Bayesian Modeling of Traffic Volume using Hierarchical Structures

1 Traffic Forecast modeling

1.1 Activity-Based Model (ABM) and Its Significance

An Activity-Based Model (ABM) is an advanced simulation tool used in transportation planning. Unlike traditional four-step travel demand models, ABMs simulate individual travel behavior based on the daily activities of households and persons. The core principle of an ABM is to replicate real-world travel patterns by modeling the sequence of activities performed by individuals, considering factors such as:

- Socio-economic characteristics (e.g., age, employment, household size).
- Travel preferences and constraints (e.g., mode of transportation, trip purpose).
- Time and spatial interactions (e.g., work, school, recreational activities).

By capturing these detailed interactions, ABMs provide a more realistic representation of travel behavior and traffic dynamics.

1.1.1 The Role of the Atlanta Regional Commission (ARC)

The Atlanta Regional Commission (ARC) is a leader in transportation planning and modeling in the United States. The ARC's ABM is widely regarded as one of the best implementations of this model due to:

- **High-Quality Data Inputs:** The ARC integrates comprehensive data sources, including household travel surveys, land use data, and economic forecasts, ensuring a robust model foundation.
- **Granular Resolution:** The ABM for Atlanta captures fine-grained details at the Traffic Analysis Zone (TAZ) level, offering high spatial accuracy.
- **Dynamic Temporal Analysis:** It models travel behavior across different time periods, enabling detailed insights into peak and off-peak traffic patterns.
- **Scenario Testing:** The ARC uses its ABM to evaluate future scenarios, such as changes in land use, economic growth, and transportation investments, making it a valuable tool for policy and infrastructure planning.

These strengths make the ARC's ABM a benchmark for transportation modeling in the U.S., providing reliable and actionable insights for traffic forecasting.

1.2 The Importance of Traffic Volume Forecasting

Accurate traffic volume forecasting is critical for:

- **Infrastructure Planning:** Forecasting helps transportation planners design road networks and transit systems that can handle future demand, reducing congestion and improving efficiency.
- **Safety Improvements:** Predicting traffic patterns allows for proactive identification of high-risk areas, guiding investments in safety measures.
- **Environmental Impact Assessment:** Understanding future traffic volumes supports the evaluation of vehicle emissions and the development of sustainable transportation policies.
- **Economic Growth:** Effective traffic management fosters smoother transportation of goods and services, promoting regional economic development.

1.3 Data Resources

The data for this project is sourced from the Atlanta Regional Commission, covering the Atlanta Metropolitan area. To align with the project's focus, the dataset is aggregated to include only Fulton County. The dataset can be found in the following links:

- Download the model or other associated files: http://abmfiles.atlantaregional.com/.
- Download the model shapefile: https://opendata.atlantaregional.com/datasets/arc-model-transportation-analysis-zones-2020.
- Review the model documentation: https://atlregional.github.io/ARC_Model/index.html.

1.4 Data Preparation Using ArcGIS

To ensure the dataset was well-structured and spatially accurate, data preparation was conducted using ArcGIS. This process involved selecting, spatially joining, and exporting variables of interest for predictive modeling. The detailed steps are as follows:

2 Steps in Data Preparation

1. **Selection of Traffic Analysis Zones (TAZs):**
   - All TAZs relevant to the study area (Fulton County, GA) were selected by location in ArcMap.
   - This ensured that the spatial extent of the data matched the geographic region of interest.
2. **Spatial Join of Volume Variables:**
   - Traffic volume variables (e.g., *Sum_V_EA, Sum_V_AM, Sum_V_PM*) were spatially joined to the selected TAZs.
   - The **sum function** was applied during the spatial join to aggregate traffic volume data for each TAZ.
   - This step ensured that the total traffic volume for each TAZ was accurately calculated.

