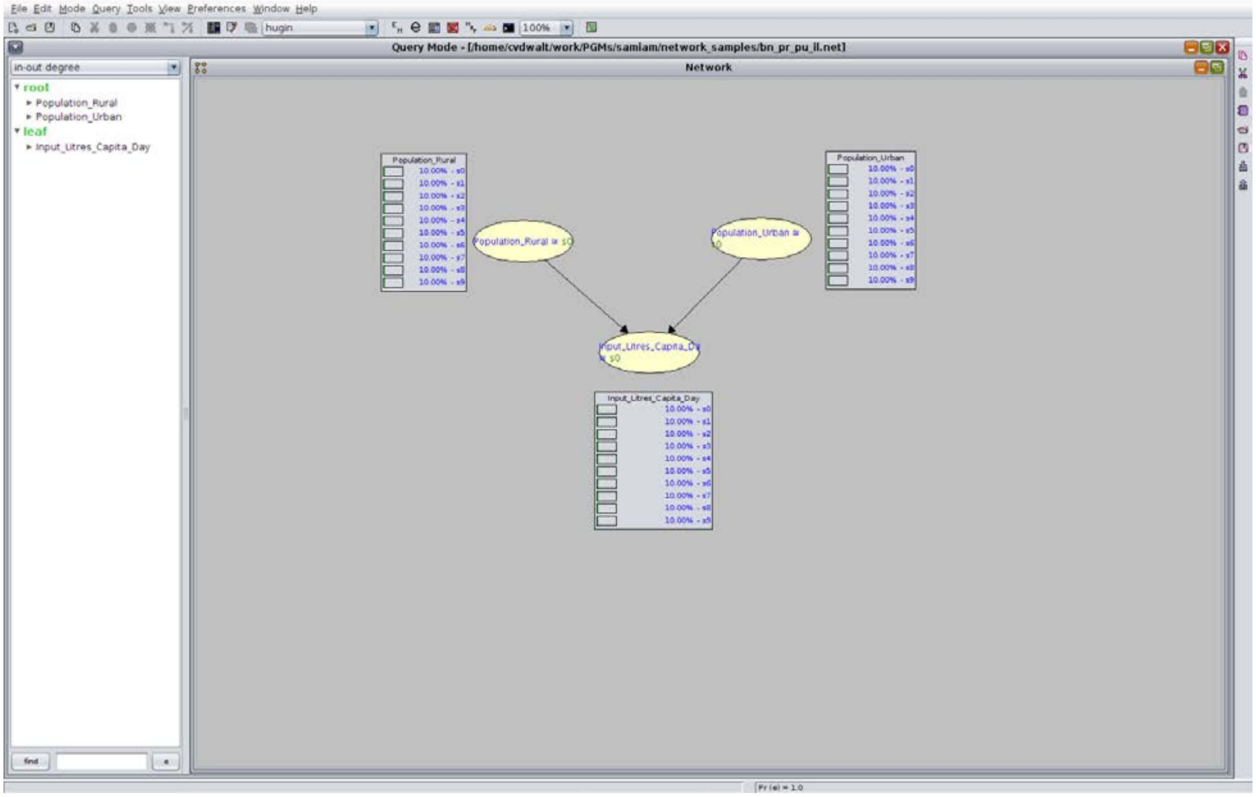


Figure 2. Population rural-population urban-input litres Bayesian network prior to training

The BN makes use of discrete data and thus all variables were binned into ten discrete intervals (equally spaced between the minimum and maximum value of each variable). For the BN illustrated in Figure 1, s0 represents the bin with the lowest value and s9 represents the bin with the highest value. The bin intervals for all variables in the dataset can be observed in Figure 1. We summarise the bin intervals for each variable in the BN in Table 2.

