To appear in Proceedings of the 82nd Meeting of the Transportation Research Board, Washington, D.C., ©2002, UTA

Learning a Causal Model from Household Survey Data Using a Bayesian Belief Network

Session Title: Sensor Fusion and Information Integration Committee: A5008 on Artificial Intelligence

Francisco J. Torres
Senior Transportation Planner
North Central Texas Council of Governments
Arlington, TX 76005-5888
Phone: (817) 277-1655
Fax: (817) 640-3028
Email: ftorresv@dfwinfo.com

Manfred Huber
Assistant Professor
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019-0015
Phone: (817) 272-2345
Fax: (817) 272-3784
Email: huber@cse.uta.edu

Number of words = 4,977
Number of tables and figures =10
Total =7,477

ABSTRACT

A Bayesian Belief Network (BBN) is a modeling and knowledge representation structure used in Artificial Intelligence (AI) that consists of a graphical model depicting probabilistic relationships among variables of interest. This graphical model is a valuable tool to represent the causal relationships that exist in a given set of variables. Because the number of possible BBNs for a given dataset is exponential with respect to the number of variables, learning a BBN from data is a difficult and resource-consuming task. This paper presents the implementation of a greedy algorithm that automatically constructs a BBN from a dataset of cases obtained from a household survey. The resulting BBN shows the dependencies that exist among key variables that are associated to the trip generation process.

INTRODUCTION

A Bayesian Belief Network (BBN) is a directed acyclic graph that is used to represent the relation among a predefined group of variables. In this graph, nodes represent the variables, and the dependencies between variables are depicted as directional links corresponding to conditional probabilities. The construction of a BBN consists of determining the structure, as parent-child link between nodes, as well as the probability distribution associated to these relations. A detailed description of BBNs can be found in (*1*) and (*2*).

According to (*2*) BBNs, when used in conjunction with statistical techniques, have the following advantages for data analysis:

1. They handle situations where some data entries are missing as a result of knowing the dependencies among variables,
2. They help to gain understanding about a problem domain once the causal relationships are learned,
3. They represent an ideal method for combining prior knowledge and data,
4. They can prevent the overfitting of data.

Traditionally, BBNs were constructed from knowledge of human experts. However, during the last decade several methods had been developed to build them directly from databases. Finding the BBN that best reflects the dependency among the variables of a given database is a difficult and resource-consuming task because it involves searching in an exponential number of possible structures.

Figure 1 is an example of a simple BBN that describes the dependencies among variables associated with the presence of traffic congestion on roads. This figure includes all elements that compose a BBN: nodes, directional links, and probability distributions

This paper presents the implementation of a very popular algorithm, referred to as K2, which was developed by Cooper and Herskovitz (*3*). For the purposes of this paper the K2 algorithm was implemented in a database manager environment. The performance and accuracy of this implementation was tested and verified using synthetic data created from a known BBN. In addition, the algorithm was applied to data from the 1996 Dallas-Fort Worth (DFW) Household Activity Survey containing very reliable data from 3,624 households. This dataset contains numerous variables that describe the characteristics of the households, their dwellers and vehicles, as well as the patterns of the trips made by the occupants. From these variables, a subset of the most commonly used in the transportation planning discipline was selected. This subset was defined as the input of the K2 algorithm. The final product of this implementation is a BBN that describes the causal relationship of all variables associated with trip generation. For this purpose, the trips produced by a household were classified as: Home-based work (HBW), Home Non-work (HNW), and Non-home-based (NHB). Each trip type was considered as a node in the BBN.

The following section presents a background about learning a BBN from data. Subsequently, the K2 algorithm is described. Then the implementation of this algorithm is explained and the results obtained given the input of household survey data are shown. Finally, a description of improvements and extensions of the implementation is presented.

BACKGROUND

Determining the BBN that optimally reflects the dependencies between all variables in a data file is of extreme difficulty because the number of available structures is exponential with respect to the number of variables or nodes. According to (*4*) the number of possible BBN structures is given by the following recursive function:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} \; 2^{i(n-i)} f(n-i) \quad \textbf{(1)}$$

,where n is the number of nodes in the BBN

Based on this formula, there are 3 possible structures for n=2, 25 possible structures for n=3, 29,000 for n=5 and approximately 4.2×10^{18} for n=10. Searching all possible structures for the optimal network becomes intractable even for a small number of nodes.

Cooper and Herskovitz (*3*) developed, in 1992, an algorithm that produces a BBN from a data file of cases as its only input. This algorithm was named K2, and has been a major point of reference for all work related to constructing BBNs that have been developed afterwards. K2 finds the optimal structure through a greedy search of a reduced space of possible networks. The greedy criterion is based on a scoring function that represents the probability of a structure given the data.

One of the main objectives of the work presented in this document is to implement a process that efficiently searches for an optimal BBN that describes trip generation based on demographic and infrastructure data. Within this application domain, one of the few documents that deal with constructing a BBN from household data is an article by Ramoni and Sebastiani (*5*) related to the British Household Survey. These authors considered 13 variables in their analysis, all of them related to the characteristics of the households and their occupants. However, no data related to trips generated by the household members was considered in their work. In contrast, the research described in this paper includes variables more closely related to the trip generation process.

