Effectiveness of Bayesian Updating Attributes in Data Transferability Applications

Taha H. Rashidi (Corresponding Author), Ph.D.
School of Civil and Environmental Engineering
University of New South Wales
H20 CV113, UNSW, Sydney 2031
Phone: +61 2 938505063
Email: *rashidi@unsw.edu.au*

Joshua Auld, Ph.D.
Department of Civil and Materials Engineering
University of Illinois at Chicago
842 W. Taylor St. Chicago, IL 60607
Phone: 312-996-0962
Fax: 312-996-2426
Email: *auld@uic.edu*

Abolfazl (Kouros) Mohammadian, Ph.D.
Associate Professor
Department of Civil and Materials Engineering
University of Illinois at Chicago
842 W. Taylor St. Chicago, IL 60607
Phone: 312-996-9840
Fax: 312-996-2426
Email: *kouros@uic.edu*

Abstract

This paper presents the findings from an analysis of several Bayesian updating scenarios in the context of data transferability. Bayesian updating has been recognized as having great potential for use in the transportation field, especially in the simulation of travel demand and other transportation-related data. For local areas where comprehensive data collection is too costly and infeasible, Bayesian updating can be used to synthesize travel demand data in a process generally referred to as data transferability. Bayesian updating has been occasionally employed for transferring travel data; however, various aspects and disadvantages of its use have been insufficiently studied. This work addresses some issues regarding Bayesian updating techniques in data transferability, including a comparison of the use of conjugate and non-conjugate formulations in the updating models, their relative effectiveness, and the impacts of the quality of the prior information on the final results. The study shows that in general, updating small local samples of travel attribute data with prior information from national data sources provides an improved estimate of local travel attributes when compared to using the local sample only. However, it was found in this study that the inclusion of all the available historical data in the prior distributions does not necessarily improve the quality of the updating results. Therefore, a careful analysis of the applicability of the prior information to the desired context is necessary when using a Bayesian updating formulation. The National Household Travel Survey 2001 (NHTS) and the Nationwide Personal Transportation Survey 1995 (NPTS) are utilized for the demonstration exercises in this study.

Keywords: Bayesian updating, Conjugate distributions, Non-conjugate distributions, Informative prior distribution

1- Introduction

Travel demand models tend to be data intensive. The data requirements for the estimation and calibration of such models are generally satisfied through conducting disaggregate travel surveys at either the household or individual levels. However, conducting a sufficiently large disaggregate travel survey is a time and money consuming task which can be unaffordable for many small and mid- size cities and areas. As a result, small and mid-sized cities have traditionally transferred models developed for other regions and the transferred model parameters are then calibrated using local characteristics. There is a rich literature behind these model updating methods in the transportation field. Recently, data transferability models have been more frequently employed by small and mid size local areas as an alternative (1, 2). The most commonly used formulation in transferability modeling is the Bayesian updating method (3, 4). A simple conjugate normal-normal Bayesian updating procedure is a typical formulation that has been employed for updating, where both the prior distribution to transfer from and the posterior transferred result are assumed to have a normal distribution for the parameter of interest (3). Unfortunately, there are many attributes of the Bayesian updating method which have generally been overlooked in these updating models and data transferability studies in the transportation field, which could potentially limit their effectiveness and applicability. For example the effectiveness of non-conjugate distributions, non-informative priors and many other alternative types of Bayesian updating formulations have not been studied and discussed in the literature.

This study examines several Bayesian updating scenarios in which different issues about Bayesian updating are addressed. One fundamental aspect of Bayesian updating is the capability of incorporating prior information about the dependent variable. However, it is possible that the use of inappropriate prior information may result in deceptive findings and would not necessarily improve the final result. Therefore several updating scenarios with different levels of prior information, including current national information and out of date national information were used to investigate this possibility. Another issue is the determination of whether the use of more complex Bayesian updating formulations such as the inclusion of random effects or non-conjugate prior distributions will produce a better model fit. The results of the study generally show that Bayesian updating is a tool that should be cautiously employed. It can improve the model fitness and lead to better results; however, it can also lead to unintended consequences and reduced model performance if employed improperly. Therefore, the Bayesian updating method should be used with great care and consideration, and the strengths and weaknesses of the method should be taken into account.

The major objective of this study is to first demonstrate the potential of Bayesian updating to improve data collection efforts when large samples are unavailable, while also analyzing some of the commonly used types of Bayesian updating attributes in order to find a yardstick for validating the strengths and weaknesses of each of them in a real data transferability applications.

The rest of the paper is organized as follows. Initially, a literature review of travel data transferability models and Bayesian updating procedure is presented. Following that, data sources that are utilized in this study are introduced. Then, the modeling methodology and results are presented and discussed. Finally, conclusions and future research directions are presented.

2- Literature Review

Transferring models or data distribution parameters either spatially to other locations or temporally for forecasting and transferability has become a subject of interest in many fields, including transportation. The Bayesian updating method, which gives a “posterior” or updated probability distribution of some variable, model parameter, etc. of interest through the combination of a current sample of data regarding the attribute combined with some “prior” knowledge of its distribution, presents an approach to reliably transfer models in a scientifically valid way (5). In recent years, this approach has drawn much attention primarily due to advances in computational tools and the availability of off-the-shelf packages that enable researchers to utilize techniques such as Markov Chain Monte Carlo (6 and 7) and Gibbs Sampling (8) for non-conjugate distributions. Previously, only conjugate distributions, i.e. combinations of prior and sample likelihood types that result in the prior and posterior having the same distribution types, were employed for updating purposes. For instance, Mahmassani and Sinha (9) used a normal-normal Bayesian updating approach to update the trip generation of origin destination tables. Atherton and Ben-Akiva (3) also employed a normal-normal conjugate Bayesian updating approach for updating the work-trip modal split model of Washington D.C. by a sample obtained from New Bedford, Massachusetts. These two studies utilized the Bayesian updating method to examine the spatial transferability of the model parameters.

