Effectiveness of Bayesian Updating Attributes in Data Transferability Applications

Taha H. Rashidi, Joshua Auld, and Abolfazl (Kouros) Mohammadian

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 issues about Bayesian updating techniques in data transferability, including a comparison of the use of conjugate and nonconjugate formulations in the updating models, their relative effectiveness, and impacts of the quality of the prior information on 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 compared with using the local sample only. However, this study found that the inclusion of all available historical data in the prior distributions does not necessarily improve the quality of the updating results. Therefore, careful analysis of the applicability of the prior information to the desired context is necessary when a Bayesian updating formulation is used. The 2001 National Household Travel Survey and the 1995 Nationwide Personal Transportation Survey were used for the demonstration exercises in this study.

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 the individual level. However, conducting a sufficiently large disaggregate travel survey is a time- and money-consuming task that can be unaffordable for many small and midsize cities and areas. As a result, small and midsized cities have traditionally transferred models developed for other regions, and the transferred model parameters are then calibrated by using local characteristics. These model updating methods in the transportation field have a rich literature behind them. Recently, data transferability models have been more frequently employed by small and midsize 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; in it, both the prior distribution from which to transfer and the posterior transferred result are assumed to have a normal distribution for the parameter of interest (*3*). Unfortunately, many attributes of the Bayesian updating method have generally been overlooked in both these updating models and data transferability studies in the transportation field, and this fact could potentially limit their effectiveness and applicability. For example, the effectiveness of nonconjugate distributions, noninformative 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 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 nonconjugate 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 so as to find a yardstick for validating the strengths and weaknesses of each of them in 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 used in this study are introduced. Then, the modeling methodology and results are presented and discussed. Finally, conclusions and future research directions are presented.

T. H. Rashidi, School of Civil and Environmental Engineering, University of New South Wales, Room CV113, Building H20, Sydney, New South Wales 2031, Australia. J. Auld and A. Mohammadian, Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 West Taylor Street, Chicago, IL 60607. Corresponding author: T. H. Rashidi, rashidi@unsw.edu.au.

Transportation Research Record: Journal of the Transportation Research Board, No. 2344, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 1–9.
DOI: 10.3141/2344-01

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, or other element of interest through a combination of a current sample of data about the attribute and some prior knowledge of its distribution—presents an approach for reliable transfer of models in a scientifically valid way (*5*). Recently, this approach has drawn much attention primarily because of the advances in computational tools and the availability of off-the-shelf packages that enable researchers to use techniques such as Markov chain Monte Carlo (*6, 7*) and Gibbs sampling (*8*) for nonconjugate 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 approach for updating the work-trip modal split model of Washington, D.C., by a sample obtained from New Bedford, Massachusetts. These two studies used the Bayesian updating method to examine the spatial transferability of the model parameters.

Several other transferability studies have used the Bayesian updating method in transportation research. Wilmot and Stopher (*10*) transferred travel attributes like trip rates, mode shares, and trip-length prior distributions from the 1995 Nationwide Personal Transportation Survey (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*) used local sociodemographic data for individuals and households from the census with household travel attributes to generate synthetic household travel attributes (*11, 12*). In a similar study, Stopher et al. introduced a set of homogeneous clusters for which they used a normal–normal conjugate Bayesian updating approach for the transferability models (*13*). The normal–normal conjugate distribution was the only formulation that was used in all 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 by 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 for which data collection is extremely costly (*14*). Concerns about insufficient capability of data transferability models in capturing local and regional properties have promoted the necessity for using updating methods (*14, 16*). Bayesian updating, as a robust updating method, brings in properties of sample data at the local-area level to the transferred data in a straightforward, efficient fashion (*17*). Javanmardi et al. can be consulted for a practical application of a data transferability framework with a Bayesian updating component (*18*). Other than data from a local-area level, 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 (*19–22*). Some of these studies considered nonconjugate distributions such as Poisson gamma (*19*). It is possible to model transferability by using the simplified normal–normal distribution, but the validity of this assumption should always be tested. For cases in which the normal–normal is inappropriate, nonconjugate Bayesian formulations should be considered. Therefore, nonconjugate Bayesian updating formulations should always be considered as an option (*23*).