K2 ALGORITHM

The K2 algorithm searches, given a data set D, for the BBN structure B_S* that maximizes the probability of the structure given the data, P (B_S | D). This probability can be calculated according to the following theorem, which is proven in (*3*)

Theorem. Let Z be a set of n discrete variables (nodes) , where a variable X_i in Z has r_i possible value assignments: (v_{i1}, …., $V_{i,ri}$). Let D be a database of m cases, where each case contains a value assignment for each variable in Z. Let B_s denote a belief network structure containing just the variables in Z, and B_p the conditional probabilities. Each variable X_i in B_S has a set of parents, which are represented with a list of variables Π_i. Let w_{ij} denote the jth unique instantiation of Π_i relative to D. Suppose there are q_i such unique instantiations of Π_i. Define N_{ijk} to be the number of cases in D in which variable X_i has the value v_{ik} and Π_i is instantiated as w_{ij}. Let $N_{ij} = \sum_{k=1}^{r} N_{ijk}$. If given a BN model, the cases occur independently and the density function f(Bp|Bs) is uniform then it follows that

$$P(B_S \mid D) = P(B_S) \prod_{i=1}^{q_i} \prod_{j=1}^{n} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad \textbf{(2)}$$

The K2 algorithm determines the optimal structure through a greedy procedure that identifies if a node can increase the network probability by adding a new parent to it, given the prevailing structure. This process is performed sequentially for all nodes in the network given a specific order. At each step, it is only necessary to calculate the comparative probability of adding a parent to a node with respect to the probabilities associated to all potential parents. Based on this, the incremental probability of adding a new parent, given the existing parents π_i of a node i , is defined as the function $g(i,\pi_i)$, which is calculated, based on equation 2, as:

$$g(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \quad \text{.......... (3)}$$

The basic structure of the K2 algorithm is described with the following pseudocode:

```
for i = 1 to n
    π_i = { }
    gn[ i ] = g( i , null )
    finish = false
    while not finish and | π_i | < u
        gn_new = - ∞
        for j = 1 to |pred|
            if  g(pred[ j ] , π_i ) > gn[ i ]  then
                gn_new = g(pred[ j ] , π_i )
                z = j
            if gn_new  > gn [ i ] then
                gn [ i ] = gn_new
                π_i = π_i  U { z }
            else  finish = true
```

Where

n	Number of nodes in the network
π_i	Array of parents of node i
u	Maximum allowable parents that any node can have
gn	Array that stores the maximum values of g() associated each node
pred	Array of predecessor nodes to each node i
z	Prospective parent with the highest probability
g()	Function calculated as indicated by eq.3

The K2 algorithm makes the following assumptions

1. All variables are discrete
2. There are no missing values for any variable
3. The cases are independent
4. Prior probabilities over hypotheses are uniform
5. The parent selection process is based on a fixed ordering of the nodes so that only the predecessors of a node can be selected as parents.
6. The frequencies of the variables, *N*, needed to compute the probabilities are integers.

The computation time of this algorithm is $2^{n(n-2)/2}$. Its output consists of list of parents for each node in the network.

The flowchart for the K2 algorithm is shown in FIGURE 2.

IMPLEMENTATION

For the purpose of the implementation it was observed that, when working with files of more than 1000 cases, the probability values calculated in Equation 3 were negligible most of the times. Due to this fact and considering that the K2 algorithm uses this value only to make comparisons, it was decided to use the natural logarithm of g (i,π_i) instead. This modification also resulted in less computation time because all multiplications and divisions were transformed into additions and subtractions.

The K2 algorithm was implemented inside a Relational Database (*Microsoft® Access 2000*) environment where all the necessary calculations are performed through the execution of SQL commands. This environment was selected because most of the calculations involve summations of frequencies for all variables under different variable grouping combinations, which adapt easily to the following SQL command.

$$SELECT\ SUM\ (ln(N_i))\ GROUP\ BY\ X_i,\ldots.X_n$$

The K2 algorithm is part of an executable program that creates several tables that are needed for the calculation of the BBN. The only input for this program is a file of cases. The file can be in any of the following formats: worksheet (Excel), ASCII, or DBF.

The first column of the input file corresponds to a unique id that represents a case number. The remaining columns are the variables that are represented as nodes in the BBN.

The input file can have any name but inside the database is named ***Cases***. This table is used to create a second table, called ***Nodes*** that stores the basic information associated with each node. A sequential node number is assigned to each variable based on the column order observed in the ***Cases*** table. The maximum and minimum occurrences of each variable are stored in columns or fields of the ***Nodes*** table.

A table named ***Complete*** is created to store the unique combinations and their frequencies. Because some combinations are presented more than once in the ***Cases*** table, the number of records in this table is greater than that of the ***Complete*** table. This implies that all frequencies associated with these multiple combinations are integer values greater than one.