There have been a number of other transferability studies in transportation utilizing the Bayesian Updating method. Wilmot and Stopher (10) transferred travel attributes like trip rates, mode shares, and trip-length prior distributions obtained from the 1995 NPTS Survey to the North Central Texas Council of Governments Survey of 1996. They validated their transferability process with data from the Baton Rouge Personal Transportation survey conducted in 1997. In another related study, Greaves and Stopher (11) utilized local socio-demographic data for individuals and households from the census with household travel attributes to generate synthetic household travel attributes (11 and 12). In a similar study, Stopher *et al.* (13) introduced a set of homogeneous clusters for which they utilized a normal-normal conjugate Bayesian updating approach for the transferability models. The normal-normal conjugate distribution was the only formulation which was used in all of these studies. Similarly, Zhang and Mohammadian (14) studied two travel demand attributes, the trip count and average trip distance per person. They defined 11 homogeneous clusters and developed models using a gamma-normal conjugate Bayesian updating method with the Gibbs Sampling method used for parameter estimation.

As data transferability approaches attract more attention, at the same time the tendency to use Bayesian updating methods on data transferability models increases (15). Data transferability models suitably substitute the necessity for collecting household travel survey where data collection is very costly (17). Concerns about insufficient capability of data transferability models in capturing local and regional properties have promoted the necessity for using updating methods (14 and 16). Bayesian updating, as a robust updating method, bring in properties of local-area-level sample data to the transferred data in a straightforward and efficient fashion (18). Javanmardi *et* al's paper can be consulted for a practical application of data transferability framework with a Bayesian updating component (19) Other than local-area level data, Bayesian updating requires a prior estimate of the travel characteristics of interest from some other comparable sources of data. However, prior information should be up-to-date, accurate and relevant; otherwise it can be spurious and misleading.

Outside of travel demand modeling, Bayesian updating has also been used in other transportation applications such as safety and risk analysis (20, 21, 22 and 23). Some of these studies considered non-conjugate distributions such as Poisson-gamma (20). It is possible to model transferability using the simplified normal-normal distribution, but the validity of this assumption should always be tested. In cases where the normal-normal is inappropriate, non-conjugate Bayesian formulations should be considered. Therefore, non-conjugate Bayesian updating formulations should always be considered as an option (24).

Bayesian updating can also be applied to hierarchical models, where updating is performed on model hyper-parameters or for models that have parameters updated in more than one dimension. Such models are referred to as hierarchical Bayesian updating models. An example of a well-known hierarchical Bayesian updating model for a two dimensional problem is reported by Gelfand *et al.* in which they modeled the weight of rats on various days after their birth (25). In this model 30 observations for five time cross sections have been observed. The paper is also a classical example of the hierarchical Bayesian updating models where the hyper-parameters are the parameters of the probability density function of the first level parameters. It has been argued that these more sophisticated multilevel Bayesian updating models can provide better fit to the data (26). However, the merit of this argument should be probed in each case. In summary, the effectiveness of some of the different specifications of the Bayesian updating method which have recently been employed in a growing number of data transferability applications need to be examined. The literature shows there is a need for examining both conjugate and non-conjugate formulations and determining the appropriate use of each. The different manner in which the prior information and the quality of the prior information utilized in Bayesian updating impact the posterior distribution also needs to be addressed. Finally, whether the multilevel Bayesian updating approach can improve the quality of the model also requires evaluation. This paper, therefore, attempts to address these questions through several transferability exercises with known distribution data, as discussed in the following sections.

3- Data

The data used in this study was obtained from the National Household Travel Survey 2001 (NHTS) and the Nationwide Personal Transportation Survey 1995 (NPTS). The NPTS 1995 and NHTS 2001 were both sponsored by the Federal Highway Administration (FHWA), Bureau of Transportation Statistics (BTS) and National Highway Traffic Safety Administration (NHTSA). The two datasets contain detailed information about the socio-economic attributes and travel characteristics of nationally distributed households.

There are a total of 42,033 households in the final 1995 NPTS dataset. About half of the households are in the national sample and the other half belong to five add-on areas, namely, New York State, Massachusetts, Oklahoma City, Tulsa, and Seattle. It was a telephone survey that was conducted using Computer-Assisted Telephone Interviewing (CATI) technology. Detailed data on all travel of each household is collected over a 14-day period among which one day is selected for more detailed travel information collection survey.

The NHTS 2001 is a similar survey that consists of 69,817 households among which 43,779 are from the add-on samples and the remaining 26,038 households were collected at the national level. The nine add-on areas surveyed in NHTS 2001 are: Baltimore, Des Moines, Hawaii, Kentucky, Lancaster PA, New York State, Oahu (Honolulu MPO), Texas and Wisconsin. Like the NPTS 1995, the NHTS 2001 was a telephone interview.

All the models in this study are developed for the household total number of work trips per day. Work trips are estimated by including to and from work trips along with the *related to* *work* trips in NHTS and NPTS. To be consistent, NPTS 1990 definitions are used to categorize the trip purposes in both datasets.

4- Methodology and Results

Bayesian updating methods are typically used to transfer data, such as model parameter, key travel demand data distribution parameters, etc, from one context to another. For instance, they are used to synthesize data for a small region using the available data in the large metropolitan area or national level. The effectiveness of incorporating data from previous years is seldom considered. The Bayesian updating methods have also been utilized in transferring model parameters from one context to another context (9) in addition to transferring data. However, the effectiveness of Bayesian updating has usually been presumed in these applications without verification. In this study however, the efficacy of Bayesian updating in different scenarios is evaluated directly. Intuitively, it would seem that if more information is included in the prior distribution, it would improve the posterior results. However, this may not necessarily be the case. Therefore, different levels of data availability and application of various Bayesian updating techniques are studied and discussed in more details in this work.