Bayesian updating can also be applied to hierarchical models, for which updating is performed on model hyperparameters, or to 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, as reported by Gelfand et al., in which they modeled the weight of rats on various days after the rats' birth (*24*). In this model, 30 observations for five-time cross sections have been observed. That paper is also a classical example of the hierarchical Bayesian updating models in which the hyperparameters 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 a better fit to the data (*25*). 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 that have recently been employed in a growing number of data transferability applications need to be examined. The literature shows a need for examining both conjugate and nonconjugate formulations and for determining the appropriate use of each. The different manner in which prior information and the quality of prior information used in Bayesian updating affect 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.

DATA

The data used in this study were obtained from the 2001 National Household Travel Survey (NHTS) (*26*) and the 1995 NPTS (*27*). The 1995 NPTS and the 2001 NHTS were both sponsored by the Bureau of Transportation Statistics, FHWA, and NHTSA. The two data sets contain detailed information about the socioeconomic attributes and travel characteristics of nationally distributed households.

The final 1995 NPTS data set included 42,033 households. About half the households were in a national sample, and the other half belong to five add-on areas, namely, New York State; Massachusetts; Oklahoma City, Oklahoma; Tulsa, Oklahoma; and Seattle, Washington. It was a telephone survey conducted by computer-assisted telephone interviewing. Detailed data on all travel of each household were collected over 14 days, among which 1 day is selected for collection of even more detailed travel information.

The 2001 NHTS 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 are at the national level. The nine add-on areas surveyed in NHTS 2001 are Baltimore, Maryland; Des Moines, Iowa; Hawaii; Kentucky; Lancaster, Pennsylvania; New York State; Oahu (Honolulu metropolitan planning organization), Hawaii; Texas; and Wisconsin. Like the 1995 NPTS, the 2001 NHTS was a telephone interview.

All the models in the current study were developed for the household total number of work trips per day. Work trips were estimated by including to and from work trips along with the related to work trips in the 2001 NHTS and 1995 NPTS. To be consistent, 1995 NPTS definitions are used to categorize the trip purposes in both datasets.

METHODOLOGY AND RESULTS

Bayesian updating methods are typically used to transfer data, such as model parameter, key travel demand data distribution parameters, and so forth, from one context to another. For instance, they are used to synthesize data for a small region by 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.

General Background

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.

$$p(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)} \tag{1}$$

$$p(x|\text{data}) = \frac{P(x) \cdot P(\text{data}|x)}{P(\text{data})} \tag{2}$$

where

x = unknown model parameters,
data = sample with which parameter is updated,
$P(x)$ = prior distribution of x,
$p(x|\text{data})$ = posterior distribution of x, and
$P(\text{data}|x)$ = 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 so as to provide posterior belief about the unknown parameter given the new data. This approach has been compared with the approach used by individuals in the learning process. The combination of specific probability density functions selected for the prior and the 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 nonconjugate formulations, for which no closed-form solution for the posterior parameters exists, numerical methods like Markov chain Monte Carlo with Gibbs sampling must be used.

In this study, Bayesian updating is used, as shown in Equation 2, in which the data to be updated are the parameters that determine the distributions of the parameters of the distributions of work trip counts. The parameters of the work trip distributions 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 lambda in the exponential case) also has a mean and standard deviation associated with it. This assumptions means that, in effect, the model is updating hyperparameters of the distribution rather than the distribution parameters directly.