The data that defines the structure of the BNN is calculated at the end of the program execution. This data is stored in the ***Parents*** table, which contains, among others, two columns that are associated to the nodes; the first one indicates a given node number in the network and the second stores the node number of the parent found for the node indicated in the first column. Each node has as many records in this table as parents in the BBN. The other columns of this table store values related to the probability associated with the inclusion of the parents of each node.

Due to the characteristics of the computational environment, it was possible to override the restriction of the K2 algorithm related to the maximum number of parents, represented by u, that a node could have.

HOUSEHOLD SURVEY RESULTS

The sections of this chapter present a description of the results obtained from the K2 algorithm for input files with different characteristics. First, the ability of the algorithm to produce accurate BBNs was tested using synthetically generated data. Then a BBN was constructed based on the data from the DFW household survey.

For the analysis, the complete list of variables considered is presented below.

1. <u>County</u>. Demarcation where the household is located.(1=Collin; 2=Dallas; 3=Denton; 4=Tarrant; 5=Other)
2. <u>Dwelling Unit.</u> Single or multiple.
3. <u>Size</u>. The number of persons who permanently live in a household. A range of 1 to 6 is considered for this variable.
4. <u>Adults</u>. Number of persons whose age is greater or equal to 18 years. A range of 0 to 6 adults is associated with this variable.
5. <u>Workers</u>. Number of the members of the household who develop an activity outside of it and who are either employed or perform regular volunteer work. At most, all persons in the household who are older than 16 years might be considered as workers. The range for this variable is exactly the same as for Adults.
6. <u>Income.</u> This variable represents the income quartile group to which the household belongs. The following four values are considered: 1= low; 2= low medium; 3 = medium high; 4 = high.
7. <u>Vehicles.</u> Number of vehicles that are available for the exclusive use of the members of the household, within a range of 0 to 6.

8. Transit. This binary variable indicates if there is transit available, within walking distance, for the household occupants. This is true if the household is located within half a mile of a transit stop, whether it is bus, light rail or commuter train; in this case, the variable is assigned a value of 1, otherwise a zero is stored.
9. HBW Trips. Number of trips that are made by all members of the household during the 24 hours of a typical working day, between home and a work related activity. A range of integer values from 0 to 15 is considered.
10. HNW Trips. Trips made between home and a non-work related activity, such as shopping, school, etc. The range is similar to HBW trips.
11. NHB Trips. Trips for which none of the end points is the home. The range is similar to HBW trips.

In cases for which the maximum value of an observed variable was greater than the established range, the observed value was truncated to that of the range upper limit. For example, a value of 7 vehicles was set equal to 6 in the input file. Considering this, the values of the input file were still equal to the observed data in 99% of the cases.

Input File from a Known Structure

In the first test the K2 algorithm was applied to a file containing synthetic data that was created based on a predefined BBN structure. For this purpose, only the following variables were considered: size, adults, workers, income, vehicles, transit, and trips (equal to the summation of HBW, HNW, and NHB).

The input file used for the tests related to this task contained 1000 records for which the values of the variables were randomly generated based on predefined probability distributions and the structure presented in FIGURE 3.

In order to test the reliability of the estimations obtained, the data set with 1000 cases was used to generate 20 random subsets of 500 records that where subsequently used to construct BBNs. In all 20 cases, the data subsets generated the original BBN structure, indicating the robustness of the algorithm. However, there was a small difference in the maximum values of the *g* function calculated for each node. The coefficient of variation of the incremental probability values, defined as the standard deviation divided by the mean, was 2.9%. TABLE 1 presents a sample of a Parents table obtained from this K2 implementation.

File of Cases from a Household Survey

The input file with the data collected from the 1996 Household Survey for the DFW area has 3,634 records. TABLE 2 presents, as an example, some of the data records used for this analysis. A process of splitting the file into two sets, similar to that described in the previous section, was implemented. In all tests there was no significant variation in the characteristics of the BBN obtained. In fact, the structure remained basically unchanged in all tests that were performed. The resulting BBN is presented in FIGURE 4. The thickness of the links in this figure represents the strength of the dependencies among the corresponding variables.

An interesting result from the analysis performed by the K2 algorithm is that the trips produced by a household in the DFW area are not directly caused by accessibility variables such as vehicles or transit available. This could imply that people do not make trips because they have means available but rather their trips are motivated by the need to perform basic activities, such as going to work or to school, for example.

Regarding accessibility measures, Transit was related only to the County and the Dwelling Unit type. The relation is explained because, in 1996, the DFW transit network served only four counties, and the supply was concentrated in Dallas and Tarrant counties. In this region, as in all metropolitan areas, the transit service tends to favor corridors where population density is higher and multiple dwelling units, such as apartment complexes, are located. The number of vehicles available in a household is related not only to Dwelling Unit type as in the case of Transit, but also to Adults. On the other hand, the BNN indicates that the auto ownership does not depend on the household income. This finding can be supported by the fact that there is not transit service in many cities of the DFW area and riding a car is the only motorized mode of transportation available. Another interesting fact derived from the constructed BBN is that Vehicles and Transit are not related.

