International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 24(2) Special, December, 2024, pp 163-179
DOI: https://doi.org/10.3329/ijss.v24i20.78222
© 2024 Dept. of Statistics, Univ. of Rajshahi, Bangladesh

Optimizing Transport Predictive Modeling with Simulation-Based Statistical Inference

Quyen Tran[1] and Mamunur Rashid[2*]

[1,2]Department of Mathematical Sciences, DePauw University, Greencastle, IN, United States

[*]Correspondence should be addressed to Mamunur Rashid (Email: mrashid@depauw.edu)

[Received October 23, 2024; Accepted December 1, 2024]

Abstract

Simulation-based statistical inference (SBI) leverages computer simulations to help scientists understand and analyze complex data. This paper explores how SBI techniques can be used to analyze transportation data. We use modern computational methods, including machine learning models, to improve the accuracy of predictions and decision-making in transportation planning. Our study focuses on two SBI methods, Approximate Bayesian Computation - Markov Chain Monte Carlo and Synthetic Likelihood, to create synthetic data for training machine learning models. These models show the potential of SBI to handle uncertain transportation data. It also highlights the practical benefits of SBI in making better decisions for transportation systems.

Keywords: Simulation-based statistical inference, simulation-based inference, Synthetic Likelihood, Approximate Bayesian Computation, machine learning, predictive modeling, computational methods, transportation planning.

AMS Classification: 62F15.

1. Introduction

Simulation-based statistical inference (SBI) leverages computer simulations to help scientists understand and analyze complex data. Simulators' flexibility has made them indispensable tools for predicting system behavior across numerous scientific and engineering disciplines. SBI represents a significant advancement in the methodological evolution of

statistical practice, finding applications in fields such as statistics, computer science, engineering, and robotics.

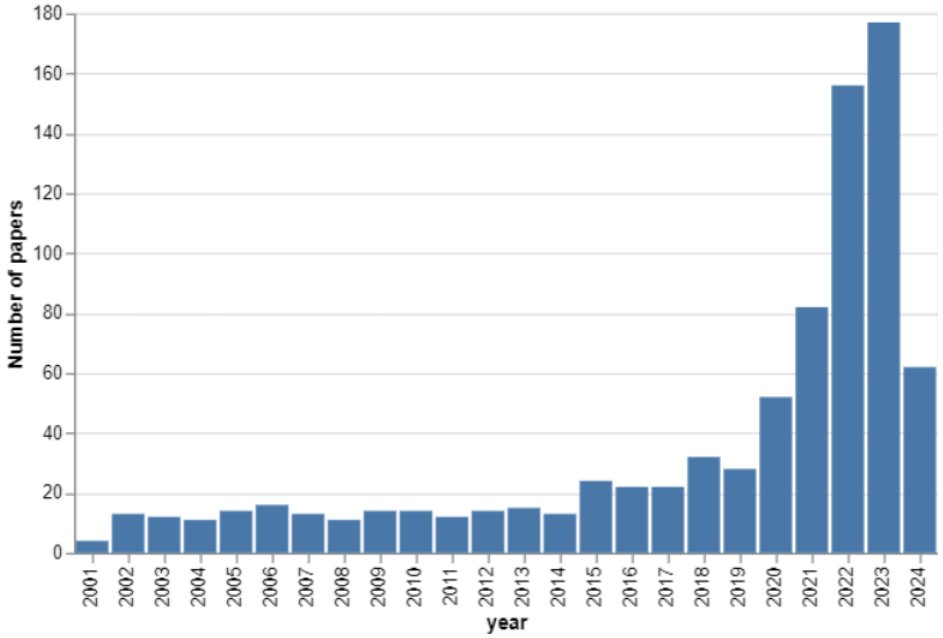
In the past, implementing SBI faced various challenges, including limitations in computational power and the complexity of high-dimensional data. However, with the rise of artificial intelligence (AI) and machine learning, people can now work with more complex data, significantly improving the quality and range of their inferences.

Transportation systems are vital in modern societies, impacting economic development, environmental sustainability, and social equity. Analyzing transport data is essential for understanding travel patterns, optimizing infrastructure investments, and improving the efficiency of transportation networks. However, traditional statistical inference methods may face challenges in capturing the complexity and uncertainty inherent in transportation systems. This research explores applying simulation-based statistical inference techniques to analyze transport data, leveraging their strengths to enhance predictive modeling and decision-making processes in transportation planning and management.

2. Literature Review

2.1. Trend of SBI Articles in Recent Years

Figure 1: Number of Simulation-based Inference Papers by Year [20]



From 2001 to 2019, the number of SBI papers published each year remained low, generally below 40 papers per year. However, starting in 2020, there was a notable increase in the publication of SBI papers, with a significant rise observed between 2021 and 2023. In the current year, 2024, the trend is predicted to continue with an increasing number of publications.

The sharp rise in publications suggests that SBI has become a crucial area of research in recent years. This could be due to advancements in computational power, especially in AI and machine learning, which have led to the development of new methods and broader recognition of SBI's utility across various fields.