3. **Spatial Join of Speed Variables:**
   - Traffic speed variables (e.g., *Avg_SPD_EA, Avg_SPD_AM, Avg_SPD_PM*) were spatially joined to the selected TAZs.
   - The **average function** was applied during the spatial join to calculate the mean traffic speed for each TAZ.
   - This provided an aggregated measure of traffic speed, capturing the overall efficiency of traffic flow within each zone.
4. **Spatial Join of Socio-Economic Data:**
   - Node data containing socio-economic variables (e.g., *Sum_POP, Sum_EMP, Sum_HSHLD*) were spatially joined to the TAZ shapefiles.
   - This step ensured that socio-economic information was accurately linked to each TAZ, enabling a comprehensive analysis of demographic and economic factors influencing traffic.
5. **Exporting Data:**
   - All spatially joined variables were exported into Excel files for further analysis and feature engineering.
   - This step facilitated integration with predictive modeling tools and ensured a seamless transition between spatial and statistical analyses.

2.1 Significance of the Data Preparation Process

- **Spatial Accuracy:** The use of spatial joins ensures that all traffic and socio-economic data are accurately aligned with their corresponding TAZs, providing high geographic fidelity.
- **Data Aggregation:** Aggregating volume data by sum and speed data by average ensures that the dataset captures key characteristics of each TAZ while reducing noise from finer-scale variations.
- **Comprehensive Dataset:** By combining traffic volume, speed, and socio-economic data, the final dataset supports robust feature engineering and predictive modeling.
- **Seamless Integration:** Exporting the data into Excel facilitates its use in machine learning workflows, enabling efficient analysis and forecasting.

2.1.1 Outcome

The data preparation process in ArcGIS provided a geographically accurate and well-structured dataset that integrates traffic and socio-economic variables at the TAZ level. This forms the foundation for the predictive modeling scenarios described in subsequent sections.

3 Data Description

The following variables have been used in this project, as represented in the correlation chart:

3.0.1 Traffic Volume Variables

- **Sum_V_EA (Early AM Traffic Volume):** Total traffic volume during the early morning period. Used to understand traffic flow patterns during pre-peak hours.
- **Sum_V_AM (AM Traffic Volume):** Total traffic volume during the morning peak period. Reflects commuter traffic patterns and congestion during peak hours.
- **Sum_V_MD (Mid-Day Traffic Volume):** Total traffic volume during mid-day hours. Useful for understanding traffic flow in non-peak daytime periods.
- **Sum_V_PM (PM Traffic Volume):** Total traffic volume during the evening peak period. Captures traffic patterns as people return home from work.
- **Sum_V_EV (Evening Traffic Volume):** Total traffic volume during the evening hours after the peak. Reflects late-night traffic patterns.
- **Sum_V_TOTD (Total Daily Traffic Volume):** Aggregate traffic volume across all time periods in a day. Serves as a comprehensive measure of daily traffic flow.

3.0.2 Traffic Speed Variables

- **Avg_SPD_EA (Early AM Average Speed):** Average traffic speed during early morning hours. Helps evaluate road conditions and flow efficiency in pre-peak hours.
- **Avg_SPD_AM (AM Average Speed):** Average traffic speed during the morning peak period. Indicates congestion levels during peak commuting hours.
- **Avg_SPD_MD (Mid-Day Average Speed):** Average traffic speed during mid-day hours. Reflects traffic flow efficiency during non-peak daytime periods.
- **Avg_SPD_PM (PM Average Speed):** Average traffic speed during the evening peak period. Highlights congestion levels during evening commutes.
- **Avg_SPD_EV (Evening Average Speed):** Average traffic speed during evening hours. Shows road usage efficiency in late-night periods.
- **Avg_SPEED (Overall Average Speed):** Aggregate average speed across all time periods in a day. A general indicator of traffic conditions and road performance.

3.0.3 Socio-Demographic Variables

- **Sum_POP (Population):** Total population in a given Traffic Analysis Zone (TAZ). Represents the potential source of traffic generation within the zone.
- **Pop_Acres (Population Density):** Population per acre within the TAZ. Indicates the concentration of residents in the area, impacting traffic patterns.
- **Sum_HSHLD (Households):** Total number of households within the TAZ. Acts as a proxy for the residential base contributing to local traffic.
- **Sum_EMP (Employment):** Total employment in the TAZ. Represents the job density in the area, influencing commuting patterns and traffic volume.