A variable that is commonly used in many trip generation models is Income. In the BBN, Income is closely related to the Dwelling Unit, Adults, and Workers, but it does not appear as a cause or motivation for the generation of trips. Although it is a logical conclusion that HBW trips are closely related to Workers, it is remarkable that in the BBN this is the only variable related to HBW trips. It was expected that the BBN would show a relation with other variables such as Income, Size, or Vehicles, as is the case in traditional trip generation models. HNW trips are related to HBW and also to the variable Size, which represents the number of people living in the household. Finally, according to the BBN, NHB trips are only related to the HNW trips.

An example of a resulting probability distribution table for Workers is presented in TABLE 3.

EXPERIMENTS

To demonstrate the capabilities of the BBN as a tool to identify relationships among a set of variables, it was compared with the results of a multiple linear regression analysis using the information of all households (3,624). A separate regression analysis was performed for each of the three trip types considered: HBW, HNW and NHB. In all three cases, the following variables were considered as independent variables: Size, Adults, Workers, Income, Vehicles, Transit, County, and Dwelling. For HNW, HBW was also considered as an independent variable, while both variables were added to the set of independent variables used to predict the linear regression model for NHB.

Variable Relations

Considering a 95% confidence interval the only significant coefficient for HBW was the one associated with Workers, with a t-statistic value of 41.61. For HNW, the significant coefficients and their corresponding t-statistic values (in parenthesis) are: Size (45.2), HBW (-8.92), Vehicles (7.79), Workers (- 5.71), Adults (-5.24), and Income (2.26). Finally the significant coefficients for NHB are those associated with: HNW (19.52), Workers (10.22), Size (-4.42), Income (4.42), HBW (-4.1), and Vehicles (2.32). The relationships among variables indicated by these results are compatible with the ones represented by the links in the BBN constructed from the same data: Workers->HBW, Size->HNW, HBW->HNW, and HNW->NHB. It can be seen that only the relations with the most significant values appear in the BBN.

Prediction Capability of the BBN

An effort was made to test the predictive capability of the BBN. For this task, three different techniques were used:

Maximum Likelihood (ML), for which the predicted value is defined as the one with the highest probability.
Expected Value (EV), calculated as the summation of the products of each possible value and its probability.
Rounded EV (REV), determined based on the EV according to the following rule: if the fractional part of the EV is less than 0.5 then the REV is equal to the integer part of EV, otherwise REV is equal to the integer of EV plus one.

In order to obtain an unbiased result, a BBN was constructed based on a subset of the households. This subset was defined through the random selection of half of the household records. The predictive capability of the resulting BBN was then tested using the remaining half of the records.

TABLE 4 shows the correlation and coefficient of determination values that were obtained by comparing the actual values with those estimated by the BBN.

For all techniques the correlation and determination values are higher for HBW and HNW than for NHB. On the other side, it is evident that the EV and REV methods provide better estimations than ML. In fact, the R2 values for the former methods are similar to those obtained from the linear regression model (47.94%, 46.58%, and 18.77%, for HBW, NNW, and NHB, respectively).

The capacity of the BBN to predict the actual value of variables in an individual record was also tested. Since the actual values are integers and the EV technique estimates non-integers, only the REV and the MD techniques were considered for these comparisons. The results are presented in TABLE 5

In all cases the prediction rate of the BBN is significantly higher than for randomly imputed values. In this case, ML results in better predictions than REV, illustrating the advantage of ML to impute variables in individual records and of EV to predict the population average of a variable.

MODIFICATIONS AND ENHANCEMENTS PROPOSED TO THE K2 ALGORITHM

Some modifications can be introduced in order to enhance the basic implementation of the K2 algorithm and to overcome some of the restrictions introduced by its assumptions. At this time, the authors of this paper are working on the improvement of the K2 algorithm. The improvements that are considered are briefly discussed below.

Missing Data

Traditionally, according to (*6*), the problem of dealing with incomplete data in databases has been addressed with some of the following measures:

- Ignore variables with missing data.
- Remove records with missing data.
- Re-Weigh the data.
- Impute missing data.

Since, as noted before, BBNs explicitly identify the relationship between all variables in a dataset, they can be used to deal with the problem of missing data. As part of the research being conducted by the authors, a method consisting of an iterative implementation of the K2 algorithm is being tested. For this implementation, the probabilities associated with the incomplete records are calculated based on the structure defined by the algorithm in a previous iteration. The missing items are then imputed based on the probabilities of the incomplete records using the maximum likelihood method. This process can be executed repeatedly until the explanatory power of the estimated BNN no longer improves. Preliminary experiments with this algorithm and the synthetically generated data where 5% of the data records had data items randomly removed, have demonstrated that this algorithm is capable of inferring missing data records with a precision of 40%. However, additional experiments have to be performed to evaluate the precision and computational overhead of this algorithm. **FIGURE 5** shows a diagram of the proposed method to deal with missing data using a BBN.

