Modeling Time-of-Day Car Use Behavior: A Bayesian Network Approach

Dawei Li (corresponding author)
Jiangsu Key Laboratory of Urban ITS, Southeast University
Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic
Sipailou 2, Xuanwu District, Nanjing 210096, China
(e-mail) lidawei@seu.edu.cn
Tel: +86-25-8379-5642

Tomio Miwa
EcoTopia Science Institute &
Green Mobility Collaborative Research Center,
Nagoya University, Japan
(e-mail) miwa@civil.nagoya-u.ac.jp

Takayuki Morikawa
Graduate School of Environmental Studies &
Green Mobility Collaborative Research Center,
Nagoya University, Japan
(e-mail) morikawa@nagoya-u.jp

Abstract: In this research, a Bayesian network (BN) approach is proposed to model the car use behavior of drivers by time of day and to analyze its relationship with driver and car characteristics. The proposed BN model can be categorized as a tree-augmented naive (TAN) Bayesian network. A latent class variable is included in this model to describe the unobserved heterogeneity of drivers. Both the structure and the parameters are learned from the dataset, which is extracted from GPS data collected in Toyota City, Japan. Based on inferences and evidence sensitivity analysis using the estimated TAN model, the effects of each single observed characteristic on car use measures are tested and found to be significant. The features of each category of the latent class are also analyzed. By testing the effect of each car use measure on every other measure, it is found that the correlations between car use measures are significant and should be considered in modeling car use behavior.

Keywords: car use, Bayesian networks, latent class, machine learning, GPS data

1. Introduction

The use of private cars is a major cause of congestion and air pollution (Goodwin, 1996). The modeling of car use behavior is essential if we are to measure the environmental and social costs of urban transportation systems, evaluate the effect of certain transportation policies (e.g. road pricing) on reduced car use, and predict travel demand.

Car use behavior depends on many factors. There is a heavy dependence on the characteristics of a household's principal driver, such as age, gender, and occupation. Previous research has shown that workers, young people, and males are likely to drive more (Hensher, 1985; Mannering, 1983; Train, 1986). Vehicle characteristics also affect car use behavior. For example, Van Wissen and Golob (1992) determined the relationship between car usage and choice of fuel type. Household characteristics, such as household size and housing quality, are another category of factors that affect car use patterns (Borgoni et al., 2002). At a more macroscopic level, land use and transport policies also affect car use behavior (Gärling et al., 2002; Jakobsson et al., 2002; Kitamura et al., 1997). From a psychological and behavioral perspective, car use patterns are also affected by attitudes, motives, and habits (Gärling et al., 1998; Gardner and Abraham, 2008; Steg, 2005).

In all of the earlier research referred to above, the measures of car usage were not time-dependent. They included such indexes as annual mileage and weekly car use frequency. However, the modeling of dynamic behavior is the inevitable trend in transportation research. A statistical car use measure such as annual mileage can only roughly reflect the car use behavior of drivers; drivers with similar values of statistical car use measures may have different car use pattern.

The modeling of time-of-day car use behavior is also required for some particular applications. As an example, Harris and Webber (2012) did a statistical analysis on time-of-day car use patterns and analyzed their impact on the provision of vehicle-to-grid services. In another investigation related to electric vehicles, describing drivers' time-of-day car use was necessary to assess the effect of battery-range limitations on the popularization of plug-in electric vehicles (Pearre et al., 2011). In these two investigations, as well as in analysis by Krumm (2012), time-of-day car use behavior was described using only basic statistics; relationships with driver or other characteristics were not modeled. Full modeling of time-of-day car use behavior has not been described in the literature, according to our limited investigations. Therefore, the main contribution of this research is to model drivers' time-of-day car use behavior and analyze its relationship to driver characteristics and other characteristics.

From the perspective of the modeling techniques used, previous research either used a single car use measure, or estimated separate models for each of multiple car use measures. However, the correlations between various car use measures are obvious: people who use their cars more frequently are more likely to accrue higher annual mileage. Furthermore, in a time-of-day analysis, each car use measure will be calculated for multiple time intervals. Therefore, in this research, the correlations among the various car use measures will be considered. Consequently, besides modeling the relationships between car use measures and observed characteristics, the main methodology challenge of this research is the representation of the correlations among multiple car use measures in different time intervals.

To this end, in modeling the complex relationships among multiple car use measures in different time intervals, we eschew the econometric models used in most travel behavior research. Instead we choose to use a Bayesian network (BN), a commonly applied model in the field of machine learning. The main advantages of the BN approach are explained in the following section.

This paper is organized as follows. Section 2 presents some background relating to our modeling technique using a BN. Section 3 describes the data used in this research and gives some basic statistics. Section 4 develops the BN model of time-of-day car use behavior. Section 5 analyzes drivers' car use behavior according to inferences obtained from the learned BN. Finally, Section 6 presents the conclusions of the research.

2. Bayesian networks

In this section, we describe some of the basic characteristics of Bayesian networks (BNs). Also called belief networks, Bayesian belief networks, Bayes nets, and sometimes also causal probabilistic networks, BNs are an increasingly popular method of modeling uncertain and complex relationships among multiple variables. As a popular tool for machine learning, they have also been widely used to solve practical problems in the transportation field, such as the development of Intelligent Transportation Systems (Li et al., 2011; Ozbay and Noyan, 2006; Zhang and Taylor, 2006), travel behavior analysis (Janssens et al., 2006; Li et al., 2013), and travel demand modeling (Castillo et al., 2008; Castillo et al., 2012).

Bayesian networks are directed acyclic graphs that allow efficient and effective representation of a joint probability distribution over a set of random variables. Formally, a Bayesian network for a set of random variables $\mathbf{X} = \{X_1, \dots, X_n\}$ is the pair $B = (G, \boldsymbol{\theta})$. The first component, G, is a directed acyclic graph whose nodes correspond to the random variables $X_1, \dots, X_n$ and whose links represent direct dependencies between the variables. Graph G, which is also called the structure of this Bayesian network, encodes the independence assumption: that each variable X_i is independent of its non-descendants given its parents in G. The second component of the pair, $\boldsymbol{\theta}$, represents the set of parameters that quantifies the network. It contains a parameter $\theta_{X_i|\Pi_{x_i}} = P_B(X_i \mid \Pi_{X_i})$ for each possible value x_i of X_i, and Π_{x_i} of Π_{X_i}, where Π_{X_i} denotes the set of parents of X_i in G. A Bayesian network B defines a unique joint probability distribution over $\mathbf{X}$ given by

$$P_B(X_1, ..., X_n) = \prod_{i=1}^{n} P_B(X_i \mid \Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}} \quad (1)$$

The reasons for using BNs in this research relate to the main advantages of Bayesian networks from the perspective of application (Uusitalo, 2007), which can be outlined as:

(1)*Suitable for small data sets:* there are no minimum sample sizes required to estimate a BN model. It has been demonstrated that Bayesian networks offer good prediction accuracy even with rather small sample sizes (Kontkanen et al., 1997).