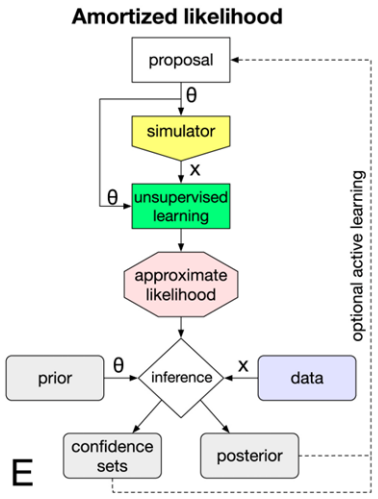
2.2. Classical Methods

2.2.1. Density Estimation

Density estimation methods are used to approximate the distribution of the summary statistics from samples generated by the simulator.

Density estimation is used to make inferences from complex statistical models. People focus on using the kernel method for non-parametric density estimation, which involves using a smooth, flexible function to estimate the probability distribution of simulated data. The kernel density estimator is chosen for its simplicity and effectiveness, allowing for accurate estimation even when the underlying data distribution is complex or unknown.

Figure 2: Density Estimation [3]



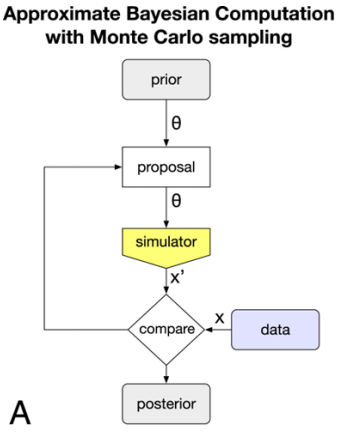
This method is particularly valuable for continuous data, where it estimates the log-likelihood function by summing the logarithms of the kernel estimates at each observed

data point. This approach helps make inferences about the parameters of implicit statistical models, providing a practical solution where traditional likelihood-based methods are not applicable.

2.2.2. Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation (ABC) is another popular method where scientists compare real-world data with simulated data. However, because it approximates results, ABC might not always be as precise as other methods like Markov Chain Monte Carlo (MCMC), especially when exact calculations are possible.

Figure 3: Approximate Bayesian Computation (ABC) [3]



It has been used for many years to tackle complex problems in genetics, where traditional methods are too slow or difficult, primarily to make inferences about population size, growth rates, and other genetic parameters.

Today, ABC is still important in genetics but is also used in other fields like epidemiology, ecology, and evolutionary biology. With the advancement of technology and improved methods, ABC has become more accurate and useful. Scientists use it to study the spread of diseases, track changes in animal populations, and understand complex biological processes.

2.3. Improvement over Classical Methods

Recent advancements in Simulation-Based Inference (SBI) techniques have significantly improved over classical methods. Neural Likelihood Estimation (NLE) utilizes neural networks to estimate likelihood functions from simulated data, offering flexibility in modeling complex, high-dimensional distributions [12]. Normalizing Flows transform simple probability distributions into complex ones through invertible transformations, enabling effective density estimation for intricate data structures [13]. Feature Selection Through Likelihood

Marginalization (FSLM) efficiently identifies important data features without repeated full estimations, reducing computational costs [2]. Approximate Bayesian Computation using Markov Chain Monte Carlo simulation (DREAM(ABC)) employs the Differential Evolution Adaptive Metropolis algorithm for efficient sampling in high-dimensional spaces [16]. The sequential Monte Carlo (SMC) method improves sampling from sequences of probability distributions, improving exploration and efficiency [17]. Integrating Graph Neural Networks (GNNs) with Approximate Bayesian Computation (ABC) automates the extraction of summary statistics, boosting accuracy and computational efficiency [6]. These advancements leverage machine learning and computational power to handle complex data and improve inference quality.

2.4. Application in transportation

In transportation, SBI methods, including Approximate Bayesian Computation and Density Estimation, are applied to large, high-dimensional datasets to estimate travel time reliability and route choice preferences. Manole and Niles-Weed (2024) used high-dimensional data points from particle collisions in high-energy physics experiments at the Large Hadron Collider (LHC) [9]. Unnikrishnan, Kochar, and Figliozzi used travel time data from the Portland, OR metropolitan region, with 17,491 observations after removing outliers [19]. These methods improve the accuracy of utility function coefficient estimations that describe travelers' route choices and improve traffic flow [9], [7], [19].

ABC is utilized for likelihood-free inference in high-dimensional problems where the likelihood function is computationally prohibitive. It estimates parameters without explicit likelihood functions, offering flexibility and scalability. However, it is computationally intensive and sensitive to the choice of summary statistics and tolerance levels.

Density estimation methods improve the modeling of background events, providing detailed insights into distributional properties. Techniques such as density ratio extrapolation adjust for shifts between auxiliary and background distributions, while empirical optimal transport coupling refines background estimates [9]. Non-parametric methods like Kernel Density Estimation offer flexibility without assuming a specific form for the distribution but can suffer from boundary issues and performance dependence on bandwidth choice [19]. Integrating these methods ensures more accurate and feasible background modeling, which is crucial for identifying rare events in high-energy physics experiments.