4-1 General Introduction

The foundation of the Bayesian updating method rests on the use of Bayesian probability. The Bayesian view of probability can be seen in contrast with the Frequentist view which has been the prevailing view in probability theory in the past. The Bayesian probability paradigm incorporates a personal degree of belief in the form of the prior probability distribution and can be updated as new information is received by the observer. One central advantage of the Bayesian view is its capability of taking into account the prior available information in the current decision. The Bayes Theorem which is the groundwork of the Bayesian updating method essentially relates the conditional probabilities of two events. This theorem is valid in the Frequentist view as well while Bayesian statistics can be also applied to unknown parameters. Equation 1 shows the Bayes Theorem formulation and Equation 2 presents the Bayesian statistics formulation.

$$p(B|A) = \frac{P(B).P(A|B)}{P(A)} \qquad [1]$$

$$p(x|data) = \frac{P(x).P(data|x)}{P(data)} \qquad [2]$$

Where,

x	= the unknown model parameter(s)
data	= the sample the parameter is updated with.
p(*x*)	= the prior distribution of *x*
p(*x*|*data*)	= the posterior distribution of *x*
p(*data*|*x*)	= the likelihood of *x* given *data*

It is clear that the Bayes Theorem is used to unite new data and prior information about an unknown parameter in order to provide posterior belief about the unknown parameter given the new data. This approach has been compared to the learning approach used by individuals in the learning process. The combination of specific probability density functions selected for the prior and likelihood may result in a closed form posterior formulation, which is referred to as a conjugate distribution. The estimation of the posterior distribution parameters when conjugate distributions like normal-normal are used is straightforward as closed form solutions generally exist. However, for non-conjugate formulations where no closed form solution for the posterior parameters exists, numerical methods like Markov Chain Monte Carlo with Gibbs Sampling methods must be utilized.

In this study, Bayesian updating is used as shown in Equation 2, where the data to be updated are the parameters which determine the distributions of the parameters of the work trip count distributions. The work trip distribution parameters are assumed to be normally distributed, so that each parameter describing the work trip distribution (the mean and standard deviation in the normal case or the lambda in the exponential case) also has a mean and standard deviation associated with it. Note that this means that, in effect, the model is updating *hyper-parameters* of the distribution, rather than updating the distribution parameters directly.

4-2 Sample Size

In order to make the various evaluations of the Bayesian updating methods, a simulated transferability approach is used. The household daily work trip count is modeled in using a small data sample obtained from a known full sample, and updated using a prior distribution from the full national sample of the NHTS. This, then, allows the updated distributions to be compared to the actual distributions from the full data set from which the small sample was drawn to evaluate the performance of the model. For example, a sample can be drawn from the New York add-on and used to estimate the posterior work trip distribution for the New York region with the updating procedure using the national level distribution as the prior (simple Bayesian updating). The results of the updating procedure using the sample from New York can then be compared to the actual full-sample distribution parameters from New York (estimated from the full add-on sample), which will show how well the updating procedure performs for this region. To begin then, the minimum sample size which is required for the updating in each case is approximated. Using the NHTS 2001, different sample sizes are tested and compared against each other based on their Sum Square Error (SSE). The SSE measure is estimated based on the difference between the observed and simulated number of daily work trips per household. Several samples are randomly drawn from the NHTS 2001 for each sample size value and their mean values are compared with the actual mean values of the population through the SSE calculation (sum of the squared difference between sample and population mean for each random sample). Intuitively, it is clear that the larger the sample size is the more likely the sample mean value is close to the actual mean value (by way of the central limit theorem). Nonetheless, we are interested in having smaller samples for updating, due to the cost of collecting larger samples. Additionally, since the main purpose of using the Bayesian updating method in transferability is to employ the minimum available information, smaller sample sizes are preferred. Therefore, there is a tradeoff between the sample size and the accuracy of the model. Figure 1 shows the results of a series of simulation runs where 30 different samples with various sizes were used to obtain the optimum minimal sample size for Bayesian updating. One can observe the reducing pattern of SSE as the sample size increases. As shown in Figure 1, a sample size equal to 55 was selected as the optimum sample size in this study, because the accuracy measure (SSE) does not change considerably for sample sizes larger than 55. This is comparable with the 75 sample size that Zhang and Mohammadian (14) suggested. Generally, if a large sample is at hand then a transferability application becomes less useful, while doubling the effect of prior information with the information that we can get from the sample using a Bayesian updating method can result in acceptable estimations and forecast.

[Fig 1]

4-3 Conjugate Normal-Normal Bayesian Updating with a Non-Informative Standard Deviation Prior

The first scenario tested was conducted using a typical conjugate normal-normal Bayesian updating procedure, where a non-informative prior distribution is assumed for the standard deviation parameter of the work trip distribution. The Bayesian updating is performed using a sample of 55 individuals selected randomly for seven add-on areas, namely, Baltimore, Des Moines, Kentucky, Lancaster, New York, Texas and Wisconsin. The prior distribution of the mean parameter for the household total number of daily work trips distribution is obtained from the NHTS 2001 data.