(2)*Capable of dealing with missing data and incorporating latent variables:* the parameters of the model can be estimated from incomplete data using an Expectation-Maximization (EM) algorithm (Lauritzen, 1995; Spiegelhalter et al., 1993), a technique that will be used in this research. As a special case of missing data, latent variables can also be incorporated in a BN.

(3) *Explicit uncertain inferences:* the relationship between actions, knowledge and uncertainty can be considered explicit in the structure and parameters in a BN (Jensen and Nielsen, 2007). The updating of probabilities when some evidence is obtained for certain variables can be very easily done using certain inference algorithms.

(4)*Structural learning possible:* in addition to defining the structure G of a BN manually based on subject matter knowledge, it is also possible for the structure to be learned from data as along with the parameters (Heckerman et al., 1995).

(5)*Combining different sources of knowledge:* expert knowledge and previous analysis can be used *a priori* for developing a BN, which is then updated with data, thereby yielding a synthesis of old knowledge and new data (Heckerman et al., 1995).

To develop a BN is to specify the structure and parameters, which are both learned from data in this research. In the sections that follow, a BN for driver time-of-day car use behavior will be developed.

3. Data description and basic statistics

This section describes the data set used in this research. Descriptive statistics are used to show the daily and time-of-day car use patterns.

The GPS data used in this study was collected from private vehicles. In recent years, benefiting from the popularity of vehicle navigation system, GPS data has become an important resource and has been used in many investigations of travel behavior analysis. However, because of privacy issues, most research has been based on probe data collected from commercial vehicles rather than private vehicles, so individual drivers' characteristics could not be considered explicitly.

In this study, the data was collected from private cars in Toyota city, Japan in 2011 as a part of the Green Mobility project. Unlike some other Japanese cities in Japan (such as Tokyo), which are heavily dependent on public transportation systems, Toyota is a city on wheels. More than 200 drivers participated in this survey. On-board equipment installed in their private cars recorded their driving behavior (e.g. acceleration) as well as the GPS trajectory data. The data were uploaded to the Internet by the participants every week. The data were collected over a period of about 10 months (2011/3 to 2011/12). It should be noted that not all participants took part in the survey for this whole period. After some basic data cleaning work, data collected from 153 drivers remained for use in this study. Table 1 gives a basic description of driver and car characteristics. This is clearly a small incomplete data set: the overall size of the data set is not that large, and the occupations of some drivers are unknown. This gives good reason to apply a BN-based model for this study.

Unlike most previous research, which has used *vehicle travel mileage* (VTM) to analyze car-usage behavior, two alternative measures are used in this work: *car use rate* and *mean car use time*.

Car use rate describes the car-usage frequency of a driver in terms of the number of specific time intervals (e.g. 17:00-20:00 on working days) in which a driver uses the car compared with the number of specific time intervals for which the driver participates in the survey.

Mean car use time describes the intensity of car usage. For each driver, the *mean car use time* is the mean time for which the car is used during the specific time intervals in which the car was used.

Using GPS data, the dynamic trajectory of a car can be obtained, so travel mileage and car use time can be obtained. There are two main reasons for us to choose time rather than mileage for analysis: (1) the accuracy of car use time is higher than travel mileage, because a map-matching process is needed to obtain mileage; (2) car use time is directly related to a driver's overall time use, which is increasingly important in activity-based demand modeling.

First, we set two specific time intervals: 0:00-24:00 on working days and 0:00-24:00 on holidays. These time intervals, in fact, allow us to describe the daily car-usage behavior of drivers. Figure 1 show the frequency distribution histograms and cumulative curves of these two measures, respectively.

Figure 1 shows that the mean of daily *car use rate* on working days is slightly higher than that on holidays (*p-value* of *Z*-test is 0.13), while the mean of daily *mean car-usage time* on holidays is significantly more than that on working days (*p-value* is less than 0.01). Combined with the histograms, as expected, this indicates that drivers tend to use their cars more frequently on working days than on holidays, but use their cars more intensively on holidays. From Figure 1(b), it can also be seen that few drivers use their cars for more than 2 hours on any particular day.

These daily car use measures demonstrated that differences between car use on working days and holidays should be taken into account. However, the time interval of 24 hours means that the dynamic car use behavior of drivers cannot be taken into account. Drivers with similar daily car use measures may have very different time-of-day car use patterns. The main purpose of this research is to model the time-of-day car use behavior of drivers.

Using GPS data, the time-of-day car use behavior can also be analyzed. For time-of-day analysis, one day will be divided into many time intervals. In a paper, it is not feasible to illustrate car use behavior in each of these time intervals similarly to Figure 1. For this reason, we use the mean proportion of cars in use during each time interval to represent the car use patterns. This measure, which is also used by Chioke and Michael (2012), is defined as the mean ratio of number of cars in use during a time interval to the number of drivers participating in the survey on that day. In this case, the time interval is set at one minute. Figure 2 shows the results of this time-of-day analysis.

The difference in car usage between working days and holidays is now very clear. Morning and evening peak hours are clearly defined on working days, whereas on holidays more drivers choose to use their cars between 9:00 and 17:00, when usage is flat hours on working days. These differences are much more obvious than in daily car use statistics.

In conclusion, the descriptive statistics given in this section show that, drivers' car use behavior differs significantly between working days and holidays. The time dependent features of car use are significant and should be considered in modeling.

4. BN model of time-of-day car use behavior

The previous section gives a statistical analysis of car-usage patterns. In this section, a BN model is developed, with a structure learned from the data, to describe the relationships between time-of-day car use behavior and driver and car characteristics.

4.1 Nodes in the proposed BN

Before learning the structure, the nodes (i.e. the variables) included in the BN should be specified. The variables included in the proposed BN are shown in Table 2; this includes all of the characteristics given in Table 1. A latent class node is added to the proposed BN to describe the unobserved heterogeneity of drivers. According to the statistics given in the previous section, we divided each working day and each holiday into four time intervals: (1) 6:00-9:00: morning peak hours; (2) 9:00-17:00: daytime flat hours; (3)17:00-20:00: evening peak hours; and (4) 20:00-6:00: night off-peak hours. With two car use measures for each time interval, i.e. *car use rate* and *mean car use time*, there are a total of 16 nodes describing drivers' car use behavior in the proposed BN model. The notation used is ***WR#***, ***HR#***, ***WT#*** and ***HT#*** (**#**=1~4).

It should be noted that some of these variables are continuous. Bayesian networks can only deal with continuous variables in a limited manner (there are many constraints on variable distributions and BN structure), and the usual solution is to discretize them (Friedman and Goldszmidt, 1996). Although the discretization can only roughly capture the characteristics of the original distribution, it is efficient if the relationships between the variables are non-linear and complex. Since no satisfactory automatic discretization methods for BNs have been found (Myllymäki et al., 2002), as in most previous research, the discretization process is carried out manually based on experience. Table 2 shows the discretization of variables describing driver and car characteristics.

The two car use measures are both continuous variables that need to be discretized. The equal width method is used in this research. Since the meaningful range of *WR#* and *HR#* is [0, 1], with cut points 0.33 and 0.67, they are divided into three categories. The left half of Table 3 shows the means of *WR#* and *HR#* for each category.