Sample Size

To make the various evaluations of the Bayesian updating methods, a simulated transferability approach is used. The count of household daily work trips is modeled by using a small data sample from a known full sample and updated by using a prior distribution from the full national sample of the 2001 NHTS. This procedure, then, allows the updated distributions to be compared with 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 sample and used to estimate the posterior distribution of work trips for the New York region with the updating procedure by using the national level distribution as the prior (simple Bayesian updating). The results of the updating procedure that uses the sample from New York can then be compared with the actual parameters of the full-sample distribution from New York (estimated from the full add-on sample), which will show relative performance of the updating procedure for this region. To begin then, the minimum sample size that is required for the updating in each case is approximated. By using the 2001 NHTS, different sample sizes are tested and compared against each other on the basis of their sum square error (SSE). The SSE measure is estimated on the basis of the difference between the observed and the simulated number of daily work trips per household. Several samples are randomly drawn from 2001 NHTS 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, the larger the sample size is, the more likely it is that the sample mean value will be close to the actual mean value (by way of the central limit theorem). Nonetheless, the authors are interested in having smaller samples for updating because of the

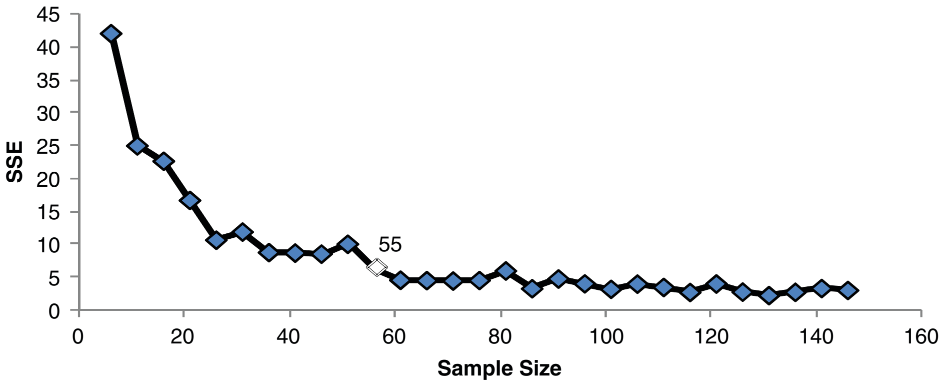

FIGURE 1 Simulation test to find optimum sample size.

cost of collecting larger samples. Furthermore, because the main purpose of using the Bayesian updating method in transferability is to employ the minimum-available information, smaller sample sizes are preferred. Therefore, a tradeoff exists between the sample size and the accuracy of the model. Figure 1 shows the results of a series of simulation runs for which 30 samples of 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 Figure 1 shows, a sample size of 55 was selected as optimum in this study because the accuracy measure does not change considerably for sample sizes larger than 55. This size is comparable to the sample of 75 that Zhang and Mohammadian (*14*) suggested. Generally, if a large sample is at hand, then a transferability application becomes less useful; in contrast, doubling the effect of prior information with the information available from the sample by using a Bayesian updating method can result in acceptable estimations and forecast.

Conjugate Normal–Normal Bayesian Updating with Noninformative Prior for Standard Deviation

The first scenario tested was conducted by using a typical conjugate normal–normal Bayesian updating procedure, in which a noninformative prior distribution is assumed for the standard deviation parameter of the work trip distribution. The Bayesian updating was performed by means of 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 distribution of the household total number of daily work trips was obtained from 2001 NHTS 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:

$$x[i] \sim \text{normal}(\mu, \sigma) \qquad (3)$$

where μ ~ normal(1.84, 0.2275) and σ ~ gamma(0.001, 0.001).

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

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. The average of these Bayesian updating runs of the add-ons are shown in the second column of Table 2. The SSE of the updated means and standard deviations from the observed values are calculated over all the iterations and presented in the last column of the Table 2. 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, as Table 1 shows, a substantial improvement occurred in mean value SSEs when updating with an informative prior was performed. The SSE values of the parameters of the estimated work trip distribution from both the updating procedure and directly from the sample were 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. However, this outcome is not necessarily the case for any single iteration, for which 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.