Ordering of Variables

As stated in one of the assumptions of the K2 algorithm, an ordering of the nodes has to be established to define the structure of a BBN. In many situations this restriction is implicitly enforced by the nature of the problem that has to be solved. This often applies when the causal relations among the variables are known. An example of this is to consider that the number of its occupants causes the number of trips produced by a household. Another situation for which ordering is important is when there is a need to predict a parameter as a function of many independent variables. In this case, the node that represents the dependent variable must be at the end of the ordering sequence.

However, for many problems no information about the ordering of the nodes is available. In this case, it would be extremely valuable if the algorithm could also provide a greater insight into how all variables are related without obeying any predefined order. On this regard, the use of a Genetic Algorithm (GA) to determine the optimal node ordering that yields the maximum $P(B_S|D)$ is the approach that is currently investigated .

CONCLUSIONS

BBNs represent a very valuable AI tool that can be used to determine the existing dependencies among a group of variables. However, the benefits of this graphical model have not yet been divulged to many areas, such as Transportation Planning, that could benefit from the capability of BBNs to establish causal relations among several independent variables. Currently, linear regression is the technique mostly used by professionals in the field of Transportation Planning.

In this paper, an algorithm, referred to as K2, was implemented to construct BBNs from data. This implementation was performed inside a relational database environment. The input for the K2 algorithm is a dataset of records with different values assigned to the variables considered. The output consists of a description of the causal relations among all variables. The results of the K2 algorithm were tested and verified using synthetic data generated from predefined BBNs .The execution time of the implementation is very reasonable, at least for the size of the datasets and the number of variables considered in this paper.

Besides describing the implementation of the K2 algorithm, this paper proposes extensions aimed at increasing the capabilities and versatility of this computation tool. In particular, a proposed method to deal with the problem of incomplete datasets is briefly described and the implementation of a GA for the purpose of further automating the process of BBN construction is proposed. The authors believe that these improvements will produce a valuable tool that can be used to impute missing data in incomplete datasets.

Finally, this document presents an interesting application of a BBN related to the trip generation process. For this purpose, data from the 1996 Dallas-Fort Worth Household Activity Survey was used. A dataset of 3624 household records was used and 11 variables were considered. The results of an analysis of the BBN constructed from this data indicate that neither accessibility variables, such as transit or vehicles, nor income, have causal links with the trip generation variables. Number of workers and occupants are the variables most strongly related to trip generation.

Overall, the results presented in this paper show that BBNs provide a powerful alternative to linear regression techniques when used as a prediction tool or to discover relations among variables. Furthermore, the experiments indicate the usefulness of BBNs in other, important applications in the Transportation Planning field, including the imputation of missing values in survey records and the generation of sample population records for use in micro simulation studies.

ACKNOWLEDGEMENTS

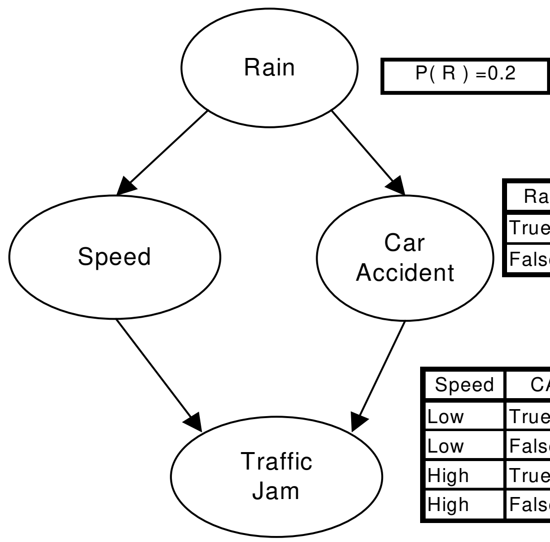
The authors thank the Transportation Department of the North Central Texas Council of Government for allowing them to use the data collected from the 1996 Dallas-Fort Worth Household Activity Survey.

REFERENCES

1. S. Russel and P. Norvig (1995) Artificial Intelligence. A Modern Approach. Prentice Hall.
2. D.Heckerman (1996) A Tutorial on Learning with Bayesian Networks. Technical Report , Microsoft Research MSR-TR-95-06.
3. G. F. Cooper and E. Herskovitz (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning, Vol. 9, pp. 309-347.
4. R.W. Robinson (1977). Counting unlabeled acyclic graphs. In C.H.C. Little (Ed.) Lecture notes in mathematics, 622: Combinatorial mathematics.
5. M. Ramoni and P. Sebastiani. On the Use of Bayesian Networks to Analyze Survey Data. Research in Official Statistics 4(1):53-64, 2001
6. A.J. Richardson. TRB Transportation Research Circular E-C008. Transport Surveys: Raising the Standard.

List of Tables

List of Figures



	Prob(CA|R)	
Rain	S=Low	S=High
True	0.7	0.3
False	0.2	0.8

	Prob(CA|R)	
Rain	CA=True	CA=False
True	0.2	0.8
False	0.1	0.9

		Prob(TJ| S, CA)	
Speed	CA	TJ=True	TJ=False
Low	True	0.8	0.2
Low	False	0.6	0.4
High	True	0.3	0.7
High	False	0.1	0.9

FIGURE 1 Example of a BBN to predict the probability of a traffic jam.