WT# and *HT#* are greater than 0. The discretization process of *WT#* and *HT#* has two steps:

Step 1 *WT#* and *HT#* are normalized by scaling between 0 and 1. For example, Y is the set of all drivers' *WT1* in the training data set, $y_i \in Y$ is the *WT1* of driver *i*, then $\boldsymbol{x_i}$ is normalized by

$$Normalized(y_i) = y_i / \max(Y) \tag{2}$$

Step 2 with cut points 0.33 and 0.67, the normalized *WT#* and *HT#* are divided into three categories. The right half of Table 3 shows the means of *WR#* and *HR#* for each category.

4.2 Structure learning

The structure of a BN can be learned from the data. If there is no constraint on the structure during the structure learning process, the learned BN is called an unrestricted network. Since a latent class variable is included, an unrestricted network, in which there is no possibility of giving a class variable special status, is not the best choice in this case. Therefore, we use a special type of BN called a tree-augmented naive (TAN) Bayesian network (Friedman et al., 1997).

In a TAN Bayesian network, there is one class variable. The other variables are called attributes. The class variable has no parents and each attribute has parents consisting of the class variable and at most one other attribute. This restriction on the number of parents is to reduce the complexity and parameter count. The structure of a TAN Bayesian network can be learned by a modified Chow-Liu algorithm (Chow and Liu, 1968; Friedman et al., 1997). Since this is a well-developed algorithm in the field of machine learning, the details will not be described in this paper. *BNT,* an open-source toolbox for Matlab (Francois and Leray, 2004; Murphy, 2001) is used for the learning of the TAN Bayesian network for car use behavior. The learned structure is shown in Figure 3.

4.3 Parameter learning

Maximum likelihood estimation (MLE) is the most commonly used method of parameter learning. MLE can be used if the dataset is complete. However, our data set has missing data and, in the proposed TAN Bayesian network, a latent variable is included. Therefore the parameters of the proposed TAN Bayesian networks are learned from data using the Expectation Maximization (EM) algorithm (Lauritzen, 1995). This algorithm basically alternates between two steps:

E Step*:* complete the dataset by using the current parameter estimates $\hat{\boldsymbol{\theta}}$ to calculate expectations for the unobserved data.

M Step*:* use the completed dataset to find a new maximum likelihood estimate, $\hat{\boldsymbol{\theta}}'$, for the parameters. This estimate is then used in the next iteration of the expectation step.

This algorithm has been described abundantly in the literature, so details will not be given here. Due to the large number of parameters and limited meaning of the parameters themselves (the analysis is based on inferences, not parameters), the estimated parameters are not shown in this paper.

5. Inference and analysis

The probabilistic inference task in a BN is the task of computing the posterior marginal unobserved variable given a (possibly empty) set of evidence. Let $\mathbf{e}$ denote the evidence values for a set of variables $\mathbf{X}^o = \mathbf{X} \setminus \mathbf{X}^u$ (i.e. $\mathbf{X}^u$ is the set of unobserved variables). For a variable $X_i \in \mathbf{X}^u$, the probability $X_i = x_i$ on condition $\mathbf{e}$ is

$$P(X_i = x_i \mid \mathbf{e}) = \frac{P(X_i = x_i, \mathbf{e})}{P(\mathbf{e})} = \frac{\sum_{\mathbf{X}^u \setminus X_i} P(\mathbf{X}^u \setminus X_i, X_i = x_i, \mathbf{e})}{\sum_{\mathbf{X}^u} P(\mathbf{X}^u, \mathbf{e})} \tag{3}$$

$P(\mathbf{X}^u \setminus X_i, X_i = x_i, \mathbf{e})$ and $P(\mathbf{X}^u, \mathbf{e})$ in fact form one term of the joint probability distribution over **X**, which is calculated using Eq. (1).

In order to update the posterior marginal distribution efficiently, certain algorithms are needed. The inference algorithm used in this paper is a junction tree algorithm (Shafer and Shenoy, 1990). A junction tree algorithm can be seen as the mother of all exact inference algorithms for BNs and has been widely explained in the literature (Kjærulff and Madsen, 2008). With this algorithm, the independence properties of the BN are analyzed to establish a set of clusters and to construct a corresponding junction tree over the clusters. When updating the BN, it is necessary to calculate the posterior distributions of more than one variable. Using a junction tree to solve the inference problem, some processes can be shared, so efficiency is improved.

5.1 The effect of each single characteristic on car use measures

As is usual in the process of behavior analysis, we first explore the relationship between driver characteristics and the car use measures. In most research based on regression or economic models, qualitative behavior analysis depends on the sign and significance test results of estimated parameters. However, in this research, behavior analysis is based on inferences and evidence sensitivity analysis of the estimated BN model.

Let Z denote one of the six characteristics (i.e. rows 2-7 in Table 2) included in the proposed BN. For each characteristic Z and its possible value z_i, the posterior marginal probability distribution of each discretized car use measure S, $P(S \mid \mathbf{e} = \{Z = z_i\})$, is calculated using Eq. (3), as well as the prior probability $P(S \mid \mathbf{e} = \varnothing)$. The inference results are shown in Tables 4 and 5. In these two tables, the last columns show $P(S \mid \mathbf{e} = \varnothing)$, while the other columns show $P(S \mid \mathbf{e} = \{Z = z_i\}) - P(S \mid \mathbf{e} = \varnothing)$. As an example, the value “5” in the first cell of Table 4 (shaded) means that the probability of *WR1*=1 will increase by 5% if *Gen*=1 (the driver is female) is known, compared with the situation without any evidence (49%, shown in the last column of this row). Similarly to a significance test, we use the distance between $P(S \mid \mathbf{e} = \varnothing)$ and $P(S \mid \mathbf{e} = \{Z = z_i\})$ to investigate how sensitive each car use measure is to driver characteristics and those of their cars (Jensen and Nielsen, 2007). This distance is denoted by d and calculated as

$$d = \left| \frac{P(S \mid \mathbf{e} = \varnothing)}{P(S \mid \mathbf{e} = \{Z = z_i\})} - 1 \right| \tag{4}$$

If d is large enough (d >5% in this research), it indicates that evidence **e** is sufficient. In Tables 4 and 5, if d<=5%, corresponding cells are filled with signs without values. The main finding from these two tables is that all of the characteristics have significant effect on at least one car use measure. Given below are some specific findings from the sensitivity analysis.

The probability of a low car use rate in all time intervals increases with evidence that the driver is female (i.e. *Gen*=1). This same evidence also means the probability of low mean car use time also increases. Because the majority of drivers in the training dataset are male, evidence that the driver is male does not significantly affect the posterior marginal probability distribution of car use measures.

Regarding the effect of age, it is found that evidence *Age*=1 and *Age*=2 has similar effect on the posterior marginal probability distributions of car use measures. Evidence that the driver is older than 50 years (i.e. *Age*=3) decreases the probability of a low car use rate during all time intervals except during the flat hours on working days. During the night off-peak hours on working days, evidence *Age*=3 significantly increases the probability of high car use rate. The probability of medium mean car use time significantly increases with evidence *Age*=3.