The normal distribution is commonly used for modeling travel attributes. In addition, the normal-normal is also a conjugate formulation and therefore its application for parameter estimation is more convenient. Equation [3] presents the mathematical formulation of the likelihood and the priors used in this simple Bayesian updating model:

$$\begin{aligned} &x[i] \sim Normal(\mu, \sigma), \text{ where} \\ &\quad \mu \sim Normal(1.84, 0.2275), \\ &\quad \sigma \sim Gamma(0.001, 0.001) \end{aligned} \qquad [3]$$

It should be noted that parameters of prior distributions for the mean and standard deviations of the household work trip count distributions (i.e. the hyper-parameters) presented in Equation [3] are calculated by bootstrapping several times from the population and fitting a distribution to the bootstrapped sample in the case of the mean, and through the use of a standard non-informative distribution for the case of the standard deviation. Following that, the prior is estimated by finding the best fitted normal distribution fitted to the estimated parameters. Then the likelihood and priors shown in Equation [3] are separately updated with seven samples of 55 households which were randomly selected from the NHTS 2001 add-ons. The updating was done a total of 10 times for each add-on using a new random sample for each iteration. All the updating exercises in this study are performed using the WinBUGS software with 10,000 iterations. Table 1 presents the updated mean values for each add-on area.

[Table 1]

The first column in Table 1 presents the observed mean and standard deviation values at each add-on area. The updating process was repeated 10 times and the average of these Bayesian updating runs of the add-ons are shown in the second column of Table 2. The sum square error of the updated means and standard deviations from the observed vales are calculated over all of the iterations and presented in the last column of the above Table. Generally, the above mentioned Bayesian updating approach is not data hungry or time consuming. Instead, if a small sample is at hand, then, the available limited prior information of Equation 3 can be updated with the sample. The sample itself can be used without updating for travel demand modeling, however, it can be discerned from Table 1 that there is a substantial improvement in mean value sum square errors when updating with an informative prior was performed. The SSE values of the estimated work trip distribution parameters from both the updating procedure and taken from the sample directly are used for comparison. This test demonstrates that on average the updated distribution is more likely to be closer to the true population than a small sample from that population. It should be noted however, that this is not necessarily the case for any single iteration, where the fit to the true population of the updated distribution may or may not be better than the random sample. Therefore, in real applications when only one updating iteration is run on one sample and the true distribution for the local area is not known, the attribute distribution determined from the updated distributions cannot be said to be "more correct" than the attribute distribution found in the random sample, but rather is "more likely to be correct", and care should be taken to interpret the results accordingly. Therefore, whenever appropriate prior information is available, there is a chance that it can complement the collected sample. However, it will be shown later that spurious or outdated prior information can distance the sample attributes from the actual population attributes.

4-4 Non-Conjugate Normal-Normal Bayesian Updating with Informative Priors

The half-informative set of priors of the previous Bayesian updating formulation is extended to a full informative set of priors by adding the standard deviation prior distribution of national data to what had been presented earlier. Note that the inclusion of informative priors for both the mean and standard deviation hyper-parameters means that the updating formulation is now non-conjugate with no closed form solution. Equation 4 shows the normal-normal Bayesian updating formulation with informative priors:

$$\begin{aligned}&x[i]\sim Normal(\mu,\sigma), \text{ where}\\ &\quad \mu\sim Normal(1.84, 0.2275),\\ &\quad \sigma\sim Normal(2.2057, 0.2788)\end{aligned} \qquad [4]$$

Similar to Section 4.3, 10 random samples are drawn from the population and updated for each add-on area using Equation 4 to explore the effectiveness of the presented updating method. Results of the updating process are presented in Table 2.

[Table 2]

Table 2 shows that, similar to what was observed in the first exercise, proper prior information can complement the collected sample data if a Bayesian updating method is employed. The numbers shown under the *Randomly Sampled* column are the average of SSE of several random samples from the actual observed values. These SSEs are in some cases 8 times larger than the SSEs reported under the *Updated* column which means small random samples cannot represent the population as well as the updated distributions and should be modified with supportive external information when available.

4-5 Non-Conjugate Exponential-Normal Bayesian Updating with Informative Priors

Although, the normal distribution is commonly used in travel demand modeling, it has also been criticized to be problematic because of some of its properties such as, potential negative values, symmetric shape and long tails. In addition to the commonly used normal-normal distribution, the case of exponential-normal distribution is considered in this study. Generally, exponential shape distributions provide better fit to the household work trips per day variable (24). Since this distribution is also non-conjugate, it should be estimated with the use of numerical methods.

$$x[i] \sim Exponential(\lambda), \text{ where}$$
$$\lambda = \frac{1}{\mu},$$
$$\mu \sim Normal(1.8504, 0.2173) \qquad [5]$$

The prior information in Equation 5 is similar to what was used in the normal-normal case of Equations 3 and 4. Again, the updating exercise is done using 10 randomly drawn samples of size 55 and is compared against the normal-normal updating results which was shown to outperform the simple random sampling alternative. Table 3 shows the exponential-normal updating results and comparisons to the previous normal-normal results.

[Table 3]

As shown, the exponential-normal non-conjugate Bayesian updating does not generally provide better fit compared to the normal-normal Bayesian updating model, based on the results of Table 3. The reported SSEs of *Normal-Normal* column are smaller than the SSEs of *Exponential-Normal* column except for the Des Moines area. Therefore, it can be concluded that normal-normal Bayesian updating formulation outperforms the exponential-normal formulation in the case of household total number of daily work trips in this case.

Figure 1 schematically shows the results of the previously discussed methods including normal-normal and exponential-normal methods and the simple random sampling scenario where the mean value SSEs are compared.

[Fig 2]

Figure 2 shows the superiority of both of the updating models to just using the un-updated random sample. The margin between the updating models is also considerable, with the simple normal-normal updating generally outperforming the exponential-normal model, with some variation. All of the updating approaches previously discussed have used Bayesian updating with prior information taken directly from the national sample of which each add-on sample was a part. In the next several sections several variations on the development of the priors are evaluated to determine their impacts on the efficacy of the general Bayesian updating formulation.