3.0.4 Key Insights

- The dataset includes both traffic and socio-demographic factors, allowing for a comprehensive analysis of how human activity influences traffic flow and congestion.
- Temporal granularity (e.g., early AM, AM, mid-day, PM, evening) enables detailed modeling of traffic patterns across different parts of the day.
- Socio-demographic factors, such as population, employment, and households, provide essential context for understanding the spatial distribution of traffic.

These variables form the basis for predictive modeling, enabling the development of models that capture both temporal and spatial traffic dynamics effectively.

4 Data Analysis and Insights

This section provides an in-depth analysis of traffic, population, and employment data to uncover patterns and relationships that influence traffic volume and speed. The goal is to leverage these insights for enhancing predictive modeling and traffic management strategies. The analysis is structured into the following key components:

- **Correlation Analysis**: Examines the relationships between traffic-related variables, such as speed, volume, population, and employment, using a correlation heatmap. This helps identify the most influential factors driving traffic patterns.
- **Temporal Analysis**: Explores trends in traffic speed, volume, and population over time, providing insights into how these variables have evolved from 2020 to 2023.
- **Periodic Traffic Trends**: Analyzes traffic dynamics across different times of the day (e.g., AM, PM, and evening periods) and across years, highlighting consistent patterns in traffic behavior.

Objective: This section aims to:

- Identify key variables and trends influencing traffic patterns.
- Provide actionable insights for improving traffic prediction models.
- Highlight areas of focus for urban planners and traffic management authorities.

By combining correlation, temporal, and periodic analyses, this section lays the groundwork for informed decision-making in traffic modeling and congestion mitigation. Each visualization is accompanied by detailed insights, offering a comprehensive view of the data.

4.1 Correlation Analysis

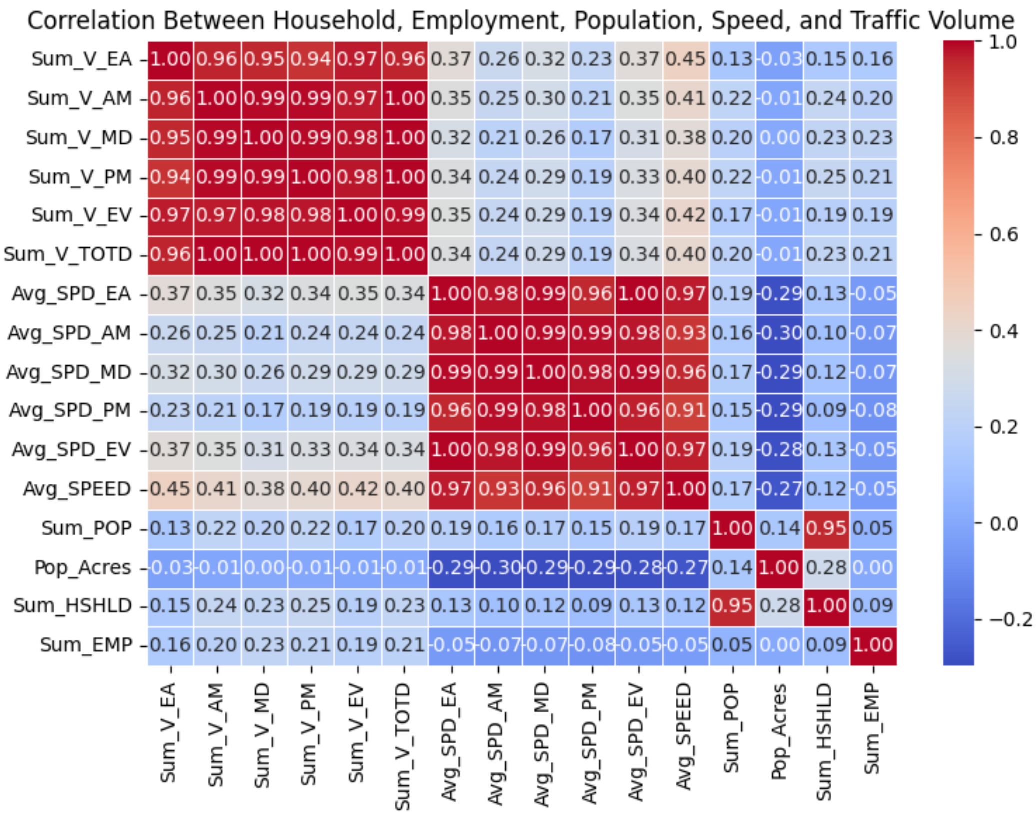

Figure 1: Correlation Between Household, Employment, Population, Speed, and Traffic Volume