SBI techniques have the potential to significantly improve transportation models. These methods offer flexible and adaptable solutions for complex, high-dimensional problems. However, their computational intensity and sensitivity to various parameters present challenges, especially when applied to smaller datasets due to limited resources.

3. Methodology

Despite the effectiveness of these methods with high-dimensional and large datasets, our resources did not permit handling such extensive data. Therefore, we applied these tech-

niques to much smaller datasets.

We used the Bureau of Transportation Statistics data on Indiana transportation from 2010 to 2021 to predict the data for 2022. We decided to do this because we could compare the prediction data with the actual data.

We conducted tests using progressively increasing data points to test the method's effectiveness. We started using from 1 data point and gradually added more until 10 data points for predictions. This approach helps us to check whether the method can be effective even when using minimal data.

This is the dataset we used:

Table 1: Indiana transportation data from the Bureau of Transportation Statistics[1]

Year	Licensed drivers	Vehicles
2010	5,550,469	5,698,010
2011	5,669,665	6,132,770
2012	5,375,973	6,004,370
2013	4,500,403	5,574,030
2014	4,448,099	6,012,600
2015	4,467,848	6,045,114
2016	4,553,259	6,140,530
2017	4,553,584	6,170,034
2018	4,589,405	6,190,736
2019	4,589,405	6,223,460
2020	4,532,708	6,199,901
2021	4,636,114	6,241,291
2022	4,653,808	6,256,479

After exploring several Simulation-Based Inference (SBI) techniques, we implemented them in our transportation prediction model.

We used two different techniques: Approximate Bayesian Computation - Markov Chain Monte Carlo (ABC-MCMC) and Synthetic Likelihood methods. The synthetic data generated by these methods was then used to train machine learning models, specifically Random Forest and Gradient Boosting Machine (GBM). We implemented all of these in R.

Both GBM and Random Forest are highly regarded machine learning models for prediction. At first, we trained both models with our data and found that the GBM provided better results. Consequently, we chose to continue using only the GBM model.

One disadvantage of these models is their requirement for a large amount of data for accurate prediction. SBI techniques are particularly useful when dealing with limited or

[1]Bureau of Transportation Statistics. (n.d.). *State highway travel*. Retrieved June 3, 2024, from `https://www.bts.gov/browse-statistical-products-and-data/state-transportation-statistics/state-highway-travel`

hard-to-find data.

We generated synthetic data using ABC-MCMC and Synthetic likelihood methods to overcome data shortage and train the GBM model. The synthetic Likelihood method produced slightly better results than the ABC-MCMC for our data, and altering the initial guess for the Synthetic Likelihood method could potentially improve its results. However, the differences were minimal, indicating consistency between the SBI methods. Increasing the number of simulations improved the results, but achieving high accuracy required significant computational power.

3.1. ABC MCMC

The ABC Markov chain Monte Carlo algorithm, also known as ABC-MCMC, was originally proposed by Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré in 2003 [10] (According to Paul Marjoram [11])

Bayesian inference is a statistical methodology that updates beliefs about model parameters based on observed data. This is typically articulated through Bayes' Law:

$$p(\theta|\tilde{Y}) \propto p(\theta)p(\tilde{Y}|\theta)$$

where $p(\theta)$ is the prior distribution of the parameters, $p(\tilde{Y}|\theta)$ is the likelihood of observing the data given the parameters, and $p(\theta|\tilde{Y})$ is the posterior distribution, which represents the updated beliefs about the parameters after observing the data $\tilde{Y}$. [16]

We used the JAGS model in R. MCMC is used within JAGS to efficiently sample from the posterior distribution. The algorithm implemented in JAGS for Bayesian inference through MCMC starts with defining a probabilistic model where the data for the year i, $Data_i$, is modeled as normally distributed:

$$Data_i \sim \mathcal{N}(\mu + \beta \times (Year_i - 2021), \tau^{-1})$$

This formulation aligns with the Bayesian framework, where the mean of $Data_i$ is influenced by μ (baseline data for the year 2021), and β (rate of change per year), τ is the precision of the distribution, related to variance through $\tau = \sigma^{-2}$. The parameters μ, β, and σ are assigned prior distributions: $\mu \sim \mathcal{N}(mean, variance)$, $\beta \sim \mathcal{N}(mean, variance)$, and $\sigma \sim \mathcal{U}(lower, upper)$ with $\tau = \sigma^{-2}$.

Parameter initialization follows, where μ, β, and σ are either set to specific starting values or randomly generated based on their prior distributions. During the MCMC sampling phase, new parameter values are proposed using either a random walk or an autoregressive approach. The likelihood of observing the given data with these proposed parameters is calculated, and the acceptance probability for the new parameters is determined using the Metropolis-Hastings rule:

$$\alpha = \min\left(1, \frac{p(\theta_{new}|Data) \cdot q(\theta_{old}|\theta_{new})}{p(\theta_{old}|Data) \cdot q(\theta_{new}|\theta_{old})}\right)$$

Based on this probability, the algorithm decides whether to accept the new parameters or retain the old ones for the next iteration. This process repeats until the parameter values converge, as evidenced by stabilization in their distributions or using diagnostic tools.