Table 2: Bin intervals of the three variables for the population rural-population urban-input litres Bayesian network

	Population_Rural		*Population_Urban*		*Input_Litres_Capita_day*
Bin	*minimum*	*maximum*	*minimum*	*maximum*	
s0	0	126,827	0	449,341	108.09
s1	126,827	253,653	449,341	898,658	161.77
s2	253,653	380,480	898,658	1,347,974	215.45
s3	380,480	507,306	1,347,974	1,797,291	269.13
s4	507,306	634,133	1,797,291	2,246,608	322.81

s5	634,133	760,960	2,246,608	2,695,925	376.49
s6	760,960	887,786	2,695,925	3,145,242	430.17
s7	887,786	1,014,613	3,145,242	3,594,558	483.85
s8	1,014,613	1,141,439	3,594,558	4,043,875	537.53
s9	1,141,439	1,268,266	4,043,875	4,493,192	591.21

We initialise the Conditional Probability Tables (CPTs) of the Bayesian network with equal prior probabilities. The probabilities for all bins (shown in the monitors next to each variable in Figure 1) in each variable are thus equal prior to training of the network.

After constructing this network, we make use of data from the above mentioned variables for all municipalities in South Africa from 1994-2014 to train the CPTs of the variables in the network. Figure 3 shows the BN and the CPT monitors after training with the Expectation Maximisation (EM) algorithm. The CPTs have thus been updated by the EM algorithm based on the data or evidence that has been provided during training.