The correlation heatmap highlights the relationships between traffic volume, speed, population, employment, and other variables. Key findings include:

- **Traffic Volume Correlations**: Strong positive correlations exist between total traffic volume (*Sum_V_TOTD*) and individual time-period volumes (e.g., *Sum_V_AM, Sum_V_PM*), indicating consistent traffic patterns across periods.
- **Speed Relationships**: Traffic speed shows weak negative correlations with population and employment, suggesting that denser areas tend to experience slower traffic speeds.
- **Population and Employment**: Population and employment densities are moderately correlated with traffic volume, underlining their role as key determinants in traffic forecasting.

This smaller correlation heatmap focuses on key variables, including total traffic volume (*Sum_V_TOTD*), average speed (*Avg_SPEED*), population (*Sum_POP*), population density (*Pop_Acres*), household counts (*Sum_HSHLD*), and employment (*Sum_EMP*). Key insights from this chart include:

- **Traffic Volume and Speed**: A moderate positive correlation ($r = 0.42$) exists between traffic volume and average speed, suggesting that higher speeds are often associated with greater traffic flow in certain conditions.

Figure 2: Correlation Between Household, Employment, Population, Speed, and Traffic Volume

- **Population and Household Counts**: There is a strong positive correlation between population (*Sum_POP*) and household counts (*Sum_HSHLD*), with $r = 0.95$. This indicates that areas with larger populations tend to have more households, which likely contribute to higher traffic volumes.

- **Employment and Traffic Volume**: Employment (*Sum_EMP*) is positively correlated with traffic volume ($r = 0.22$), reflecting the impact of job locations on daily commuting patterns.

- **Population Density and Speed**: Population density (*Pop_Acres*) exhibits a weak negative correlation with average speed ($r = -0.27$), suggesting that denser areas tend to experience slower traffic speeds, likely due to congestion or urban infrastructure constraints.

- **Traffic Volume and Households/Population**: Total traffic volume correlates positively with household counts ($r = 0.23$) and population ($r = 0.20$), emphasizing their role in traffic generation.

This chart provides a focused view of the relationships between essential traffic, demographic, and socio-economic variables, helping to identify the key drivers of traffic patterns. These insights can guide feature selection and model optimization in predictive modeling tasks.

4.2 Temporal Analysis of Key Variables

The temporal analysis reveals the following trends:

- **Daily Traffic Speed**: Average daily speeds remain relatively stable over the years, with a slight decrease from 2020 to 2023, possibly reflecting increasing urbanization and traffic congestion.