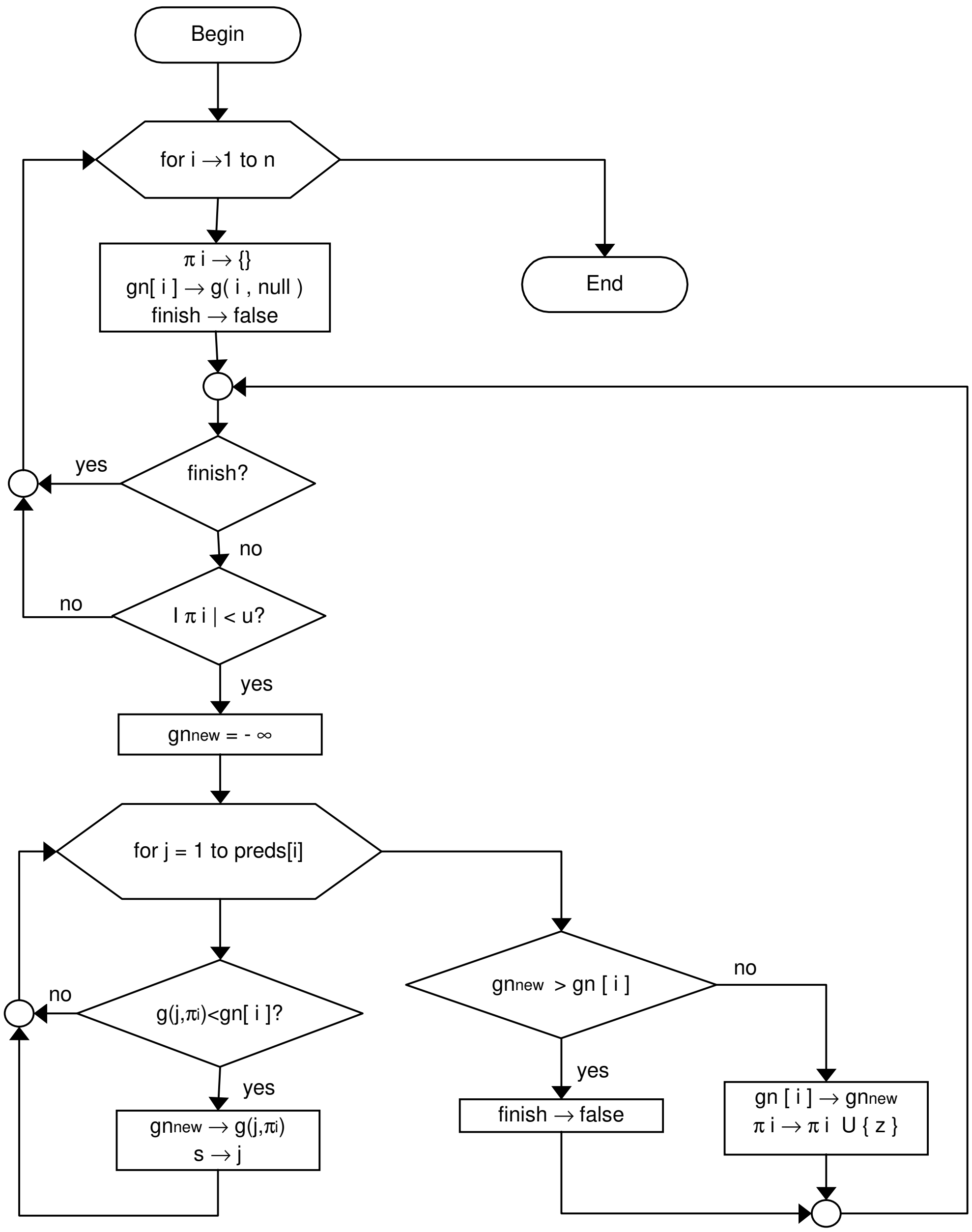


FIGURE 2 .Flowchart for the K2 Algorithm.

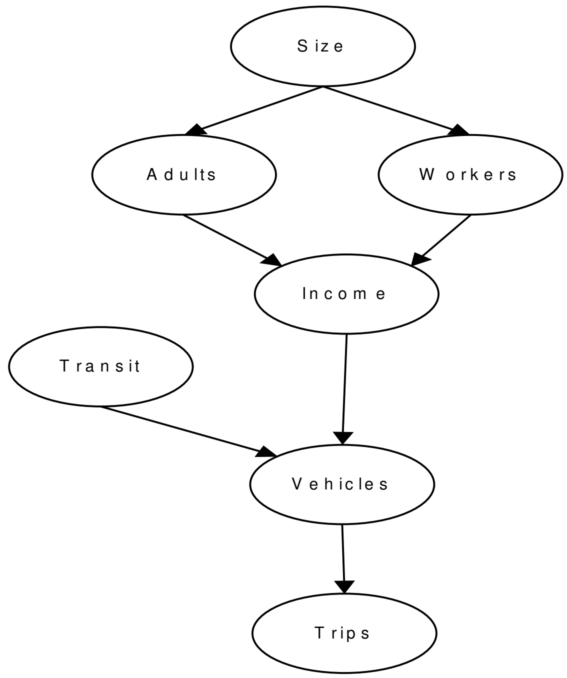


FIGURE 3. Predefined BBN structure used to create the synthetic input file.