On working days, evidence that the driver is a company employee increases the probability of a high car use rate during the day flat hours, but decreases it during morning peak hours. On holidays, company employees have a lower probability of low car use rate than organization employees. Regarding mean car use time, the company employment evidence results in a reduced probability of low mean car use time during every time interval on working days. Similarly, during the day flat hours on holidays, the probability of medium mean car use time is lower and both the probabilities of high and low mean car use time are higher.

Evidence that the car is a hybrid increases the probability of high car use rate between 6:00 and 20:00 on working days.

Evidence that *C_C*=3 decreases the probability of low mean car use time during all time intervals on both working days and holidays, while evidence *C_C*=2 has the opposite effect. The probability of high car use rate increases with evidence *C_C*=1 during the hours from 9:00 to 20:00, but decreases during the night off-peak hours on working days.

With evidence *C_D*=1, the probability of low car use rate during the daytime flat hours on working days decreases, while during other time intervals on both working days and holidays it increases. This evidence also increases the probability of low mean car use time during all time intervals on both working days and holidays.

5.2 Analysis of latent classes

The variable LC in the proposed BN model is a discrete unobserved variable with four categories that can indicate the latent class of drivers. In this section, the features of each latent class will be analyzed.

The posterior marginal probability distribution of each car use measure with each point of evidence for membership of an LC is calculated. To show the car use patterns more clearly, the posterior mean value of each measure U given evidence $\mathbf{e}$ is calculated by

$$M(U \mid \mathbf{e}) = \sum_{s=1}^{|sp(S)|} P(S = s \mid \mathbf{e}) \cdot M(U \mid S = s) \tag{5}$$

where S is the discretized car use measure and $|sp(S)|$ is the number of categories of S ($|sp(S)|=3$ in this research). $M(U \mid S = s)$ is the mean of car use measure U under the condition that the discretized category is s.

It is assumed that U is independent of any other variables when the discretized category is known, then $M(U \mid S = s)$ can be estimated using the values shown in Table 3. Using Eq. (5), the posterior mean value of the 16 car use measures are calculated and shown in Figure 4, along with the prior mean $M(U \mid \mathbf{e} = \varnothing)$ .The bars show the differences between $M(U \mid \mathbf{e} = \{LC = i\})$ and $M(U \mid \mathbf{e} = \varnothing)$.

From Figure 4, compared with the situation of no evidence, the four categories of the latent class variable can be interpreted as follows:

Latent Class 1: use their cars more frequently during the hours 9:00-17:00 and 17:00-20:00, but less frequently during morning peak hours and night off-peak hours, on both working days and holidays. The mean car use time is longer during each time interval on both working days and holidays.

Latent Class 2: use their cars more frequently during the morning peak hours, evening peak hours, and night off-peak hours on working days. Mean car-usage time is shorter during the hours 6:00-17:00 on working days. On holidays, this class has a higher car use rate, and a longer mean car use time.

Latent Class 3: use their cars less frequently. Mean car use time is also much shorter.

Latent Class 4: use their cars more frequently during night off-peak hours on both working days and holidays. Mean car use time is shorter during each time interval.

To analyze the effect of observed characteristics on latent class variables, further evidence sensitivity analysis is reported in Table 6. With evidence that the driver is female, the probability of LC=3 and 4 increases significantly, while the probability of LC=1 and 2 decreases significantly. The probability of LC=1 increases significantly with evidence that the driver is more than 50 years old. With evidence that the driver is a company employee or the car is not a hybrid, the probability of LC=1 also increases. If the driver's car is hybrid, the probability of LC=2 or 4 increases. C_C=2 makes the probability of LC=4 increase sharply. With evidence C_D=1, the probability of LC=3 increases significantly.

5.3 Correlations between car use measures

The BN model enables us to consider correlations between different car use measures. This section will explore whether the correlations are significant. For each discretized measure *S* shown in Table 3, its effect on every other discretized measure S' is scaled by

$$Eff(S,S') = \frac{\operatorname*{Max}_{s=1}^{|sp(S)|}(M(U' \mid S=s)) - \operatorname*{Min}_{s=1}^{|sp(S)|}(M(U' \mid S=s))}{M(U' \mid \varnothing)} \tag{6}$$

where *S'* is the discretization of *U'* and $M(\cdot)$ is calculated by Eq. (5).

The results are shown in Figure 5. As an example, the highlighted third rectangle from the bottom in the first column reflects the value of $Eff(WR1, WR3)$. According to the level of gray, $Eff(WR1, WR3)$ is between 10% and 20%. This makes it clear that some correlations between car use measures are significant. It is therefore necessary to consider these correlations in developing car use models.

6. Conclusions

In this research, an approach based on a Bayesian network (BN) is proposed to model the car use behavior of drivers by time of day. Two car use measures are used: *car use rate* and *mean car use time*. Each working day and holiday is divided into four time intervals, therefore there are 16 variables in total describing drivers' time-of-day car use behavior. A BN is chosen for this research because it is able to represent complex relationships between multiple random variables. The data used is GPS data collected from 153 drivers in Toyota City, Japan, using on-board equipment installed in their cars. The basic characteristics of the drivers and their cars are also available.

The proposed BN model can be categorized as a tree-augmented naive (TAN) Bayesian network. A latent class variable is included in the model to describe the unobserved heterogeneity of drivers. Both the structure and the parameters are learned from the data.

Unlike research based on regression or econometric models, in which behavior analysis looks at the sign and significance test results of estimated parameters, behavior analysis in this study depends on inferences and evidence sensitivity analysis of the estimated TAN model.

We first explore the effect of each single characteristic on car use measures. According to the evidence of sensitivity analysis, both driver characteristics and car characteristics have a significant effect on driver time-of-day car use behavior. The detailed effects of each observed characteristic are analyzed by investigating differences between posterior and prior distributions of car use measures.

Then the features of each latent class are analyzed. The four latent classes represent four different car use patterns. The effect of observed characteristics on the latent class variable is also tested and found to be significant.

The main contribution of the proposed model is that it can represent correlations among multiple car use measures in different time intervals. The final part of the analysis, therefore, is to explore the effect of each car use measure on every other measure. From the results of this analysis, it is clear that the correlations between car use measures are significant and it will be necessary to consider them when modeling car use behavior.

The scarcity of data available for this research means that only the basic characteristics of drivers and their cars are included in the proposed model. In the future, more explanatory variables, such as household characteristics, should be incorporated for a fuller picture. Further, the flexible structure of the BN means that psychological factors such as attitudes and motives can also be explicitly included in the model. There is also a possibility of combining the proposed car use model with other models, such as a car purchasing model, in the future. From the perspective of application, consideration should be given to using the proposed model for travel demand analysis in the future. In conclusion, the claim is made here that BNs, as a machine learning method now available in this "big data age" have great potential for wide use in behavior analysis.

Acknowledgements

This work described in this paper was supported by Projects of International Cooperation and Exchange of the National Natural Science Foundation of China (No. 5151101143), the Science and technology Project of Jiangsu Province, China (BK20150613), and the Fundamental Research Funds for the Central Universities. The author appreciates the comments of anonymous referees and the editor on earlier versions of the paper.