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 Noninformative Standard Deviation Priors

			Randomly Sampled		Updated	
Add-On Area	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 Moines	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
New York	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

NOTE: Numbers in parentheses are standard deviations; average number of work trips per household at national level has mean value of 1.84 and sigma value of 2.21.

Nonconjugate Normal–Normal Bayesian Updating with Informative Priors

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

$$x[i] \sim \text{normal}(\mu, \sigma) \qquad (4)$$

where μ ~ normal(1.84, 0.2275) and σ ~ normal(2.2057, 0.2788).

As in the earlier section on a noninformative prior for standard deviation, 10 random samples were again drawn from the population and updated for each add-on area by using Equation 4 to explore the effectiveness of the presented updating method. Results of the updating process are presented in 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 for the randomly sampled data are the average of SSEs of several random samples from the actual observed values. These SSEs are in some cases eight times larger than the SSEs reported for the updated data, which means that small random samples cannot represent the population as well as the updated distributions and should be modified with supportive external information when available.

Nonconjugate Exponential–Normal Bayesian Updating with Informative Priors

Although the normal distribution is commonly used in travel demand modeling, it has also been criticized as 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 (*23*). Because this distribution is also nonconjugate, it should be estimated by numerical methods.

$$x[i] \sim \text{exponential}(\lambda) \qquad (5)$$

where λ = 1/μ and μ ~ normal(1.8504, 0.2173).

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 was done by using 10 randomly drawn samples of 55 and was compared against the normal–normal updating results and was shown to outperform the simple random sampling alternative. Table 3

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-On Area	Observed	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 Moines	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
New York	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 Informative Priors

			SSE Mean	
Add-On Area	Normal–Normal	Exponential–Normal	Normal–Normal	Exponential–Normal
Baltimore	1.79 (2.01)	1.82	0.07	0.12
Des Moines	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
New York	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

shows the exponential–normal updating results and comparisons to the previous normal–normal results.

From the results in Table 3, the exponential–normal nonconjugate Bayesian updating does not generally provide better fit than the normal–normal Bayesian updating model. The reported SSEs for the normal–normal data are smaller than those for the exponential–normal except for the Des Moines area. Therefore, it can be concluded that the 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 in which the mean value SSEs were compared.

Figure 2 shows the superiority of both the updating models to simply using the unupdated 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 the updating approaches previously discussed used Bayesian updating with prior information taken directly from the national sample of which each add-on sample was a part. In the next sections, several variations on the development of the priors are evaluated to determine their impacts on the efficacy of the general Bayesian updating formulation.

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 a random variable added to the mean value of the main normal distribution was introduced. This random variable was also assumed to be normally distributed with mean zero and noninformative standard deviation. The complete formulation of the normal–normal Bayesian updating approach with random effect is presented in Equation 6:

$$x[i] \sim \text{normal}(\mu + \mu_r, \sigma) \tag{6}$$

where

$\mu \sim$ normal(1.84, 0.2275),
$\sigma \sim$ normal(2.2057, 0.2788),
$\mu_r \sim$ normal(0.0, σ_r), and
$\sigma_r \sim$ gamma(0.001, 0.001).

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

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 not being 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.