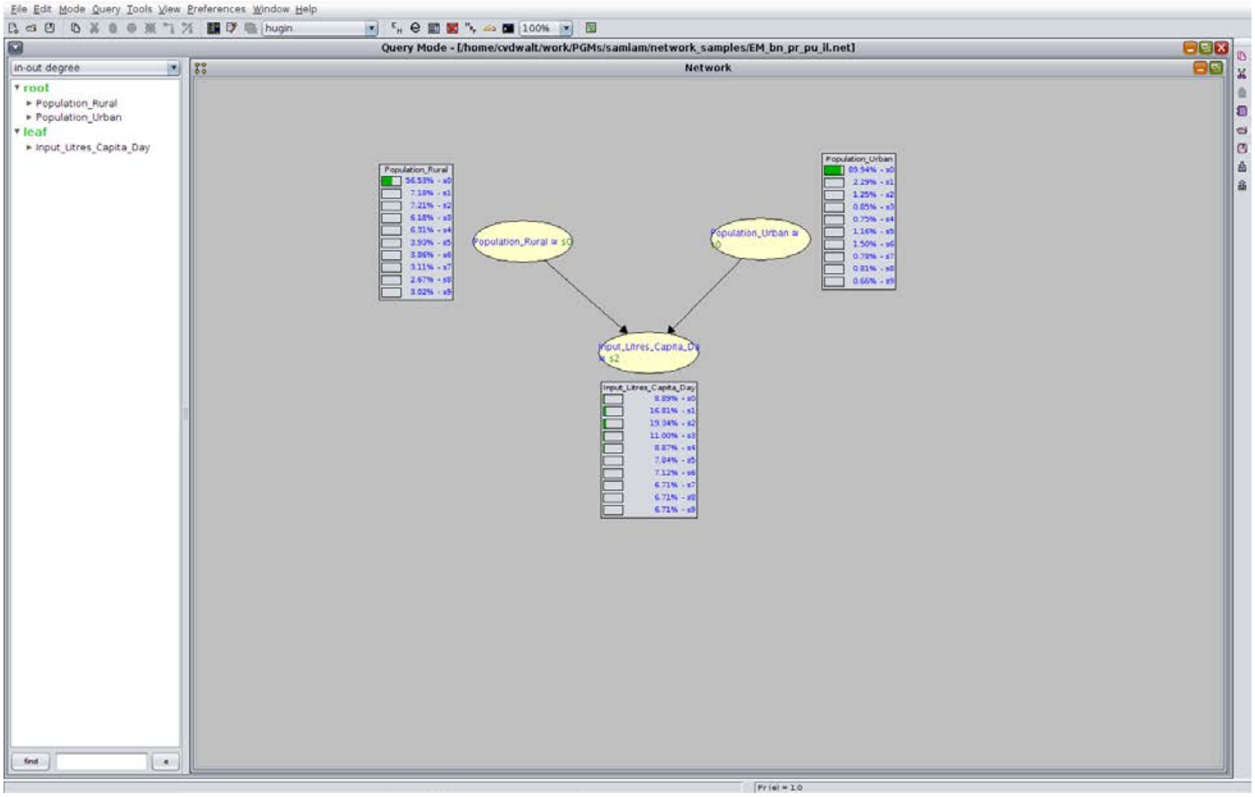


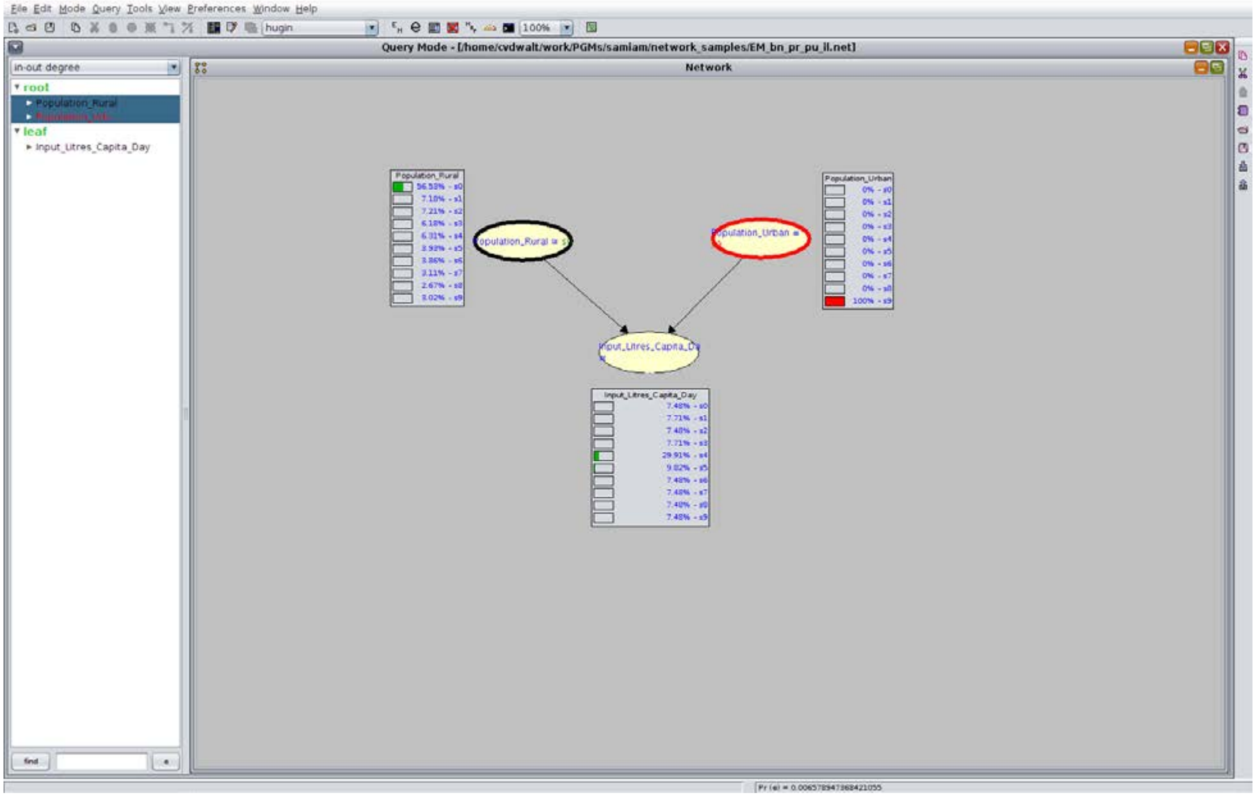
Figure 3. Population rural-population urban-input litres Bayesian network after training with data from 1994-2014.

It can be observed in Figure 3 that 56.53% of the ‘Population_Rural’ values fall in bin s0 [0, 126827] and the remaining bins contain between 7.18% and 3.02% of the values. We also observe that 89.94% of the ‘Population_Urban’ values fall in bin s0 [0, 449341] and the remaining bins contain between 0.66% and 2.29% of the values. These probabilities indicate that generally most municipalities have rural populations in the range [0, 126827] and urban populations in the range [0, 449341].

To illustrate the inference capability (what-if analysis) of this Bayesian network, we can pose the following question to be answered by the network: *Do urban populations require more input water per capita per day than rural populations?*

Intuitively, we would expect that water is more accessible in urban populations than in rural populations, and therefore the water input per capita per day for urban areas should be higher. We can validate or invalidate this intuition and answer the question posed above by explicitly setting the values for ‘Population_Rural’ and ‘Population_Urban’ in the BN.