References

Citeseer (2002) *How important are household demographic characteristics to explain private car use patterns? A multilevel approach to austrian data*.

Castillo, E., Menéndez, J.M., Sánchez-Cambronero, S. (2008) Predicting traffic flow using bayesian networks. *Transportation Research Part B: Methodological* 42, 482-509.

Castillo, E., Nogal, M., Menéndez, J.M., Sánchez-Cambronero, S., Jiménez, P. (2012) Stochastic demand dynamic traffic models using generalized beta-gaussian bayesian networks. *Intelligent Transportation Systems, IEEE Transactions on* 13, 565-581.

Chioke, B.H., Michael, E.W. (2012) A temporal assessment of vehicle use patterns and their impact on the provision of vehicle-to-grid services. *Environmental Research Letters* 7, 034033.

Chow, C., Liu, C. (1968) Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on* 14, 462-467.

Francois, O., Leray, P. (2004) Bnt structure learning package: Documentation and experiments. *Proceedings of Technical Report FRE CNRS 2645. Laboratoire PSI, Universitè et INSA de Rouen*.

Friedman, N., Geiger, D., Goldszmidt, M. (1997) Bayesian network classifiers. *Machine Learning* 29, 131-163.

Friedman, N., Goldszmidt, M. (1996) Discretizing continuous attributes while learning bayesian networks. *Proceedings of Machine Learning-International Workshop Then Conference*, pp. 157-165.

Gärling, T., Eek, D., Loukopoulos, P., Fujii, S., Johansson-Stenman, O., Kitamura, R., Pendyala, R., Vilhelmson, B. (2002) A conceptual analysis of the impact of travel demand management on private car use. *Transport Policy* 9, 59-70.

Gärling, T., Gillholm, R., Gärling, A. (1998) Reintroducing attitude theory in travel behavior research: The validity of an interactive interview procedure to predict car use. *Transportation* 25, 129-146.

Gardner, B., Abraham, C. (2008) Psychological correlates of car use: A meta-analysis. *Transportation research part F: traffic psychology and behaviour* 11, 300-311.

Goodwin, P. (1996) Simple arithmetic. *Transport Policy* 3, 79-80.

Harris, C.B., Webber, M.E. (2012) A temporal assessment of vehicle use patterns and their impact on the provision of vehicle-to-grid services. *Environmental Research Letters* 7, 034033.

Heckerman, D., Geiger, D., Chickering, D.M. (1995) Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20, 197-243.

Hensher, D.A. (1985) An econometric model of vehicle use in the household sector. *Transportation Research Part B: Methodological* 19, 303-313.

Jakobsson, C., Fujii, S., Gärling, T. (2002) Effects of economic disincentives on private car use. *Transportation* 29, 349-370.

Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., Timmermans, H. (2006) Integrating bayesian networks and decision trees in a sequential rule-based transportation model. *European Journal of operational research* 175, 16-34.

Jensen, F.V., Nielsen, T.D. (2007) *Bayesian networks and decision graphs*. Springer Verlag.

Kitamura, R., Mokhtarian, P.L., Laidet, L. (1997) A micro-analysis of land use and travel in five neighborhoods in the san francisco bay area. *Transportation* 24, 125-158.

Kjærulff, U.B., Madsen, A.L. (2008) *Bayesian networks and influence diagrams*. Citeseer.

Kontkanen, P., Myllymäki, P., Silander, T., Tirri, H., Grunwald, P. (1997) Comparing predictive inference methods for discrete domains. *Proceedings of In Proceedings of the sixth international workshop on artificial intelligence and statistics*.

Krumm, J. (2012) How people use their vehicles: Statistics from the 2009 national household travel survey. *Proceedings of Society of Automotive Engineers (SAE) 2012 World Congress*.

Lauritzen, S.L. (1995) The em algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis* 19, 191-201.

Li, D., Cheng, L., Ma, J. (2011) Incident duration prediction based on latent gaussian naive bayesian classifier. *International Journal of Computational Intelligence Systems* 4, 345-352.

Li, D., Miwa, T., Morikawa, T. (2013) Dynamic route choice behavior analysis considering en route learning and choices (forthcoming). *Transportation Research Record: Journal of the Transportation Research Board*.

Mannering, F.L. (1983) An econometric analysis of vehicle use in multivehicle households. *Transportation Research Part A: General* 17, 183-189.

Murphy, K. (2001) The bayes net toolbox for matlab. *Computing science and statistics* 33, 1024-1034.

Myllymäki, P., Silander, T., Tirri, H., Uronen, P. (2002) B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools* 11, 369-387.

Ozbay, K., Noyan, N. (2006) Estimation of incident clearance times using bayesian networks approach. *Accident Analysis & Prevention* 38, 542-555.

Pearre, N.S., Kempton, W., Guensler, R.L., Elango, V.V. (2011) Electric vehicles: How much range is required for a day’s driving? *Transportation Research Part C: Emerging Technologies* 19, 1171-1184.

Shafer, G.R., Shenoy, P.P. (1990) Probability propagation. *Annals of Mathematics and Artificial Intelligence* 2, 327-351.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G. (1993) Bayesian analysis in expert systems. *Statistical science*, 219-247.

Steg, L. (2005) Car use: Lust and must. Instrumental, symbolic and affective motives for car use. *Transportation Research Part A: Policy and Practice* 39, 147-162.

Train, K. (1986) *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*. MIT press.

Uusitalo, L. (2007) Advantages and challenges of bayesian networks in environmental modelling. *Ecological modelling* 203, 312-318.

Van Wissen, L., Golob, T.F. (1992) A dynamic model of car fuel-type choice and mobility. *Transportation Research Part B: Methodological* 26, 77-96.

Zhang, K., Taylor, M.A.P. (2006) Effective arterial road incident detection: A bayesian network based algorithm. *Transportation Research Part C: Emerging Technologies* 14, 403-417.

List of Tables:

List of Figures:

Table 1 Description of Driver and Car Characteristics

Drivers		Cars	
Gender		***Type***	
Male	141	Normal	73
Female	12	Hybrid	82
Age (Years)		***Capacity*** (Unit: *kg*)	
mean (std.)	44.56 (11.06)	mean (std.)	1225.69 (576.16)
<=35	37	<=1000	61
(35-50]	60	(1000-1500]	32
>50	56	>1500	60
Occupation		***Displacement*** (Unit: *L*)	
Company employee	40	mean (std.)	1.93 (0.46)
Organization employee	90	<=1.6	30
Unknown	23	(1.6-2]	107
		>2	46

Table 2 Description on Variables in Proposed BN

Variable	***Number of Categories***	***Description***
LC	4	Unobserved, indicates latent class of driver
Gen	2	Gender: 1. Female; 2. Male
Age	3	Age: 1. <=35; 2. (35, 50]; 3. >50
Occ	2	Occupation: 1. Company employee; 2. Organization employee
C_T	2	Car type: 1. Normal; 2. Hybrid
C_C	3	Car capacity: 1. <=1000; 2. (1000, 1500]; 3. >1500 (Unit: *kg*)
C_D	3	Car displacement: 1. <=1.6; 2. (1.6, 2]; 3. >2 (Unit: *L*)
WR#, *HR#*, *WT#*, *HT#* (#=1~4)	3	*WR#* is car use rate in the #*th* time interval on working days; *HR#* is car use rate in the #*th* time interval on holidays; *WT#* is mean car use time in the #*th* time interval on working days； *HT#* is mean car use time in the #*th* time interval on holidays but discretized into three categories.