4-6 Conjugate Normal-Normal Bayesian Updating with Informative Priors with Noise Effect

One may suggest that including a random effect in estimating the mean values can improve the Bayesian updating modeling fit. So, we introduce a random variable added to the mean value of the main normal distribution. This random variable assumes to be also normally distributed with mean zero and non-informative standard deviation. The complete formulation of normal-normal Bayesian updating approach with random effect is presented in Equation 6:

$$
\begin{aligned}
&x[i] \sim Normal(\mu + \mu_r, \sigma), \text{ where}\\
&\quad \mu \sim Normal(1.84{,}0.2275),\\
&\quad \sigma \sim Normal(2.2057{,}0.2788),\\
&\quad \mu_r \sim Normal(0.0, \sigma_r),\\
&\quad \sigma_r \sim Gamma(0.001, 0.001)
\end{aligned} \qquad [6]
$$

The results of applying the formulation shown in Equation 6 are presented in Table 4 below. The same procedure was used as followed in the previous sections with 10 iterations performed for each add-on.

[Table 4]

A comparison between the results shown in Table 4 demonstrates that the inclusion of noise in the formulation does not necessarily improve the modeling fitness. One may rationalize this conclusion as the noise random parameter is not useful if a good prior has already been employed whereas it might improve the modeling results if the used prior is not accurate. This scenario will be validated in the next section.

4-7 Conjugate Normal-Normal Bayesian Updating with Old Informative Priors with/without Noise Effect

It was mentioned in the previous section that additional information does not necessarily improve the modeling quality. The inclusion of a random effect in the mean value of a normal distribution was tested and discussed. Nonetheless, it was stated that the random effect could potentially be beneficial if the existing prior information is of low quality. This possibility is evaluated in this section. In general, Bayesian updating cannot necessarily surpass the random sampling approach unless a proper prior distribution has been selected. Inappropriate prior distribution can even be misleading which can skew the outcome of a Bayesian updating procedure to a spurious outcome. To demonstrate this potential, in this scenario the work trip prior distributions were taken from the NPTS 1995 dataset. The updating results using the outdated prior are then examined to evaluate the importance of an appropriate prior distribution and the effectiveness of a random effect in the quality of the final updating results. The same analysis procedure from the previous sections was used for both the scenario using the outdated prior information alone and the outdated prior with the inclusion of a noise effect. Each updating scenario used the normal-normal framework discussed in Section 4.4 and shown in Equation 4 where informative priors for the mean and standard deviation are used as this formulation was shown to work best of all of the different formulations tested.

[Table 5]

The proposed effectiveness of the inclusion of random effect in cases with improper priors has been somewhat shown in this case according to the results presented in Table 5. The SSE values for the mean are generally lower with the inclusion of noise, with some exceptions, although the SSE values for standard deviation are higher. Seemingly, when the priors are not reliable it can be useful to include a random effect which can capture the unobserved deviation from the actual value.

A general comparison with all the discussed updating scenarios along with the simple random sampling approach is presented in the next subsection.

4-8 Summary of Findings and Results

Based on the preceding discussions about the developed models, the goodness-of-fit of measures for the different models have been compared against each other and shown in Figure 3. The results of the comparison support the primary focus of this work, namely that caution should be used when applying different specifications of Bayesian updating method in the context of travel data transferability.

[Fig 3]

Figure 3 shows a large disparity in the ability of different Bayesian updating formulations to represent the true distribution of the work trip counts in the NHTS add-on samples. Plainly, it can be claimed that a Bayesian updating model with up-to-date and relevant prior distributions, such as the Normal-Normal or Exponential-Normal models using priors from the 2001 NHTS, can improve the information that can be extracted from a sample. However, as can be seen form Figure 3, the updating using the NPTS 95 priors provide even worse results compared to the simple random sampling. However, the inclusion of a random effect in the priors of this ineffective updating model may reduce the impact of the bad prior information selection while it deprecates the outcomes of an updating model in which suitable prior distributions have been utilized. While the specific results from this study are not generally applicable to all transferability problems, the results do show that care should be utilized to fit the updating technique selected to the available data to achieve the best possible fit, as improper prior selection can actually degrade performance and give worse results than using no updating at all.

5- Conclusion

The applications of the Bayesian updating formulation in the transportation and travel demand fields are continually growing. Improving the state of belief and knowledge about data by incorporating the existing prior information is one of the major properties of the Bayesian updating that makes this approach superior when compared to other approaches to transferability. In particular, recent advances in the areas of synthetic disaggregate population generation and travel data transferability and simulation have resulted in further development in the areas of Bayesian statistics. Data simulation studies, in particular, have employed the Bayesian updating method because in theory there is no limitation on the type of the distribution assumption for the priors. Therefore, any type of probability density function including continuous or discrete and conjugate or non-conjugate can be assumed for the prior distributions. However, in practice usually normal distribution is assumed. Furthermore, the application of Markov Chain Monte Carlo and Gibbs Sampling methods facilitated applications of Bayesian updating. This study attempts to depart from the simple Bayesian updating assumptions used in other travel data transferability studies by introducing a rarely used application of the Bayes theorem to the travel demand and data simulation field.