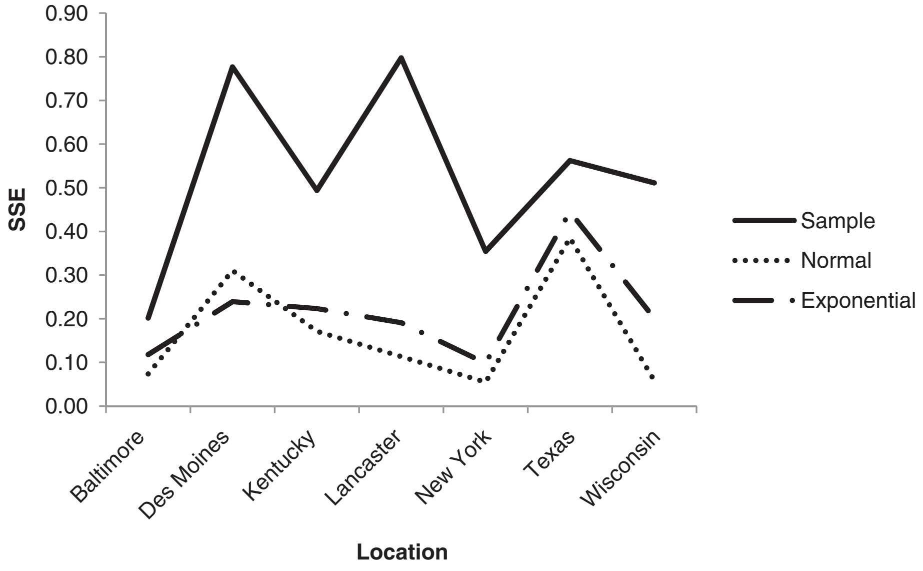

FIGURE 2 Comparison between SSE of mean values of exponential–normal Bayesian updating, normal–normal Bayesian updating, and simple random sampling methods.

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-On Area	Mean (SD)	SSE Mean	SSE Sigma	SSE Mean	SSE SD
Baltimore	1.77 (2.04)	0.07	0.20	0.10	0.29
Des Moines	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
New York	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

NOTE: SD = standard deviation.

Conjugate Normal–Normal Bayesian Updating with Old Informative Priors With and Without Noise Effect

The previous section noted that additional information does not necessarily improve 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 here. In general, Bayesian updating cannot necessarily surpass the random sampling approach unless a proper prior distribution has been selected. An inappropriate prior distribution can even be misleading and 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 1995 NPTS data set. The updating results using the outdated prior were 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 employed for both the scenario by 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 earlier and shown in Equation 4, for which informative priors for the mean and standard deviation were used, as this formulation was shown to work best of the different formulations tested.

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

Summary of Findings and Results

On the basis of 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 one is applying different specifications of the Bayesian updating method in the context of travel data transferability.

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 2001 NHTS add-on samples. Plainly, the claim came be made that a Bayesian updating model with up-to-date and relevant prior distributions, such as the normal–normal or exponential–normal models that use priors from the 2001 NHTS, can improve the information that can be extracted from a sample. However, as Figure 3 shows, the updating using the 1995 NPTS

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 and Without Noise

			Priors 1995		Priors 1995 with Noise	
Add-On Area	Priors 1995	Priors 1995 with Noise	SSE Mean	SSE SD	SSE Mean	SSE SD
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
New York	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