Table 3 Means of Categories of Discretized Car Use Measures

	Mean of Category				Mean of Category (Unit: *minute*)		
	1 (*low*)	2 (*medium*)	3 (*high*)		1 (*low*)	2 (*medium*)	3 (*high*)
WR1	0.11	0.48	0.76	*WT1*	15.52	33.83	57.21
WR2	0.16	0.49	0.79	*WT2*	26.39	51.96	86.04
WR3	0.15	0.46	0.74	*WT3*	20.97	32.76	52.92
WR4	0.13	0.48	0.74	*WT4*	17.62	33.64	58.64
HR1	0.11	0.42	0.69	*HT1*	15.14	34.25	61.60
HR2	0.20	0.49	0.78	*HT2*	31.42	60.54	95.62
HR3	0.17	0.45	0.75	*HT3*	20.56	38.02	60.79
HR4	0.13	0.44	0.70	*HT4*	22.33	42.89	70.60

Table 4 Effect of Driver Characteristics on *WR*# and *HR*#

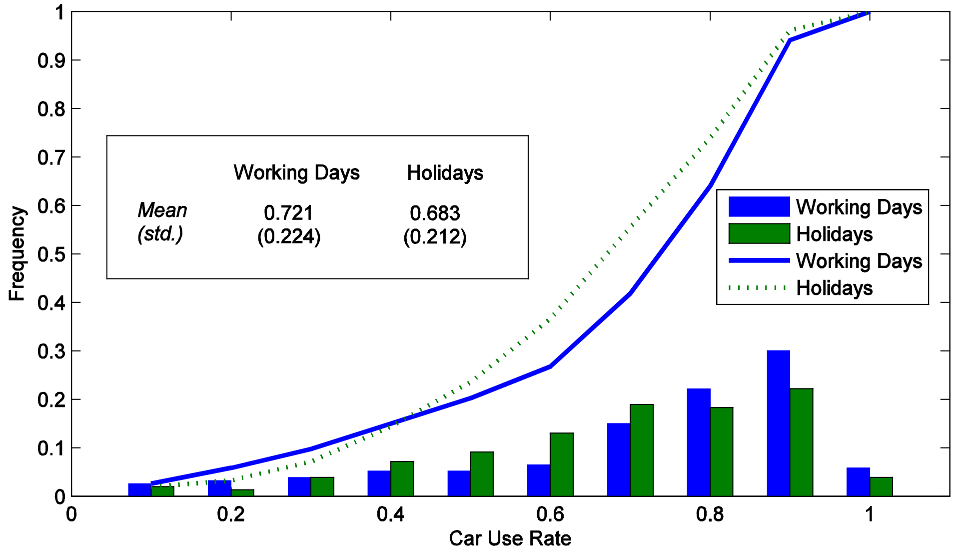
		Gen		*Age*			*Occ*		*C_T*		*C_C*			*C_D*			∅
		1	2	1	2	3	1	2	1	2	1	2	3	1	2	3	
WR1	1	5	-	7	6	-11	-	+	+	-	3	+	-4	11	-3	-3	49
	2	-2	+	-5	-5	9	3	-	4	-3	-5	-4	7	-10	3	+	29
	3	-3	+	-2	-1	3	-2	+	-4	4	+	2	-2	-	-	2	22
WR2	1	5	-	-	-	+	-	+	+	-	-3	6	+	-4	-	4	53
	2	-	+	-	-	+	-3	+	3	-2	+	-7	3	+	2	-4	33
	3	-4	+	2	1	-3	4	-2	-3	2	2	1	-3	4	-1	-	14
WR3	1	10	-	4	4	-7	-4	+	-	+	+	4	-4	7	-	-	52
	2	-7	+	-3	-3	6	4	-2	4	-4	-5	-2	6	-10	+	4	33
	3	-4	+	-1	-	1	-	+	-3	3	4	-3	-2	4	+	-3	15
WR4	1	+	-	+	+	-3	-	+	+	-	+	-3	-	6	+	-4	63
	2	-	+	+	+	-	-	+	-1	+	-	2	-	-	-	+	28
	3	-2	+	-3	-3	5	1	-0	1	-0	-3	1	2	-6	1	2	8
HR1	1	+	-	5	+	-8	-	+	-	+	+	+	-5	8	-	-	87
	2	-4	+	-5	-4	8	3	-2	2	-2	-3	-4	5	-7	2	1	12
	3	-0	+	-0	-0	0	0	-0	0	-0	-0	-0	0	-0	0	-	1
HR2	1	10	-	3	4	-6	-5	3	-	+	+	7	-4	4	-3	2	29
	2	-6	+	-4	-4	8	5	-	+	-	-4	-	4	-9	+	3	55
	3	-4	+	1	+	-2	+	-	-	+	4	-7	-	4	1	-5	16
HR3	1	12	-	4	4	-6	-6	3	-	+	-	10	-4	3	-3	3	50
	2	-10	+	-5	-4	8	5	-3	+	-	-	-7	4	-6	3	-	45
	3	-2	+	1	1	-1	0	-	-	+	2	-3	-0	2	0	-2	5
HR4	1	+	-	+	+	-	-	+	+	-	+	-	-	5	-	-	78
	2	-2	+	-2	-2	3	2	-	-	+	-2	2	1	-4	-	3	21
	3	-0	+	-0	-0	0	0	-0	0	-0	-0	-0	0	-0	0	-	1