First, we set ‘Population_Urban’ to s9 (the highest possible value). This setting thus represents the municipalities with the biggest urban populations. The BN then performs probabilistic reasoning and updates all probabilities in the network. Figure 4 illustrates the results of the BN when we set these values as proposed. We define this state as ‘urban’.



Query Mode - [/home/cvdwalt/work/PGMs/samiam/network_samples/EM_bn_pr_pu_il.net]

- root
  - Population_Rural
  - Population_Urban
- leaf
  - Input_Litres_Capita_Day

```
Population_Rural
56.53% - s0
 7.18% - s1
 7.21% - s2
 6.18% - s3
 6.31% - s4
 3.93% - s5
 3.86% - s6
 3.11% - s7
 2.67% - s8
 3.02% - s9
```

```
Population_Urban
  0% - s0
  0% - s1
  0% - s2
  0% - s3
  0% - s4
  0% - s5
  0% - s6
  0% - s7
  0% - s8
100% - s9
```

```
Input_Litres_Capita_Day
 7.48% - s0
 7.71% - s1
 7.48% - s2
 7.71% - s3
29.91% - s4
 9.82% - s5
 7.48% - s6
 7.48% - s7
 7.48% - s8
 7.48% - s9
```

```
Pr (e) = 0.006578947368421055
```

Figure 4. BN where 'Population Urban' is set to the maximum

Next, we set 'Population_Rural' to s9 (the highest possible value). This setting represents the municipalities with the biggest rural populations. Figure 5 illustrates the results of the BN when we set these values as proposed. We define this state as 'rural'.