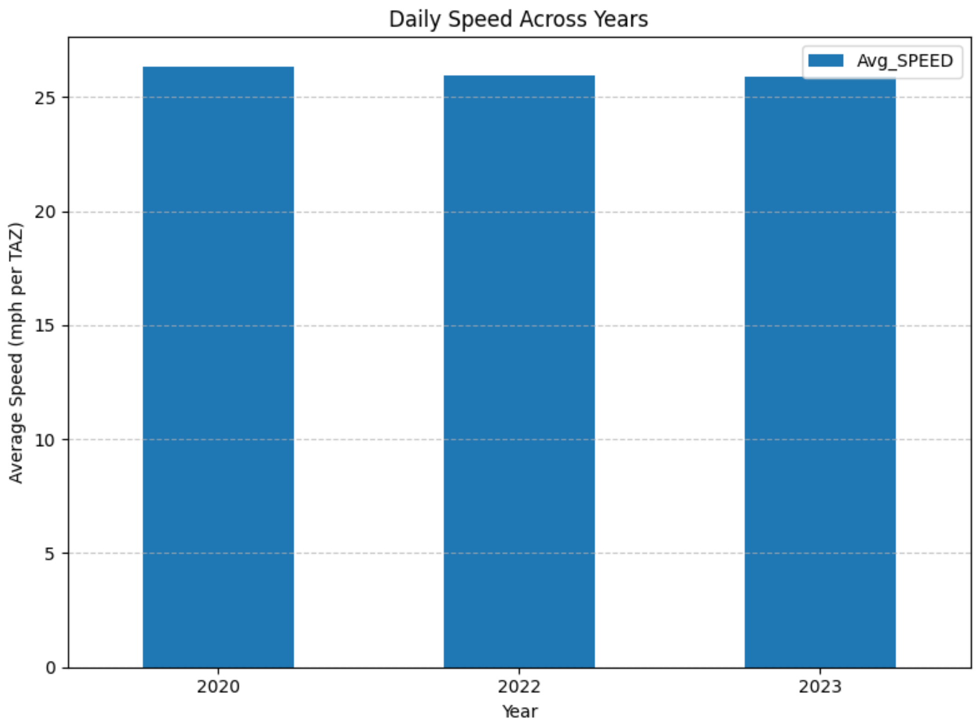

Figure 3: Daily Speed Across Years (2020, 2022, 2023)

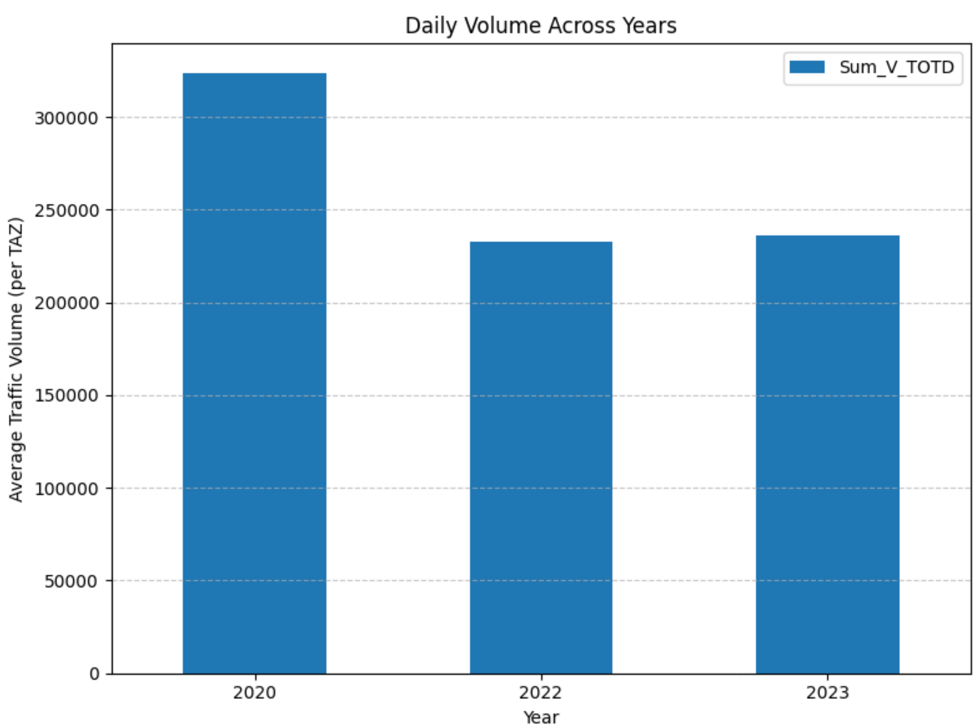

Figure 4: Daily Traffic Volume Across Years (2020, 2022, 2023)