Table 5 Effect of Driver Characteristics on *WT#* and *HT#*

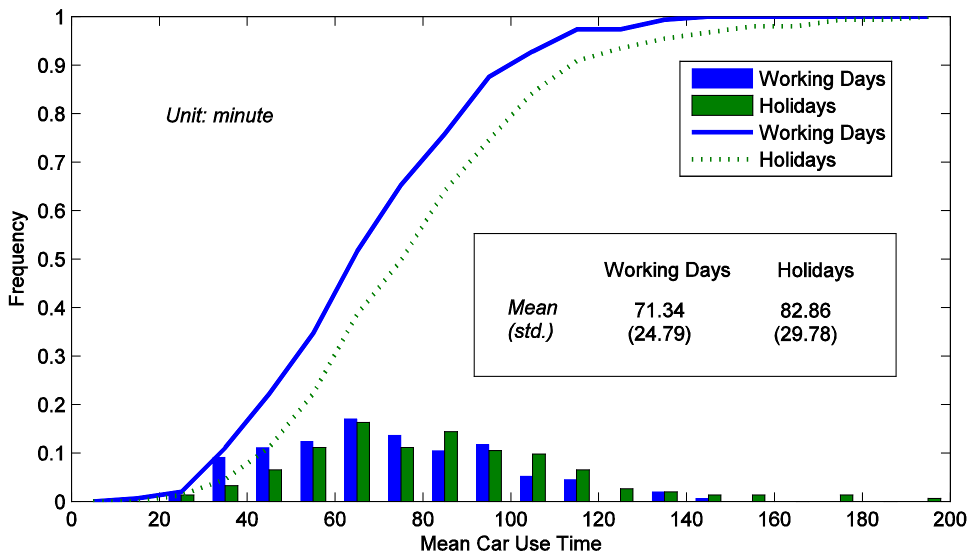
		Gen		*Age*			*Occ*		*C_T*		*C_C*			*C_D*			∅
		1	2	1	2	3	1	2	1	2	1	2	3	1	2	3	
WT1	1	6	-	7	6	-11	-3	+	-	+	3	6	-6	10	-4	-	50
	2	-6	+	-7	-6	12	3	-	+	-	-4	-5	6	-10	3	+	44
	3	0	-	+	+	-	-	+	0	-0	0	-1	+	1	0	-1	5
WT2	1	5	-	4	3	-6	-2	+	-2	+	-	8	-4	4	-3	3	32
	2	-6	+	-3	-3	5	+	-	+	-	+	-6	3	-3	+	-	53
	3	1	-	-1	-	1	+	-	1	-1	-	-2	1	-	1	-1	15
WT3	1	9	-	4	4	-6	-4	2	-2	2	-2	13	-5	2	-4	6	19
	2	-9	+	-4	-4	7	4	-	+	-	+	-14	5	-	5	-6	74
	3	+	-	+	+	-0	-	+	-	+	-	1	-	+	-	0	7
WT4	1	6	-	3	3	-5	-	+	-	+	-	10	-3	+	-3	5	46
	2	-7	+	-2	-	4	+	-	+	-	3	-12	3	+	4	-6	48
	3	0	-	-0	-0	1	+	-	-	+	-1	2	+	-1	-	1	5
HT1	1	3	-	+	+	-4	-3	+	+	-	3	-	-	6	-	-3	48
	2	-3	+	-	-	4	3	-	-	+	-3	3	+	-6	-	4	47
	3	-0	+	-0	-0	1	+	-	0	-0	0	-2	1	+	1	-1	5
HT2	1	8	-	3	3	-5	3	-2	-2	2	2	5	-5	7	-2	-2	17
	2	-5	+	-	-	+	-6	+	+	-	-	+	+	-6	-	5	69
	3	-3	+	-2	-2	3	3	-1	2	-1	+	-6	3	-1	2	-3	12
HT3	1	9	-	+	+	-2	+	-	-	+	-3	9	-2	-	-2	4	27
	2	-9	+	-	-	+	-	+	-	+	4	-9	+	+	+	-4	67
	3	+	-	-1	-1	1	1	-1	1	-1	-1	-1	1	-1	0	+	6
HT4	1	+	-	+	+	-	-	+	-	+	+	+	-	3	-	-	62
	2	-5	+	-2	-2	4	3	-	-	+	-	-2	2	-4	+	+	34
	3	2	-	1	1	-1	-0	+	1	-1	-0	1	-0	1	-1	0	4

Table 6 Effect of Driver Characteristics on *LC*

	Gen		*Age*			*Occ*		*C_T*		*C_C*			*C_D*			∅
	1	2	1	2	3	1	2	1	2	1	2	3	1	2	3	
LC=1	-16	+	-19	-16	30	14	-7	9	-8	-10	-19	20	-27	11	-	41
LC=2	-9	+	9	7	-13	-3	1	-4	3	15	-17	-6	20	+	-15	26
LC=3	14	-1	5	6	-10	-9	5	+	-	4	3	-6	11	-3	-3	19
LC=4	11	-1	5	4	-7	-2	1	-6	5	-9	32	-8	-4	-9	19	14



(a) Frequency distribution histograms and cumulative curves of daily *car use rate*.



(b) Frequency distribution histograms and cumulative curves of daily *mean car use time*.

Figure 1 Frequency distribution histograms and cumulative curves of daily car use measures.

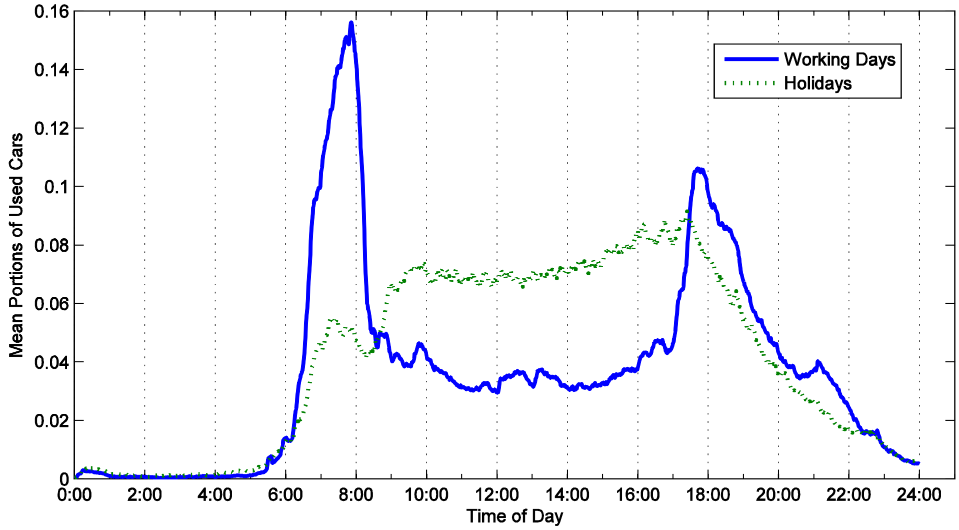


Figure 2 Mean proportion of cars in use in each time interval.