After sufficient iterations and a burn-in period, the sampled posterior distributions of the parameters are analyzed to estimate their statistics, such as mean, median, and confidence intervals. Finally, predictive checks and model validation are conducted by generating new data based on the posterior distributions and comparing these predictions with the actual observed data to assess the model's fit.

This is our algorithm for the model:

Algorithm 1 ABC - MCMC Algorithm

Inputs:
  Data(t): $N(\mu + \beta \cdot (\text{Year}(t) - 2021), \sigma^2 t^{-1})$
  Prior distributions:
    $\mu \sim N(\text{mean}, \text{variance})$
    $\beta \sim N(\text{mean}, \text{variance})$
    $\sigma^2 \sim U(\text{lower}, \text{upper})$
Initialize:
  Initialize parameters μ, β, σ^2
for each iteration t from 1 to T **do**
    Generate proposals for μ, β, σ^2
    Calculate likelihood
    Apply Metropolis-Hastings algorithm
    Update parameters
    if convergence criteria met **then**
      Analyze posterior distribution
      Predictive check and model validation
      Generate new data
      Compare predictions
      if increase check needed **then**
        Increase simulation
      end if
    else
      Continue MCMC sampling
    end if
end for
Train Gradient Boosting Machine (GBM) model with the simulated data
End

At first, we generated only 100 new data counts for each year, and the data was significantly different from the true value. We then increased the number of simulations to 10,000 new data counts for each year in the training data, resulting in a total of 100,000 simulated data points, giving us a much better result. After that, we use these data to train a GBM Model.

We chose this method because it is useful when the likelihood function is difficult to calculate directly. It allows us to generate samples from the posterior distribution without explicitly calculating the likelihood, making it suitable for complex models. The use of MCMC ensures that we can explore the parameter space efficiently

3.2. Synthetic Likelihood Method

Our approach builds on the method established by David T. Frazier, Christopher Drovandi, David J. Nott [5], and L. F. Price, C. C. Drovandi, A. Lee, D. J. Not [14]. This approach was first mentioned by M. A. Beaumont, W. Zhang, and D. J. Balding in 2002. [1]

We use the synthetic likelihood function to compare the summary statistics of the observed data to those of the simulated data. The goal is to find parameters that make the simulated data as similar as possible to the observed data.

Instead of working with the full dataset, this method allows us to reduce the data to a set of summary statistics that capture essential information. The summary statistics are the mean and standard deviation of the observed data, which are assumed to follow an approximate Gaussian distribution. This is very important because working with full datasets can be computationally expensive and complex.

The core idea of synthetic likelihood is to assume that these summary statistics follow a multivariate normal distribution. This allows us to approximate the likelihood of the summary statistics using their mean and covariance matrix. The approximation is mathematically expressed as:

$$\mathcal{L}(\theta) \approx \mathcal{N}(\hat{S}_{\text{obs}}; \mu(\theta), \Sigma(\theta))$$

where $\hat{S}_{\text{obs}}$ are the observed summary statistics, and $\mu(\theta)$ and $\Sigma(\theta)$ are the mean and covariance of the summary statistics under the model with parameters θ.

Next, the data is simulated from the model for various parameter values to estimate the mean and covariance of the summary statistics. Our data is simulated as follows:

$$\mathbf{z}_i \sim \mathcal{N}(\mu + \beta \cdot (\text{Year}_i - 2021), \sigma^2)$$

Summary statistics for this simulated data are then calculated and compared to the observed statistics using normal density functions. If the summary statistics of the simulated data are close to those of the observed data, it indicates that the model with the current parameters is a good fit. Otherwise, the model parameters need to be adjusted. The likelihood of the observed summary statistics given the simulated ones is calculated as follows:

$$L_\mu = \mathcal{N}(\hat{\mu}_{\text{obs}}; \hat{\mu}_{\text{sim}}, \sqrt{\hat{\sigma}_{\text{sim}}})$$

$$L_\sigma = \mathcal{N}(\hat{\sigma}_{\text{obs}}; \hat{\sigma}_{\text{sim}}, \sqrt{\hat{\sigma}_{\text{sim}}})$$

The total synthetic likelihood is the product of these individual likelihoods:

$$\mathcal{L}(\theta) = L_\mu \cdot L_\sigma$$

To find the parameters that best explain the observed data, we minimize the negative log-likelihood:

$$\theta^* = \arg\min_\theta - \log(\mathcal{L}(\theta))$$

This optimization uses the `optim` function in R to explore the parameter space iteratively. Using the optimized parameters θ^*, a larger synthetic dataset is generated:

$$\mathbf{z}_i \sim \mathcal{N}(\mu^* + \beta^* \cdot (\text{Year}_i - 2021), (\sigma^*)^2)$$

The simulated dataset includes polynomial terms for the year to capture non-linear trends. We then trained the GBM model based on this simulated dataset. The trained model predicted the future year 2022.