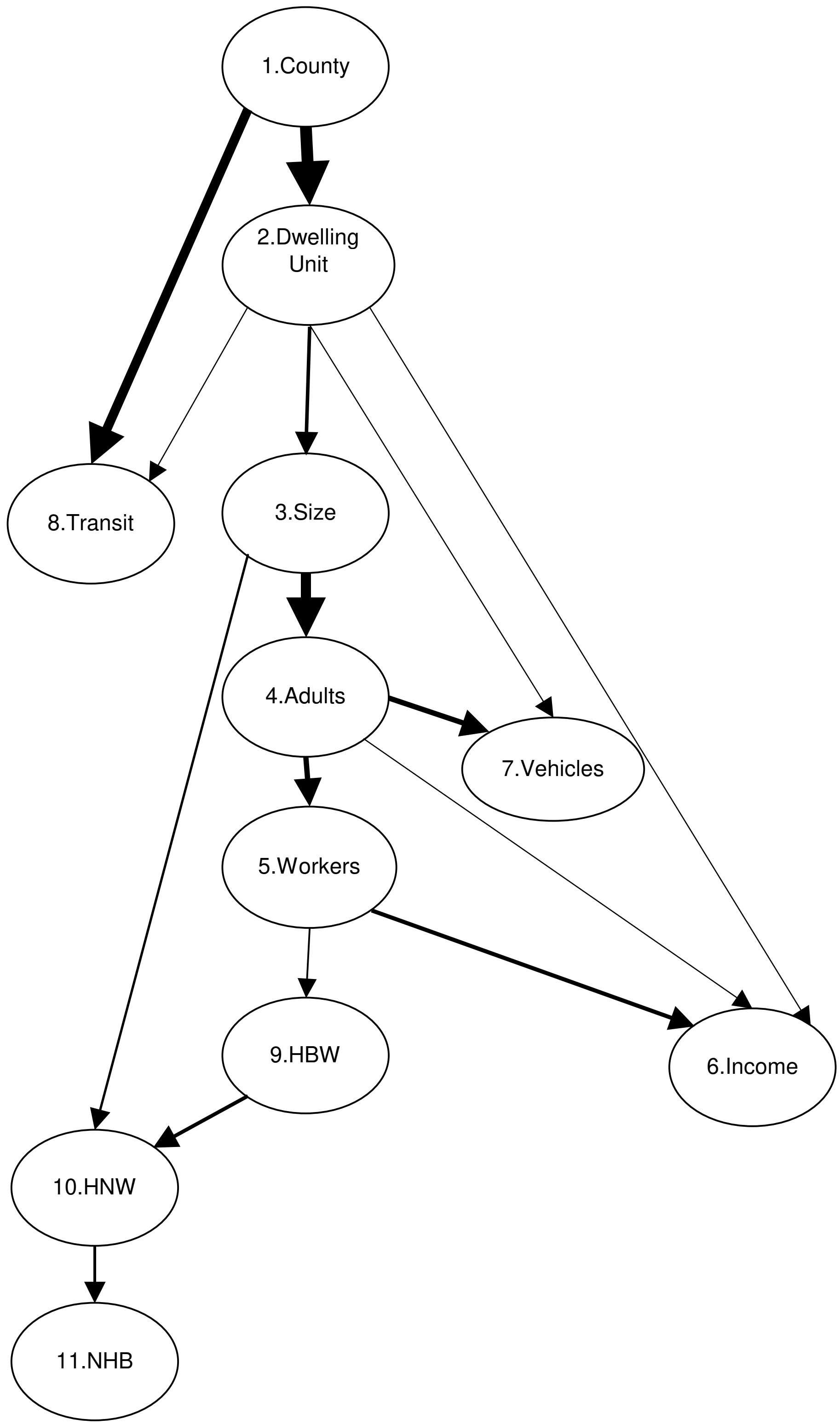


FIGURE 4. BNN constructed based on the DFW Household Survey data.