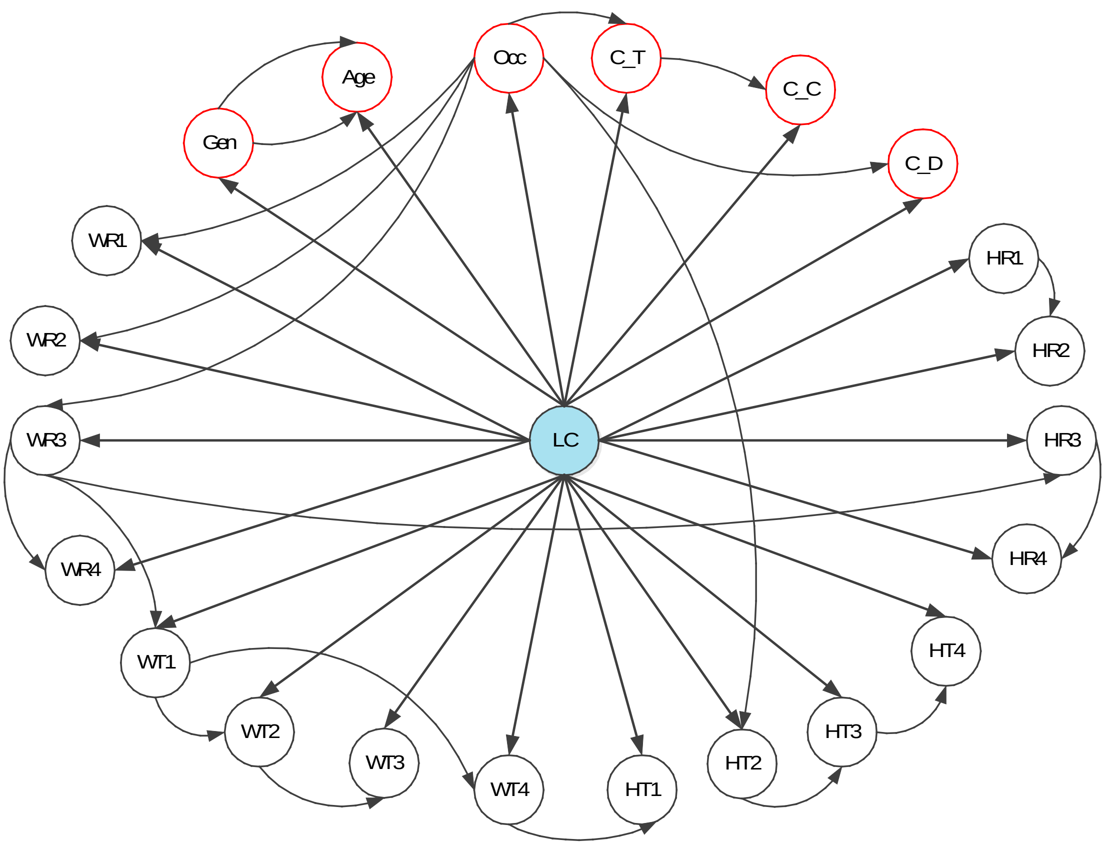
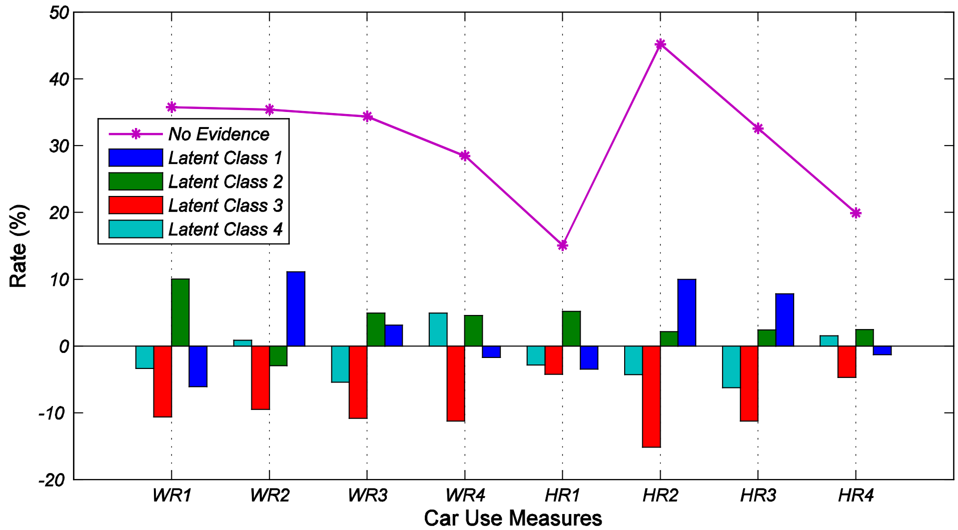
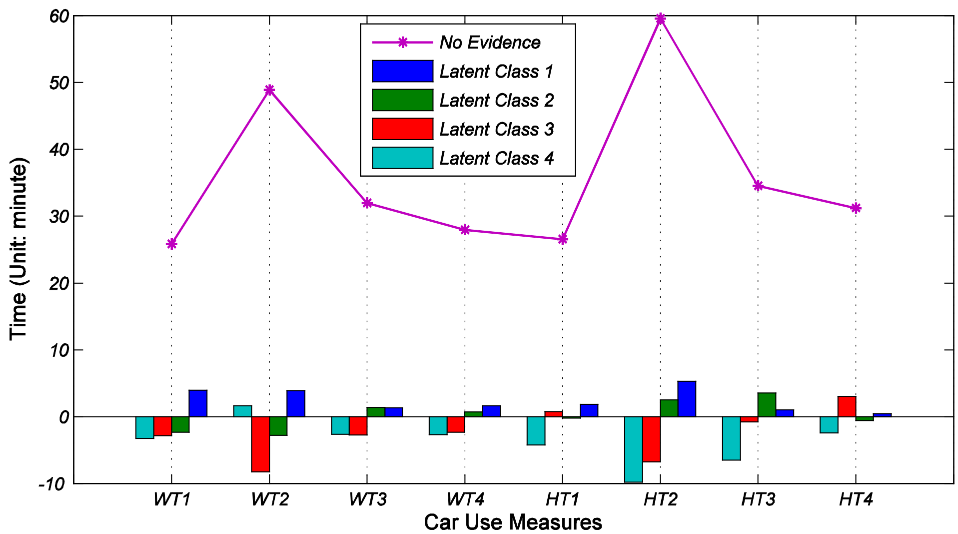


Figure 3 Learned TAN Bayesian network structure.



(a) Effect on car use rate.



(b) Effect on mean car-usage time.

Figure 4 Effect of *LC* on car use behavior.

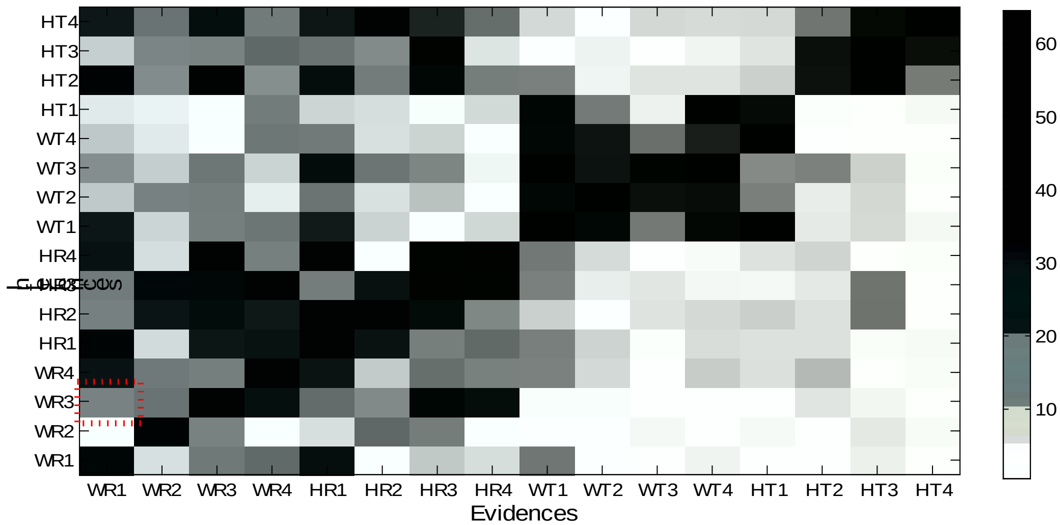


Figure 5 Correlations between car use measures.