The synthetic likelihood approach ensures that the simulated data used for training the GBM model is statistically similar to the observed data. This improves the model's predictive accuracy by providing a robust framework for parameter estimation, even when the likelihood function is not directly accessible. The synthetic likelihood method combines simulation, summary statistics, and optimization to approximate the likelihood function, making it a powerful tool for complex statistical modeling.

Initial parameters are optimized, and 10,000 simulated driver counts are generated for each year in the training data, similar to the ABC-MCMC method. This results in another 100,000 simulated data points used to train a GBM model to make predictions for the year 2022.

This is our algorithm for the model:

Algorithm 2 Synthetic Likelihood Algorithm

1: **Start**
2: Calculate summary statistics of observed data
3: Initialize parameter values
4: **while** not converged **do**
5: Simulate data from model for current parameter values
6: Calculate summary statistics for simulated data
7: Calculate synthetic likelihood
8: Adjust parameters
9: **end while**
10: Generate new data with parameters from the posterior distribution
11: Train GBM model
12: **End**

This approach is useful when we can define summary statistics that capture the essential characteristics of the data. We can ensure that our model accurately reflects the underlying data distribution by optimizing parameters to match these summary statistics.

4. Results

The findings of the study are categorically described in two sections: (i) licensed driver, (ii) vehicle

4.1. Licensed Driver

After implementing SBI techniques starting with one data point and gradually adding more, going from two data points up to ten, the table shows the results below.

Table 2: Predictive Results for Licensed Drivers for 2022

Year of Prediction	Data Used for Prediction	Predicted Value ABC - MCMC	Predicted Value Synthetic Likelihood	Actual Value (2022)	Error ABC - MCMC	Error Synthetic Likelihood
2022	2021	4,980,499	-	4,653,808	7.02%	-
2022	2020 - 2021	5,048,870	4,473,007	4,653,808	8.49%	3.89%
2022	2019 - 2021	4,965,700	4,479,193	4,653,808	6.70%	3.75%
2022	2018 - 2021	5,031,747	4,475,731	4,653,808	8.12%	3.83%
2022	2017 - 2021	4,977,580	4,476,725	4,653,808	6.96%	3.81%
2022	2016 - 2021	5,008,449	4,532,928	4,653,808	7.62%	2.60%
2022	2015 - 2021	4,997,746	4,497,615	4,653,808	7.39%	3.36%
2022	2014 - 2021	5,046,348	4,506,261	4,653,808	8.43%	3.17%
2022	2013 - 2021	4,968,533	4,482,934	4,653,808	6.76%	3.67%
2022	2012 - 2021	4,965,588	4,524,628	4,653,808	6.70%	2.78%
2022	2011 - 2021	4,967,668	4,473,072	4,653,808	6.74%	3.88%
2022	2010 - 2021	4,957,763	4,539,187	4,653,808	6.53%	2.46%

When we compared our results to a standard value, it was clear that the Synthetic Likelihood method performed better than the ABC - MCMC method. The Synthetic Likelihood method showed smaller errors, between 2.46% and 3.89%, compared to the ABC - MCMC method, which had errors between 6.53% and 8.49%. However, the Synthetic Likelihood method needs at least two data points to work.

This study also shows that both methods are effective even with minimal data. This is an encouraging aspect for situations where little data is available. We also found that adding more historical data does not always make the predictions better for either method. This suggests that the quality and relevance of the data are more important than how much data we have. Further research into how these models use historical data and the specific data types might help us improve how the models perform, possibly by focusing on newer, more relevant data if older data is less useful. However, we also observed that the error rates of both methods tend to be more consistent when using more than nine original data points.

Generating more data could significantly improve the performance of both methods. However, due to the limitations of our computing power, we are currently unable to produce additional data. In the future, improvements in modeling techniques that reduce the demand on computing resources could enable us to generate better prediction data.