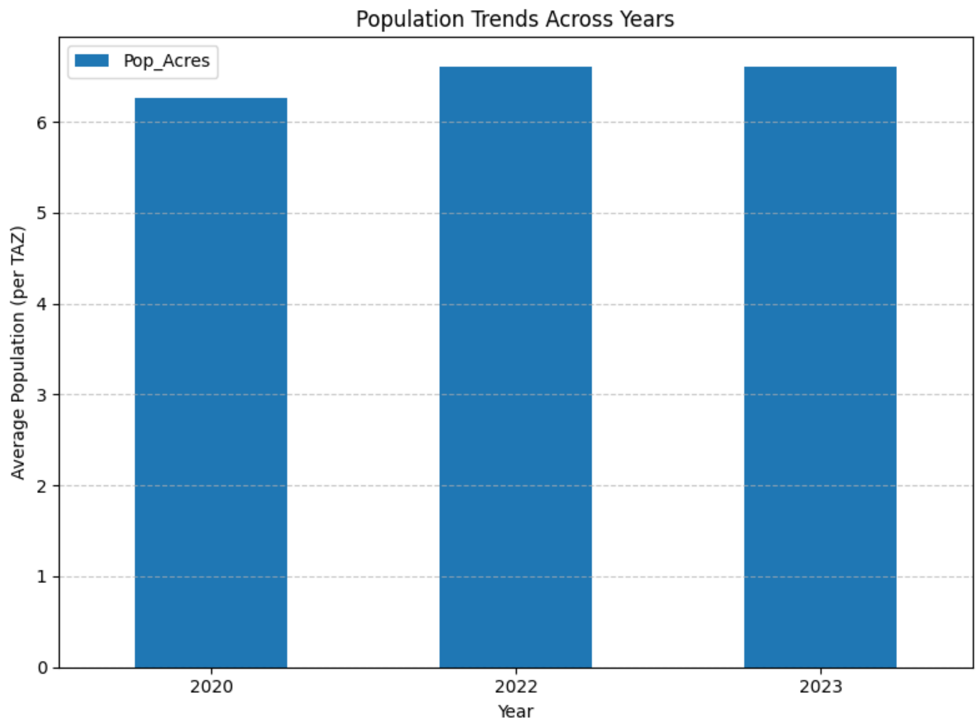

Figure 5: Population Trends Across Years (2020, 2022, 2023)

- **Daily Traffic Volume**: Traffic volumes have seen minor increases over the years, reflecting gradual population growth and economic activity.
- **Population Trends**: Population density, measured per TAZ (Traffic Analysis Zone), shows consistent growth over the years, aligning with increasing traffic volumes.

4.3 Periodic Traffic Trends

Analysis of periodic traffic trends shows:

- **Traffic Speeds**: Speeds are highest during early mornings and evenings, with a significant dip during peak traffic hours (AM and PM). This pattern is consistent across all analyzed years (2020, 2022, 2023).
- **Traffic Volumes**: Volumes peak during the AM and PM periods, reflecting commuting patterns. These trends are consistent over time, emphasizing the importance of these periods in traffic management.

4.4 Key Takeaways

- The data highlights the importance of time-period-specific variables, such as speeds and volumes, in predicting traffic dynamics.
- Increasing population and employment densities have a direct impact on traffic patterns, necessitating proactive traffic management strategies.