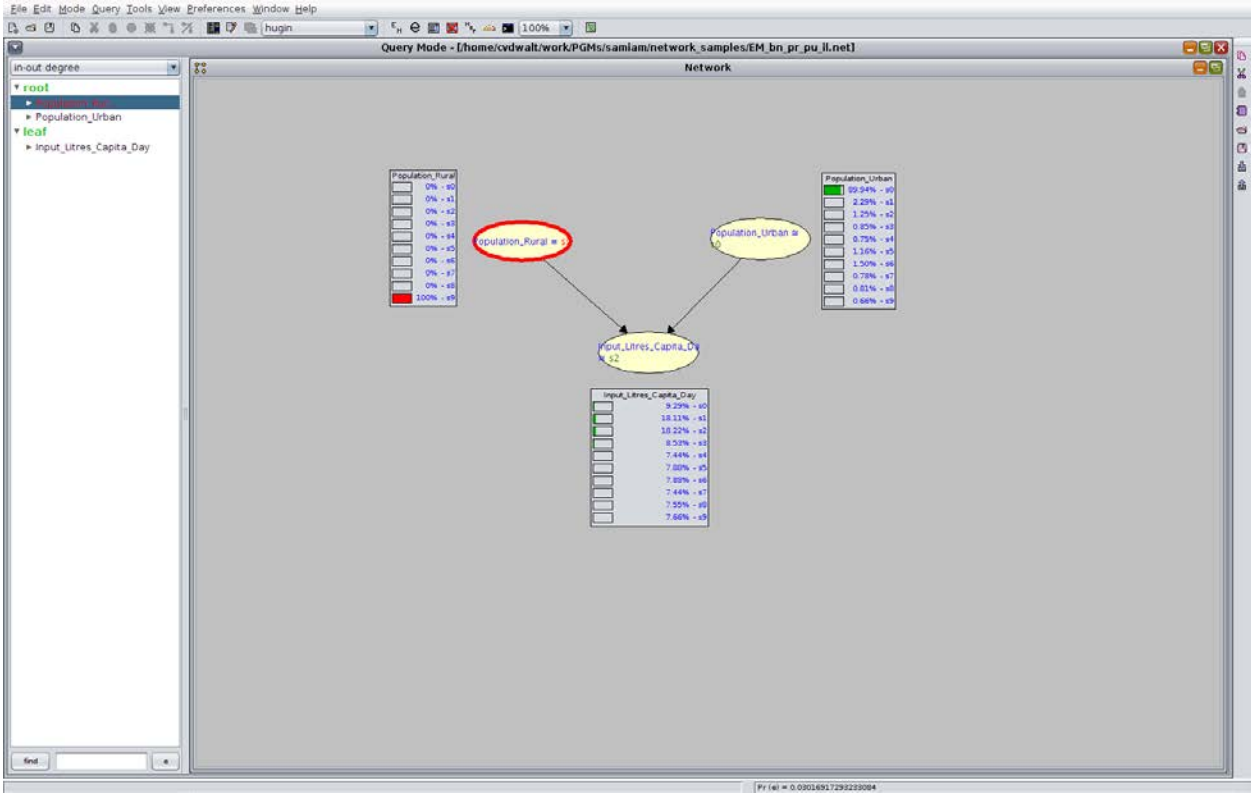
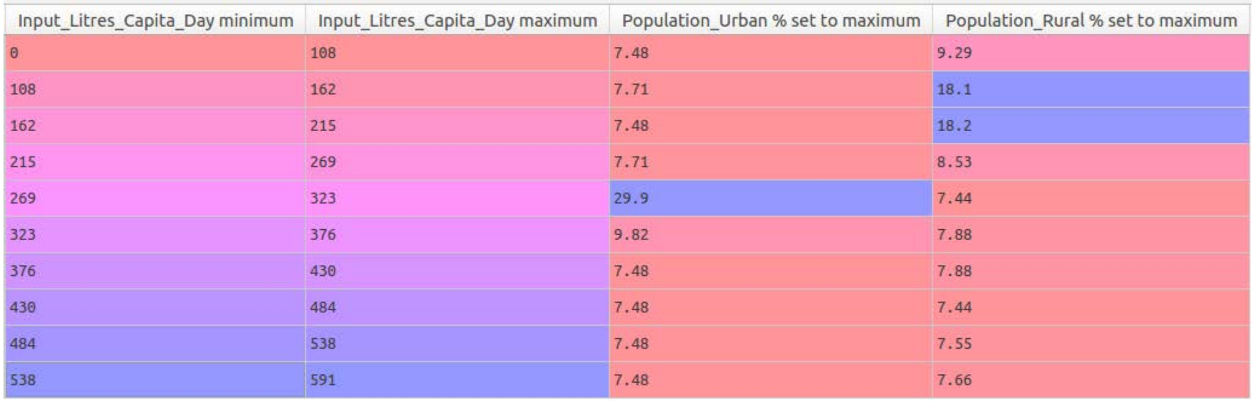


Figure 5. BN where 'Population Rural' is set to the maximum

We summarise the 'Input_Litres_Capita_Day' results of these two settings in Table 3 for comparative purposes.

Table 3. 'Input_Litres_Capita_Day' results of BN when municipality profile is set to urban and rural



Input_Litres_Capita_Day minimum	Input_Litres_Capita_Day maximum	Population_Urban % set to maximum	Population_Rural % set to maximum
0	108	7.48	9.29
108	162	7.71	18.1
162	215	7.48	18.2
215	269	7.71	8.53
269	323	29.9	7.44
323	376	9.82	7.88
376	430	7.48	7.88
430	484	7.48	7.44
484	538	7.48	7.55
538	591	7.48	7.66