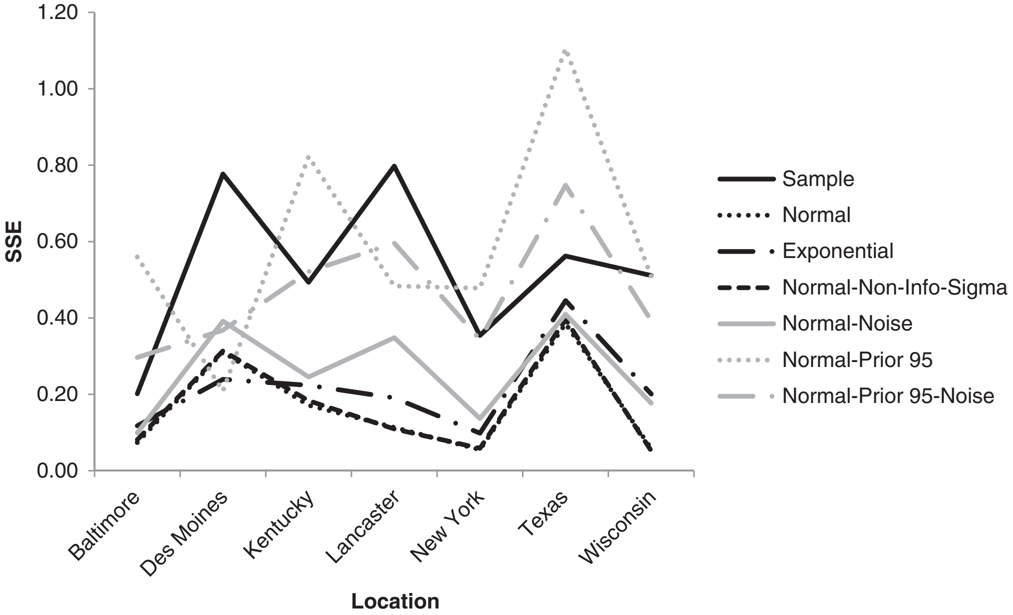

FIGURE 3 Comparison between SSEs of mean values of all introduced scenarios.

priors provides even worse results than those from 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 used. While the specific results from this study are not generally applicable to all transferability problems, the results show that care should be taken in fitting the updating technique selected to the available data to achieve the best possible fit, as improper selection of priors can actually degrade performance and give worse results than using no updating at all.

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 compared with 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 the type of the distribution assumption for the priors has no limitations. Therefore, any type of probability density function, including continuous or discrete and conjugate or nonconjugate, can be assumed for the prior distributions. However, in practice, normal distribution is usually 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 2001 NHTS. The priors in these models were obtained from the 2001 NHTS and 1995 NPTS national-level surveys and were later updated randomly by selected local samples drawn from 2001 NHTS add-ons, with 55 observations to develop posterior distributions of travel demand parameters. Both conjugate and nonconjugate formulations were tested in this study. It was found that the normal–normal conjugate formulation performed slightly better than the nonconjugate exponential–normal formulation in the case of rates of household daily work trips. However, it is recommended that both conjugate and nonconjugate formulations would be tested in other Bayesian updating practices. It was found that Bayesian updating with properly selected priors significantly outperformed a simple randomly selected sample, which should in general be the case as the sample size decreases. In contrast, it was found that an outdated, inappropriate prior may result in misleading results. The use of a more complex hierarchical Bayesian updating formulation that makes the formulation more free in parameter selection (i.e., nonconjugate normal–normal) compared with simple conjugate normal–normal with noninformative 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 inclusion of a random effect in a case in which prior distributions are quite 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: evaluation of the presented scenarios for other household travel attributes, repetition of the presented scenario on other data sets such as the 2008 NHTS by using the updated posteriors of this study as the prior distributions, and, finally, studying other types of nonconjugate distributions when other household travel attributes are examined.