The Bayesian updating method has been employed in this study to model the household total number of work trips per day. Several types of models were developed in this study for seven add-on samples of the NHTS 2001. The priors in these models were obtained from the NHTS 2001 and NPTS 1995 national-level surveys and were later updated randomly by selected local samples drawn from NHTS add-ons with the size of 55 observations to develop posterior travel demand parameter distributions. Both conjugate and non-conjugate formulations were tested in this study. It was found that the normal-normal conjugate formulation perform slightly better than the non-conjugate exponential-normal formulation in the case of household daily work trip rates. However, it is recommended that both conjugate and non-conjugate formulation would be tested in other Bayesian updating practices. It was found that Bayesian updating with properly selected priors significantly outperforms a simple randomly selected sample, which should in general be the case as the sample size decreases. On the other hand, it was found that an out-dated, inappropriate prior may result in misleading results. The use of a more complex hierarchical Bayesian updating formulation which makes the formulation more free in parameter selection (i.e. non-conjugate normal-normal) compared to the simple conjugate normal-normal with non-informative standard deviation is added to the formulation did not change the outcome in this case. Therefore, more complex formulations with limited information do not necessarily improve the goodness of fit, depending on the problem. Finally, the inclusion of a random effect in the updating formulation was tested. It was found that random effect variable in a case that prior distributions are very informative and are expected to have close relation with the target context, is not recommended. Nonetheless, in the case that outdated or inappropriate priors are at hand, inclusion of a random effect variable may alleviate the impacts of the use of the inappropriate priors.

Further improvements to the presented paper can be categorized into three chief groups: evaluating the presented scenarios for other household travel attributes, repeating the presented scenario on other data sets such as NHTS 2008 using the updated posteriors of this study as the prior distributions and finally, studying other types of non-conjugate distributions when other household travel attributes are examined.

6- References

1 Mei B., Cooney T.A., Bostrom N. R., (2005), Using Bayesian Updating to Enhance 2001 NHTS Kentucky Sample Data for Travel Demand Modeling, *Journal of Transportation and Statistics*, 8 (3): 71-82

2 Hu, P. S., Reuscher T., Schmoyer R. L., and Chin S., Transferring 2001 National Household Travel Survey, ORNL/TM-2007/013 Prepared by Oak Ridge National Laboratory for the U.S. Department of Energy. Prepared for the Federal Highway Administration, U.S. Department of Transportation, May 2, 2007.

3 Atherton, T. J., and Ben-Akiva, M. E., (1976),Transferability and updating of disaggregate travel demand models, *Transportation Research Record*, 610:12-18

4 Mohammadian, A. and Y. Zhang, 2007. Investigating Transferability of National Household Travel Survey Data. Trasportation Research Record , 1993: 67-79

5 Reuscher, T. R., Schmoyer, R., and Hu, P. S. (2002), Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales, *Transportation Research Record,* 1817: 25-32

6 Gilks, W., Richardson, S., and Spiegelhalter, D., (1996), Markov Chain Monte Carlo Methods in Practice, CRC Press

7 Stopher, P. R., Graham P., (2004), Monte Carlo simulation of household travel survey data with Bayesian updating, Road & Transport Research Journal, 13(4): 22-33

8 Casella G. and Edward G. I., (1992), Explaining the Gibbs sampler, *The American Statistician*, 46 (3): 167–174

9 Mahmassani, H. S. and Sinha, K. C. (1981) "Bayesian Updating of Trip Generation Parameters." *ASCE*, Vol. 107, No. TE5

10 Wilmot, C.G. and P. R. Stopher, (2001), Transferability of Transportation Planning Data, *Transportation Research Record*, 1768: 36-43

11 Greaves, S.P. and Stopher, P.R. (2000), Creating a Synthetic Household Travel/Activity Survey -Rationale and Feasibility Analysis, *Transportation Research Record* 1706, pp. 82-91

12 Stopher P.R., Bullock P., and Greaves S., (2003), Simulating household travel survey data: Application to two urban areas, *82nd Annual Meeting Transportation Research Board*, Washington D.C.

13 Stopher, P.R., Greaves S., and Xu, M., (2001), Using Nationwide Household Travel Data for Simulating Metropolitan Area Household Travel Data, Paper presented at *TRB Conference on 2001 National Household Travel Survey*, Washington, D.C.

14 Zhang Y., and A. Mohammadian (2008), Bayesian Updating of Transferred Household Travel Data, *Transportation Research Record,* 2049: 111-118

15 Akhil K. V., C Shinji, N S. Sathe and R E. Folsom (2010) Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey, *Transportation,* 37, pp. 825-848

16 Rashidi, T. H., and A. Mohammadian, Household Travel Attributes Transferability Analysis: Application of Hierarchical Rule Based Approach, *Transportation*, 38(4): 697-714

17 Reuscher, T. R., Schmoyer, R. L., and Hu, P. S. (2002) Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales, *Transportation Research Record,* 1817, 25-35

18 Kothuri S. (2002) Bayesian updating of simulated household travel survey data for small/medium metropolitan areas, Master's Thesis, Luisiana State University, Baton Rouge, Louisiana, USA

19 Javanmardi, M., T. H. Rashidi, and A. Mohammadian, Household Travel Data Simulation Tool: Software and Its Applications for Impact Analysis, *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2183: 9-18

20 Fu L., (2007), Traffic Safety Study: Empirical Bayes or Full Bayes?, *86th Annual Meeting Transportation Research Board*, Washington D.C.