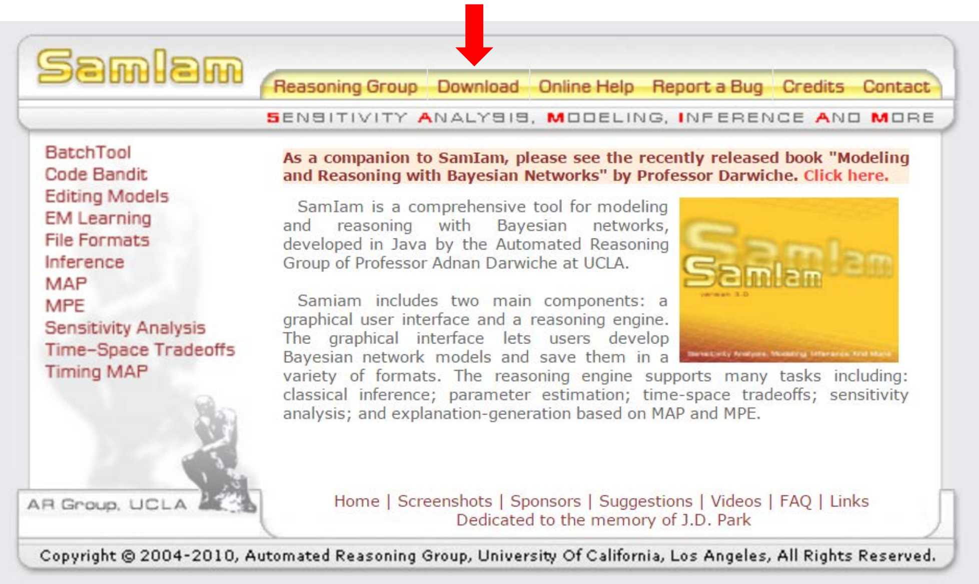
We observe in Table 3 that the percentage of municipalities in 'rural' areas have the highest probabilities for 'Input_Litres_Capita_Day' in the range [0, 215] (bins 0-2), while the percentage of municipalities in 'urban' areas have the highest probabilities for 'Input_Litres_Capita_Day' in the range [269, 376] (bins 4-5). This indicates that municipalities in 'rural' areas typically have lower ‘Input_Litres_Capita_Day’. This result thus validates our intuition that urban populations use more input litres per capita per day than rural areas.

This BN can easily be extended to take into account the interactions and causalities of a more complex model with more variables. The same methodology can also be employed to model more scenarios in any domain where either data or expert knowledge or both are available. BNs can also easily be extended to connect models from e.g. different economic sectors that share common drivers, by simply connecting the models through the common drivers (that are represented by nodes in the network).

For example, energy demand and water demand is known to share common drivers i.e. they share variables that directly or indirectly influence demand in both sectors. E.g. the GDP of a country typically drives energy demand, while the generation of energy with coal requires water. Based on the methodology presented in this document, it will thus be possible to connect a water demand and energy demand model through common drivers since BNs extend themselves naturally to this type of modelling.

4. How to load the Bayesian Network

I. Download and install SamIam – Sensitivity Analysis, Modeling, Inference and More (open source software).



As a companion to SamIam, please see the recently released book "Modeling and Reasoning with Bayesian Networks" by Professor Darwiche. Click here.

SamIam is a comprehensive tool for modeling and reasoning with Bayesian networks, developed in Java by the Automated Reasoning Group of Professor Adnan Darwiche at UCLA.

Samiam includes two main components: a graphical user interface and a reasoning engine. The graphical interface lets users develop Bayesian network models and save them in a variety of formats. The reasoning engine supports many tasks including: classical inference; parameter estimation; time-space tradeoffs; sensitivity analysis; and explanation-generation based on MAP and MPE.

Dedicated to the memory of J.D. Park

Copyright © 2004-2010, Automated Reasoning Group, University Of California, Los Angeles, All Rights Reserved.

II. Open SamIam and from “File” on the menu bar, open the file EM_bn_pr_pu_il.net in the network/bn_files folder to select the trained model.

III. Click in the menu bar “Mode” -> “Query Mode”.

IV. Click in the menu bar “Query” -> “Show monitors” -> “Show All”.

5. Forecasting model technical requirements

The technical requirements, as defined in the proposal, refer to the description of the forecasting model, including the user interface, visualisation, data requirements and outputs. The technical requirements must be aligned to the Operating Model description. An initial list of output requirements is listed below:

I. Demand
    i. Breakdown of required skills needed
    ii. Numbers of individuals with required skills needed
II. Supply
    i. Breakdown of required skills available
    ii. Numbers of individuals with required skills available
III. Scenario analysis - how the demand and supply shifts with changes in predetermined drivers e.g. population, GDP, plant type, plant age, etc.

6. Conclusion

We have illustrated the methodology of applying BNs to water demand modelling in this report. More specifically, we have proved that raw data obtained from the DWA can be processed and used to train a BN that is capable of answering a specific question by performing scenario analysis. Using this methodology it will be possible to model more scenarios in any domain where either data or expert knowledge is available. By expanding this network of three variables to more variables more complex scenarios in water demand can be modelled.