Figure 4: Predicted Values by Years of Data Used with Actual Data

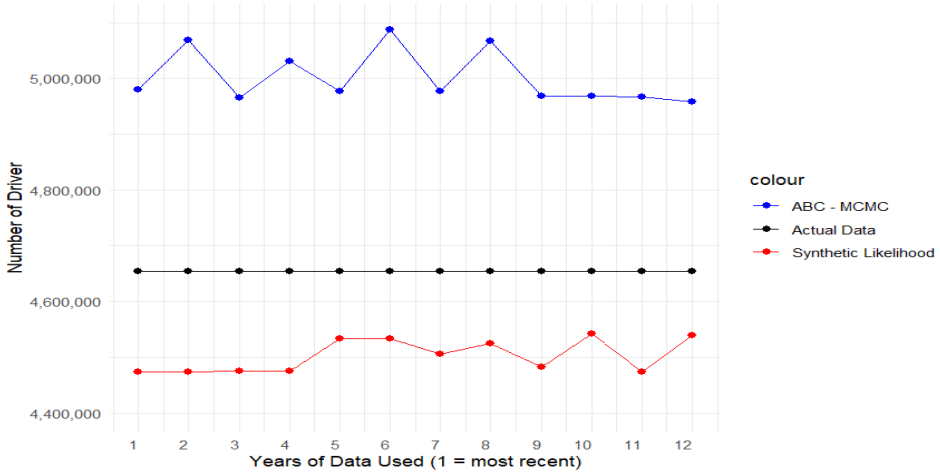
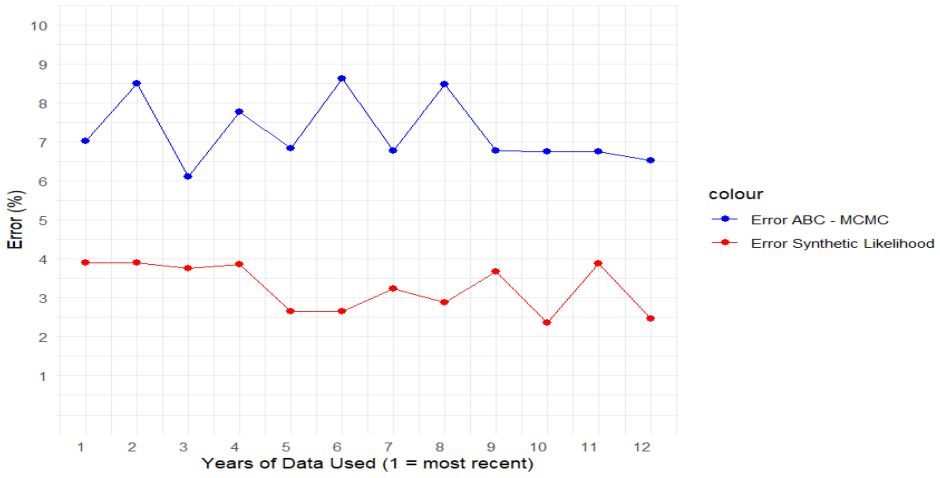


Figure 5: Error Percentages by Years of Data Used



4.2. Vehicle

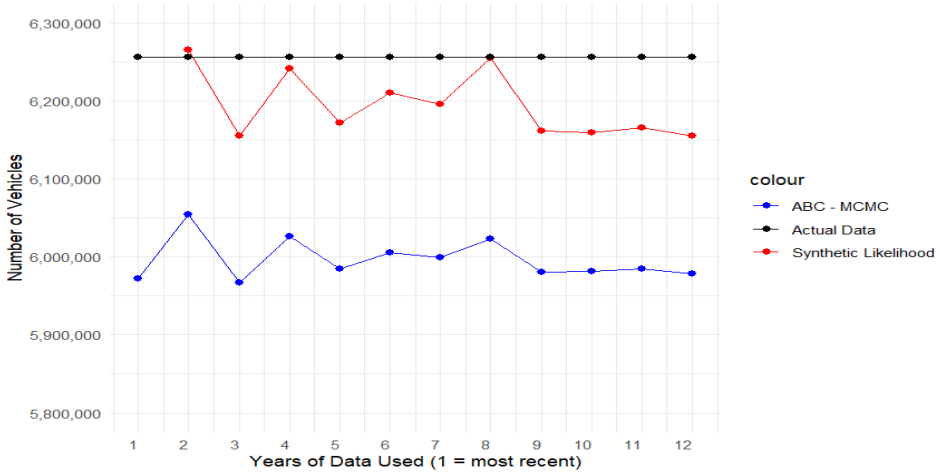
After implementing SBI techniques starting with one data point and gradually adding more, going from two data points up to ten, the table shows the results below.

Table 3: Predictive Results for Vehicles for 2022

Year of Prediction	Data Used for Prediction	Predicted Value ABC - MCMC	Predicted Value Synthetic Likelihood	Actual Value	Error ABC - MCMC	Error Synthetic Likelihood
2022	2021	5,972,366	-	6,256,479	4.54%	-
2022	2020 - 2021	6,054,325	6,265,311	6,256,479	3.23%	0.14%
2022	2019 - 2021	5,967,138	6,155,121	6,256,479	4.62%	1.62%
2022	2018 - 2021	6,025,977	6,241,047	6,256,479	3.68%	0.25%
2022	2017 - 2021	5,984,276	6,171,556	6,256,479	4.35%	1.36%
2022	2016 - 2021	6,005,393	6,210,436	6,256,479	4.01%	0.74%
2022	2015 - 2021	5,998,655	6,196,025	6,256,479	4.12%	0.97%
2022	2014 - 2021	6,022,922	6,254,881	6,256,479	3.73%	0.03%
2022	2013 - 2021	5,980,963	6,161,209	6,256,479	4.40%	1.52%
2022	2012 - 2021	5,981,432	6,159,552	6,256,479	4.40%	1.55%
2022	2011 - 2021	5,984,445	6,165,321	6,256,479	4.34%	1.45%
2022	2010 - 2021	5,978,703	6,155,011	6,256,479	4.44%	1.62%