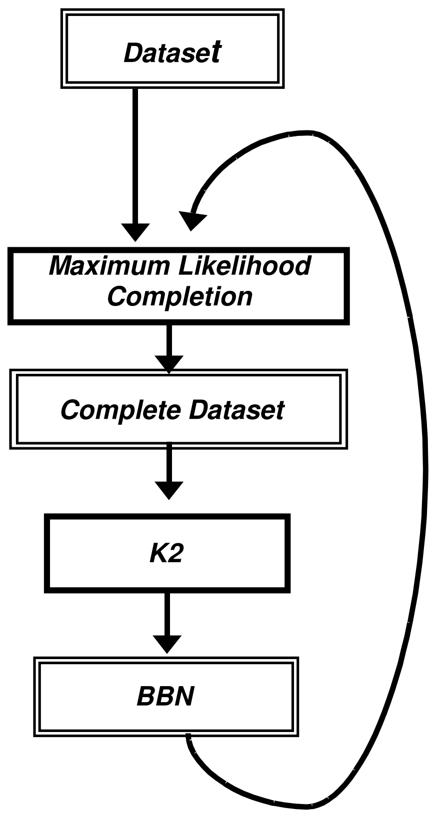


FIGURE 5 Using a BBN to impute data for incomplete datasets.

TABLE 1 Example of the Parents Table Created by the K2 Algorithm.

Node	Description	Node number of added parent	Parents	ln(g(i))	Improved?	Add Parent?	Time
1	Size	0		-1803.032758	Yes	No	3:32:55 PM
2	Adults	0		-1931.04739	Yes	No	3:32:55 PM
		1	Size,	-1186.322726	Yes	Yes	3:32:55 PM
3	Workers	0		-1943.192595	Yes	No	3:32:55 PM
		1	Size,	-1149.483261	Yes	Yes	3:32:56 PM
4	Income	0		-1390.496877	Yes	No	3:32:55 PM
		1	Workers,Adults,	-1380.519495	No	No	3:32:56 PM
		3	Workers,	-1355.593035	Yes	Yes	3:32:56 PM
5	Transit	0		-696.3680243	Yes	No	3:32:55 PM
6	Vehicles	0		-1919.334325	Yes	No	3:32:55 PM
		4	Income,	-1878.14517	Yes	Yes	3:32:56 PM
		5	Income,Transit,	-1893.63324	Yes	Yes	3:32:56 PM
7	Trips	0		-2683.690502	Yes	No	3:32:55 PM
		6	Vehicles,	-2421.76557	Yes	Yes	3:32:56 PM

TABLE 2 Sample Records of the Trip Generation Dataset Used to Construct a BBN.

ID	County	Dwelling	Size	Adults	Workers	Income	Vehicles	Transit	HBW	HNW	NHB
5	2	1	2	2	2	3	2	1	4	0	0
8	2	2	1	1	1	1	1	0	2	0	0
11	2	2	3	2	0	1	0	1	0	9	3
14	2	1	2	2	0	4	2	1	0	4	0
17	2	2	1	1	1	2	1	1	1	1	3
18	2	1	4	2	2	4	2	0	4	9	6
19	2	1	1	1	1	2	1	1	2	0	0
20	2	1	4	2	1	4	2	1	1	11	7
21	2	1	3	2	2	3	2	1	4	2	3
22	3	1	2	2	0	3	2	0	0	6	1
23	5	1	2	2	2	4	4	0	3	1	5
24	2	1	3	3	1	4	3	1	3	7	5
25	1	1	3	3	2	4	2	0	4	2	2
28	1	1	2	2	0	1	1	0	0	0	0
29	2	1	4	2	1	2	2	0	2	8	0
32	2	1	4	2	2	4	2	1	3	0	0
35	2	2	1	1	0	1	1	1	0	1	2
37	2	1	2	2	1	2	2	1	2	5	2
38	2	1	2	2	2	3	3	1	3	1	3
39	1	2	2	1	1	3	1	1	0	4	10
40	1	1	4	2	2	4	2	1	2	15	0
41	4	1	2	2	2	4	3	0	4	6	1
43	4	1	5	2	2	2	1	0	4	15	0
45	4	1	2	2	0	1	2	1	0	8	4

TABLE 3 Probability Distribution of Workers Given Adults Calculated from the BBN.

Adults	Workers	Prob(Workers | Adults)
0	0	88.9%
	1	11.1%
1	0	31.3%
	1	67.6%
	2	1.1%
2	0	13.5%
	1	35.3%
	2	50.7%
	3	0.6%
3	0	5.0%
	1	17.0%
	2	49.1%
	3	28.3%
	4	0.6%
4	0	3.3%
	1	6.7%
	2	33.3%
	3	20.0%
	4	36.7%
5	2	28.6%
	3	42.9%
	4	14.3%
	5	14.3%
6	2	100.0%

TABLE 4.Prediction Capability of the BBN.

	Correlation			Coefficient of Determination, R2		
Technique	HBW	HNW	NHB	HBW	HNW	NHB
Expected Value	68.4%	66.5%	45.1%	46.6%	44.2%	19.6%
Rounded Expected Value	67.2%	66.3%	44.8%	43.4%	43.8%	18.0%
Maximum Likelihood	68.2%	59.4%	29.4%	26.0%	20.3%	15.2%

TABLE 5. Percentage of Correctly Imputed Items

Technique	HBW		HNW		NHB	
	BBN	Random	BBN	Random	BBN	Random
Rounded Expected Value	47.1%	10.0%	15.2%	6.3%	10.9%	6.3%
Maximum Likelihood	50.5%	10.0%	35.9%	6.3%	43.2%	6.3%