21 Higle J.L. and Witkowski J.M., (1988), Bayesian Identification of Hazardous Sites, *Transportation Research Record*, No. 1185: 24-35

22 Persaud, B., Lyon C., and Nguyen T., (1999), Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record* No. 1665: 7-12

23 Heydecker B.G. and Wu J., (2001), Identification of Sites for Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference, *Advances in Engineering Software,* Vol. 32, pp. 859-869

24 Rashidi, T. H., A. Mohammadian, and Y. Zhang, (2010) Effects of Variation in Household Sociodemographics, Lifestyles, and Built Environment on Travel Behavior, Transportation Research Record, 2156: 64-72

25 Gelfand A.E., Hills S.E., Racine-Poon A., (1990) Smith A., Illustration of Bayesian inference in normal data models using Gibbs sampling, Journal of the American Statistical Association, 85: 972-985

26 Legay C., Rodriguez M.J., Miranda-Moreno L.F., Sérodes J.B., Levallois P. (2010) Multi-level modelling of chlorination by-product presence in drinking water distribution systems for human exposure assessment purposes, Environmental monitoring and assessment, pp. 624-637

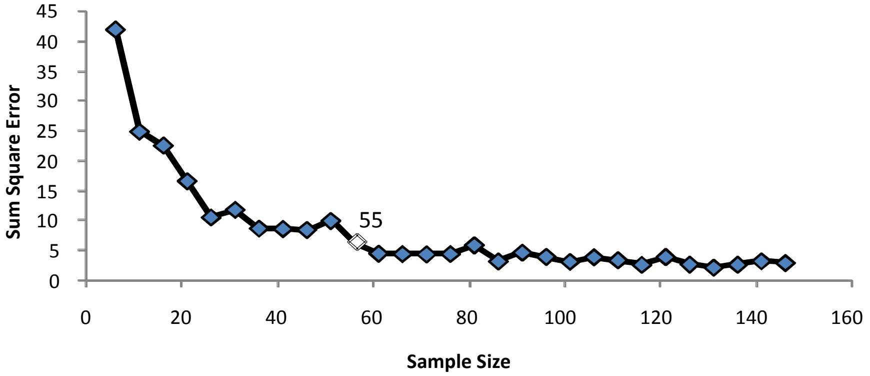

Fig 1 Simulation test to find the optimum sample size

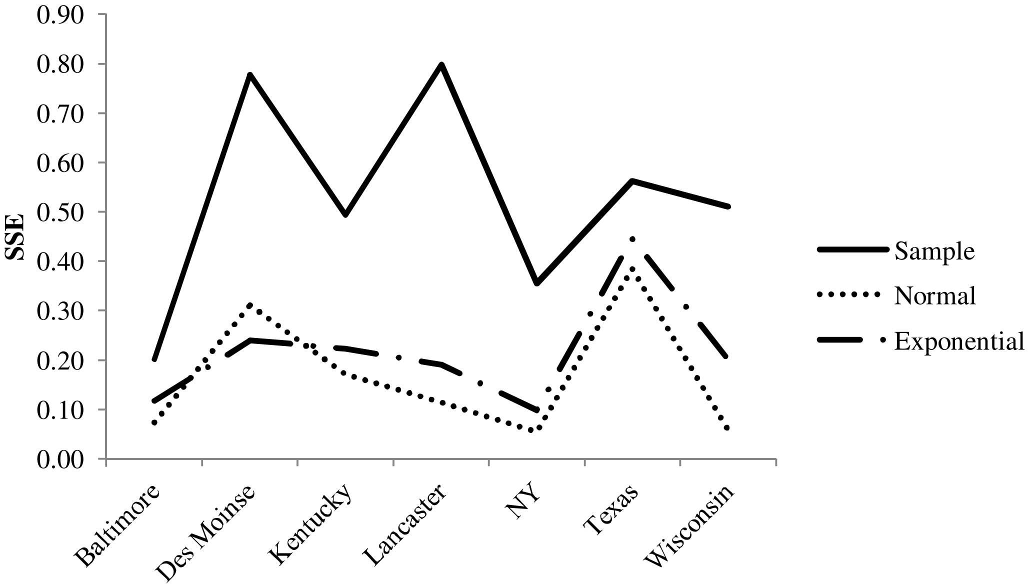

Fig 2 Comparison between SSE of mean values of the exponential-normal Bayesian updating, normal-normal Bayesian updating and simple random sampling methods