Below is the graph:

Figure 6: Predicted Values by Years of Data Used with Actual Data

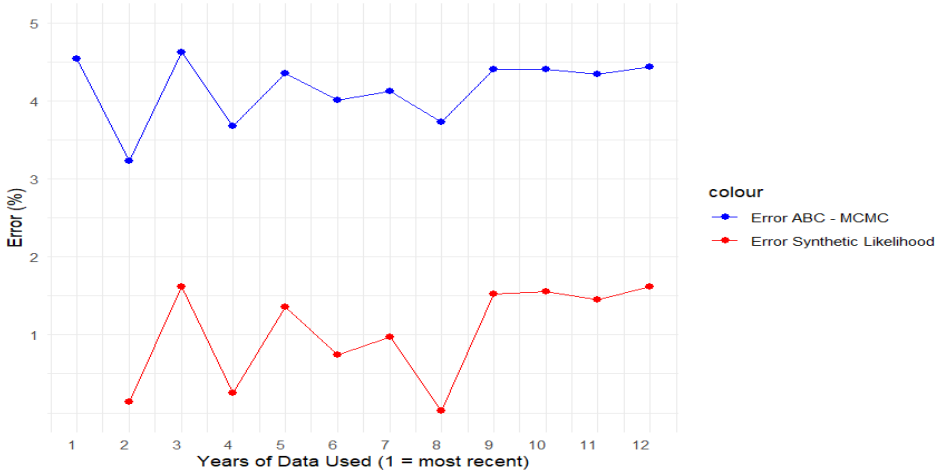


When we compared our results to an actual value, it was clear that the Synthetic Likelihood method outperformed the ABC—MCMC method. The Synthetic Likelihood method showed smaller errors, between 0.03% and 1.62%, compared to the ABC—MCMC method, which had errors between 3.23% and 4.16%. However, it is important to note that the Synthetic Likelihood method requires at least two data points to work.

Similar to the previous results, these results also show that both methods are effective even with minimal data. This is an encouraging aspect for situations where little data is available. We also found that adding more historical data does not always make the predictions better for either method. This suggests that the quality and relevance of the data are more important than how much data we have. Further research into how these

models use historical data and the specific data types might help us improve how the models perform, possibly by focusing on newer, more relevant data if older data is less useful. However, the error rates of both methods tend to be more consistent when using more than nine original data points.

Figure 7: Error Percentages by Years of Data Used



Overall, the Synthetic Likelihood method is superior for predictions in this dataset, offering very low errors. The ABC-MCMC method may require further adjustment to improve its accuracy in this dataset. Improvements in modeling techniques that reduce the demand for computing resources could enable us to generate better prediction data.

5. Advantages and Limitations

After applying two simulation-based inference (SBI) methods, ABC-MCMC and Synthetic Likelihood, combined with Gradient Boosting Machines (GBM) to Indiana transportation data, we found that both methods effectively used simulated data to capture trends and variability in historical data, leading to reliable predictions. The ABC-MCMC method estimated the distribution of model parameters, while the Synthetic Likelihood method optimized parameters to match observed data. Depending on the data, although the Synthetic Likelihood provides better results, the differences between these two methods are minimal.

These methods have several advantages. They can simulate data when we have limited or expensive data. Combining SBI methods with machine learning models (GBM) provides reliable predictions. Using simulated data enables the training of robust machine-learning models to handle large and complex datasets.

However, both methods have limitations. Both methods require significant computational resources and time. The accuracy of the predictions depends on the correctness of the model assumptions and the quality of the input data. Incorrect assumptions or poor data quality can lead to inaccurate predictions. Additionally, simulated data might not always capture real-world complexities. While real-world data would be ideal for predictions, it is often challenging to obtain in large quantities. Thus, these methods offer a viable alternative.

6. Implications for Transportation Planning

Applying Simulation-Based Inference (SBI) techniques in transportation planning offers several key benefits that can transform decision-making processes. Firstly, SBI methods allow transportation planners to generate synthetic data, enabling accurate predictions about future trends even when actual data is scarce. This capability leads to better-informed decisions regarding infrastructure investments and policy-making.

Another advantage of SBI techniques is cost efficiency. Generating synthetic data reduces the need for extensive and expensive data collection. This makes it feasible to perform detailed analyses and forecasts without incurring substantial expenses, thereby making the planning process more economical.

Using the predicted data from SBI methods, transportation planners can more effectively optimize routes, schedules, and resource allocations. This includes enhancing public transit systems, reducing carbon emissions, and promoting eco-friendly travel options. Additionally, these models help planners anticipate potential risks, leading to more resilient and reliable transportation networks. Consequently, SBI techniques contribute to more efficient transportation systems, reducing congestion and improving overall service quality.