REFERENCES

1. Mei, B., T. A. Cooney, and N. R. Bostrom. Using Bayesian Updating to Enhance 2001 NHTS Kentucky Sample Data for Travel Demand Modeling. *Journal of Transportation and Statistics,* Vol. 8, No. 3, 2005, pp. 71–82.
2. Hu, P. S., T. Reuscher, R. L. Schmoyer, and S. Chin. *Transferring 2001 National Household Travel Survey.* ORNL/TM-2007/013. Oak Ridge National Laboratory, Oak Ridge, Tenn., May 2, 2007.
3. Atherton, T. J., and M. E. Ben-Akiva. Transferability and Updating of Disaggregate Travel Demand Models. In *Transportation Research Record 610,* TRB, National Research Council, Washington, D.C., 1976, pp. 12–18.
4. Mohammadian, A., and Y. Zhang. Investigating Transferability of National Household Travel Survey Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1993,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 67–79.
5. Reuscher, T. R., R. Schmoyer, Jr. and P. S. Hu. Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1817,* Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 25–32.
6. Gilks, W., S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo Methods in Practice.* CRC Press, Boca Raton, Fla., 1996.
7. Stopher, P. R., and G. Pointer. Monte Carlo Simulation of Household Travel Survey Data with Bayesian Updating. *Road and Transport Research,* Vol. 13, No. 4, 2004, pp. 22–33.
8. Casella, G., and G. I. Edward. Explaining the Gibbs Sampler. *American Statistician,* Vol. 46, No. 3, 1992, pp. 167–174.
9. Mahmassani, H. S., and K. C. Sinha. Bayesian Updating of Trip Generation Parameters. *Transportation Engineering,* Vol. 107, No. 5, 1981, pp. 581–589.
10. Wilmot, C. G., and P. R. Stopher. Transferability of Transportation Planning Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1768,* TRB, National Research Council, Washington, D.C., 2001, pp. 36–43.
11. Greaves, S. P., and P. R. Stopher. Creating a Synthetic Household Travel and Activity Survey: Rationale and Feasibility Analysis. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1706,* TRB, National Research Council, Washington, D.C. 2000, pp. 82–91.
12. Stopher, P. R., P. Bullock, and S. Greaves. Simulating Household Travel Survey Data: Application to Two Urban Areas. Presented at 82nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2003.
13. Stopher, P. R., S. Greaves, and M. Xu. Using Nationwide Household Travel Data for Simulating Metropolitan Area Household Travel Data. Presented at TRB Conference on 2001 National Household Travel Survey, TRB, National Research Council, Washington, D.C., 2001.
14. Zhang, Y., and A. Mohammadian. Bayesian Updating of Transferred Household Travel Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2049,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 111–118.
15. Akhil, V., C. Shinje, N. Sathe, R. Folsom, P. Chandhok, and K. Guo. Small Area Estimates of Daily Person-Miles of Travel: 2001 National Household Transportation Survey, *Transportation,* Vol. 37, No. 6, 2010, pp. 825–848.
16. Rashidi, T. H., and A. Mohammadian. Household Travel Attributes Transferability Analysis: Application of Hierarchical Rule Based Approach. *Transportation,* Vol. 38, No. 4, July 2011, pp. 697–714.
17. Kothuri, S. M. *Bayesian Updating of Simulated Household Travel Survey Data for Small–Medium Metropolitan Areas.* MS thesis. Louisiana State University, Baton Rouge, 2002.
18. Javanmardi, M., T. H. Rashidi, and A. Mohammadian. Household Travel Data Simulation Tool: Software and Its Applications for Impact Analysis. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2183,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 9–18.
19. Miranda-Moreno, L. F., and L. Fu. Traffic Safety Study: Empirical Bayes or Full Bayes? Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
20. Higle, J. L. and J. M. Witkowski. Bayesian Identification of Hazardous Sites. In *Transportation Research Record 1185,* TRB, National Research Council, Washington, D.C., 1988, pp. 24–35.
21. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1665,* TRB, National Research Council, Washington, D.C., 1999, pp. 7–12.
22. Heydecker, B. G., and J. Wu. Identification of Sites for Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. *Advances in Engineering Software,* Vol. 32, 2001, pp. 859–869.
23. Rashidi, T. H., A. Mohammadian, and Y. Zhang. Effects of Variation in Household Sociodemographics, Lifestyles, and Built Environment on Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2156,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 64–72.
24. Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. Smith. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association,* Vol. 85, 1990, pp. 972–985.
25. Legay, C., M. J. Rodriguez, L. F. Miranda-Moreno, J. B. Sérodes, and P. Levallois. Multi-Level Modelling of Chlorination By-Product Presence in Drinking Water Distribution Systems for Human Exposure Assessment Purposes. *Environmental Monitoring and Assessment,* Vol. 178, No. 1–4, July 2011, pp. 507–524.
26. *2001 National Household Travel Survey.* Bureau of Transportation Statistics, FHWA, U.S. Department of Transportation, 2001.
27. *1995 Nationwide Personal Transportation Survey.* Bureau of Transportation Statistics, FHWA, U.S. Department of Transportation, 1995.

The Transportation Demand Forecasting Committee peer-reviewed this paper.