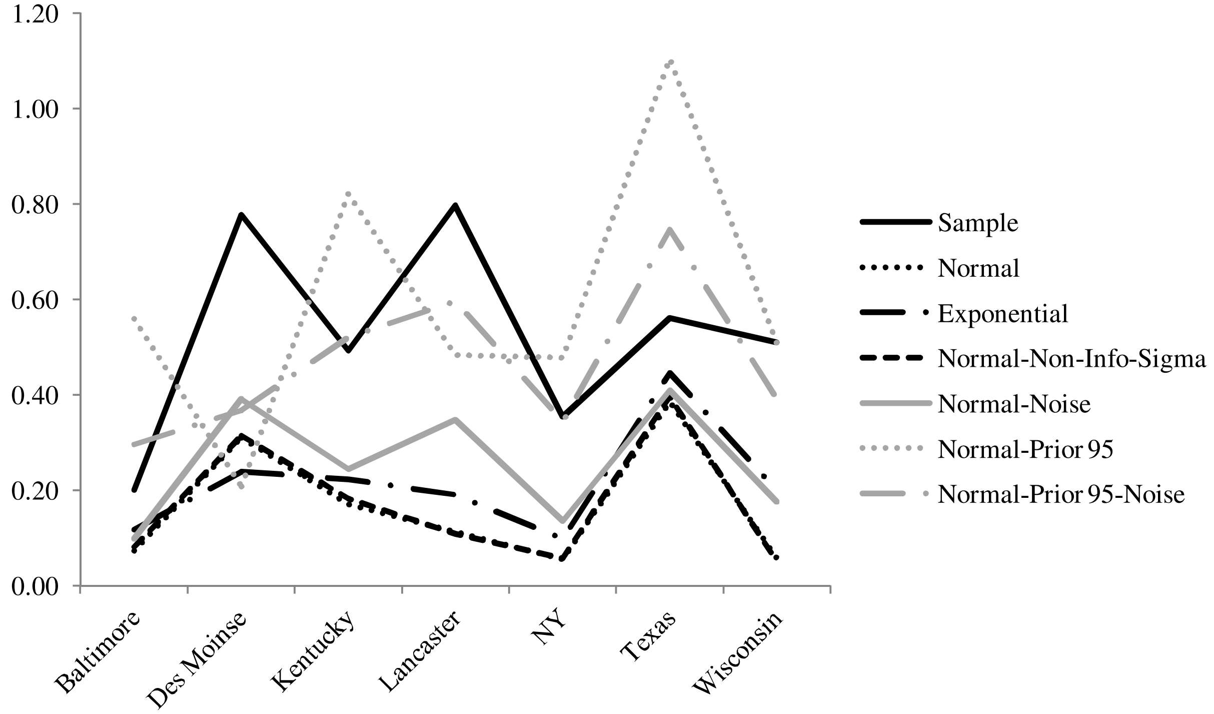

Fig 3 Comparison among the SSEs of mean values of all the introduced scenarios

Table 1 Mean and standard deviation values for average work trips per household for seven add-ons in 2001 for normal-normal distributions with informative mean and non-informative standard deviation priors

			Randomly Sampled		Updated	
Add-ons	Observed	Updated	SSE-Mean	SSE-Sigma	SSE-Mean	SSE-Sigma
Baltimore	1.73(1.91)	1.79(1.93)	0.20	0.16	0.08	0.16
Des Moinse	2.06(2.31)	1.91(2.34)	0.78	0.49	0.31	0.52
Kentucky	1.67(2.08)	1.77(2.1)	0.49	0.49	0.18	0.50
Lancaster	1.86(2.13)	1.85(2.16)	0.80	0.53	0.11	0.55
NY	1.79(2.18)	1.83(2.21)	0.35	0.59	0.06	0.65
Texas	1.55(2)	1.71(2.03)	0.56	0.98	0.40	0.94
Wisconsin	1.82(2.33)	1.83(2.36)	0.51	1.51	0.05	1.52
National	1.84(2.21)					

Table 2 Mean and standard deviation values for average work trips per household for seven add-ons in 2001 for normal-normal distributions with informative priors

			Randomly Sampled		Updated	
Add-ons	Obsereved	Updated	SSE-Mean	SSE-Sigma	SSE-Mean	SSE-Sigma
Baltimore	1.73(1.91)	1.79(2.01)	0.20	0.16	0.07	0.20
Des Moinse	2.06(2.31)	1.92(2.28)	0.78	0.49	0.31	0.21
Kentucky	1.67(2.08)	1.77(2.13)	0.49	0.49	0.17	0.24
Lancaster	1.86(2.13)	1.84(2.16)	0.80	0.53	0.11	0.21
NY	1.79(2.18)	1.82(2.19)	0.35	0.59	0.05	0.25
Texas	1.55(2)	1.72(2.07)	0.56	0.98	0.39	0.41
Wisconsin	1.82(2.33)	1.83(2.28)	0.51	1.51	0.06	0.54

Table 3 Mean and standard deviation values for average work trips per household for seven add-ons in 2001 for exponential-normal distributions with the informative prior

			Normal-Normal	Exponential-Normal
Add-ons	Normal-Normal	Exponential-	SSE-Mean	SSE-Mean
Baltimore	1.79(2.01)	1.82	0.07	0.12
Des Moinse	1.92(2.28)	1.95	0.31	0.24
Kentucky	1.77(2.13)	1.77	0.17	0.22
Lancaster	1.84(2.16)	1.86	0.11	0.19
NY	1.82(2.19)	1.85	0.05	0.10
Texas	1.72(2.07)	1.71	0.39	0.45
Wisconsin	1.83(2.28)	1.86	0.06	0.20

Table 4 Mean and standard deviation values for average work trips per household for seven add-ons in 2001 for normal-normal distributions with noise

		Normal-Normal		Normal-Normal with Noise	
Add-ons	Mean (Std. Dev.)	SSE-Mean	SSE-Sigma	SSE-Mean	SSE-Std. Dev
Baltimore	1.77(2.04)	0.07	0.20	0.10	0.29
Des Moinse	1.98(2.29)	0.31	0.21	0.39	0.20
Kentucky	1.73(2.14)	0.17	0.24	0.25	0.24
Lancaster	1.84(2.19)	0.11	0.21	0.35	0.20
NY	1.81(2.2)	0.05	0.25	0.14	0.24
Texas	1.64(2.09)	0.39	0.41	0.41	0.45
Wisconsin	1.83(2.27)	0.06	0.54	0.18	0.50

Table 5 Mean and standard deviation values for average work trips per household for seven add-ons in 2001 for normal-normal distributions with informative 1995 prior with/without noise

			Priors 95		Priors 95 with Noise	
Add-ons	Priors 95	Priors 95 with Noise	SSE-Mean	SSE-Std. Dev.	SSE-Mean	SSE-Std. Dev.
Baltimore	1.95(2.14)	1.86(2.15)	0.56	0.64	0.30	0.68
Des Moines	2.13(2.39)	2.1(2.4)	0.21	0.21	0.37	0.23
Kentucky	1.93(2.25)	1.81(2.26)	0.82	0.48	0.52	0.50
Lancaster	2.03(2.28)	1.95(2.29)	0.48	0.38	0.60	0.41
NY	2(2.31)	1.91(2.32)	0.48	0.39	0.34	0.42
Texas	1.84(2.1)	1.69(2.2)	1.11	0.52	0.75	0.73
Wisconsin	2.02(2.39)	1.93(2.41)	0.51	0.43	0.39	0.45