7. Conclusion

Our research shows the significant potential of Simulation-Based Inference techniques in enhancing predictive modeling and decision-making within the transportation sector. By applying Approximate Bayesian Computation - Markov Chain Monte Carlo (ABC-MCMC) and Synthetic Likelihood methods, we effectively generated synthetic data to train machine learning models, particularly the Gradient Boosting Machine (GBM). The results show that SBI techniques are reliable in dealing with complex and uncertain data, allowing for more accurate predictions even with limited real-world data.

Our findings demonstrate that both ABC-MCMC and Synthetic Likelihood methods offer substantial benefits in predicting transportation-related metrics such as the number of licensed drivers, vehicle counts, and highway vehicle miles traveled. The differences between those methods were generally minimal, indicating the reliability and consistency of both approaches.

However, this research also highlights certain limitations, including the need for substantial computational resources and the heavy dependency on model assumptions and data quality. Despite these challenges, integrating SBI methods with machine learning models presents a solution for making informed transportation planning and management decisions when the data is limited or expensive. Adapting SBI techniques to smaller datasets can still provide valuable insights and improvements in modeling and inference, making them a viable option even when resources are constrained.

Overall, using SBI techniques and advanced machine learning models can significantly improve the accuracy and reliability of predictions in transportation systems. This approach can assist in optimizing infrastructure investments, enhancing the efficiency of transportation networks, and promoting sustainable and equitable transportation solutions. Future research could focus on combining SBI with other methods to reduce computational resource requirements and identify better assumptions, thereby increasing the accuracy of data predictions.

References

[1] Beaumont, M. A., Zhang, W., and Balding, D. J., "Approximate Bayesian computation in population genetics," *Genetics*, vol. 162, pp. 2025–2035, 2002.

[2] Beck, J., Deistler, M., Bernaerts, Y., Macke, J. H., and Berens, P., "Efficient identification of informative features in simulation-based inference," *Proc. 36th Conf. Neural Inf. Process. Syst. (NeurIPS 2022)*, pp. 1–14, 2022.

[3] Cranmer, K., Brehmer, J., and Louppe, G., "The frontier of simulation-based inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30055–30062, 2020.

[4] Diggle, P. J., and Gratton, R. J., "Monte Carlo methods of inference for implicit statistical models," *J. R. Stat. Soc. Ser. B*, vol. 46, pp. 193–212, 1984.

[5] Frazier, D. T., Drovandi, C., and Nott, D. J., "Bayesian Synthetic Likelihood," arXiv preprint arXiv:2305.05120, 2023.

[6] Gaskell, J., Campioni, N., Morales, J. M., Husmeier, D., and Torney, C. J., "Inferring the interaction rules of complex systems with graph neural networks and approximate Bayesian computation," *Journal of the Royal Society Interface*, vol. 20, 20220676, 2023.

[7] Guarda, P., and Qian, S., "Statistical inference of travelers' route choice preferences with system-level data," *Transportation Research Part B*, vol. 179, 102853, 2024. `https://doi.org/10.1016/j.trb.2023.102853`

[8] Harchaoui, Z., and Leclercq-Samuel, F., "Statistical inference for optimal transport," Unpublished manuscript, 2021.

[9] Manole, T., and Niles-Weed, J., "Statistical inference for optimal transport," *The Annals of Applied Probability*, vol. 34, no. 3, pp. 1108-1135, 2024. https://doi.org/10.1214/23-AAP1872

[10] Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S., "Markov chain Monte Carlo without likelihoods," *Proc Natl Acad Sci U S A*, vol. 100, no. 26, pp. 15324–15328, 2003, doi: 10.1073/pnas.0306899100.

[11] Marjoram, P., "Approximation Bayesian Computation," *OA Genet.*, vol. 1, no. 3, 853, 2013, doi: 10.13172/2054-197x-1-1-853.

[12] Papamakarios, G., Sterratt, D. C., and Murray, I., "Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows," arXiv:1905.07488, 2019.

[13] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B., "Normalizing Flows for Probabilistic Modeling and Inference," arXiv:1912.02762, 2019.

[14] Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J., "Bayesian synthetic likelihood," *Journal of Computational and Graphical Statistics*, vol. 27, no. 1, pp. 1–11, 2018.

[15] Rubin, D. B., "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *Ann. Stat.*, vol. 12, pp. 1151–1172, 1984.

[16] Sadegh, M., and Vrugt, J. A., "Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM(ABC)," *Water Resources Research*, vol. 50, pp. 6767–6787, 2014.

[17] Sisson, S. A., Fan, Y., and Tanaka, M. M., "Sequential Monte Carlo without likelihoods," *Proc. Natl. Acad. Sci.*, vol. 104, pp. 1760–1765, 2007.

[18] Sisson, S. A., Fan, Y., and Beaumont, M., *Handbook of Approximate Bayesian Computation*, Chapman and Hall/CRC, 2018.

[19] Unnikrishnan, A., Kochar, S., and Figliozzi, M., "Statistical inference for multimodal travel time reliability" (Final Report 1403). *Portland, OR: National Institute for Transportation and Communities (NITC)*, 2022.

[20] Simulation-based Inference. (n.d.). Retrieved from https://simulation-based-inference.org/