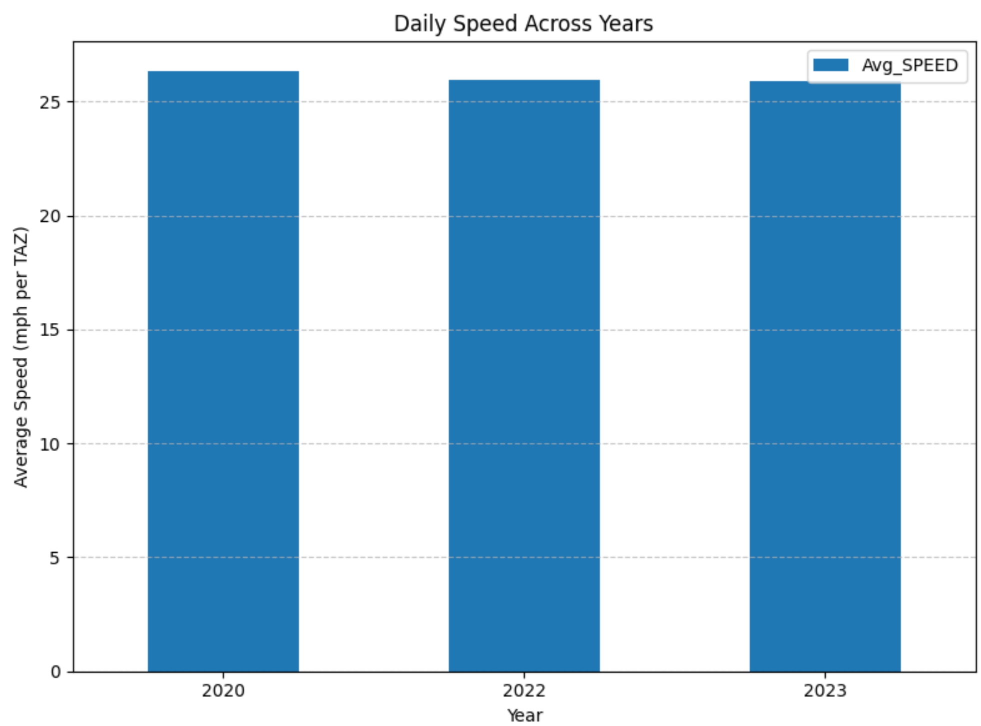

Figure 6: Traffic Speed Trends Across Periods and Years

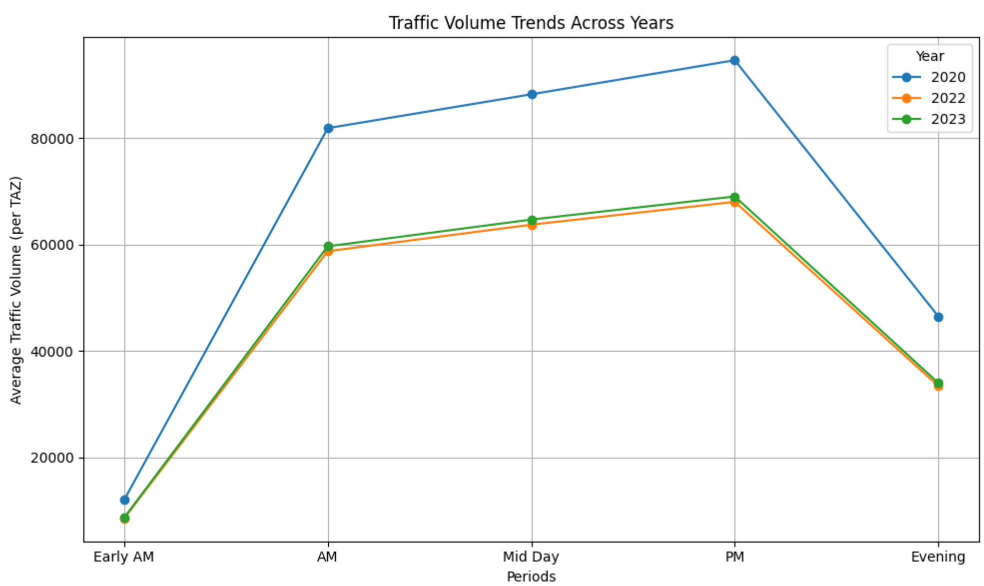

Figure 7: Traffic Volume Trends Across Periods and Years

- Future predictive models should incorporate both temporal and spatial factors to capture the nuanced relationships between these variables.

5 Feature Engineering

Feature engineering is a critical step in preparing data for predictive modeling. For this project, the following steps were implemented to create a robust dataset using Activity-Based Model (ABM) outputs from the Atlanta Regional Commission (ARC) for the years 2020, 2022, and 2023. The key feature engineering steps and their importance are detailed below:

5.1 Steps in Feature Engineering

1. **Data Combination:** Data from three years (*2020, 2022, 2023*) were merged into a single dataset, ensuring consistent columns of interest such as traffic volume, speed, and socio-demographic variables (e.g., population, household density).

2. **Handling Missing Data:** Missing values were filled with column means to ensure the dataset remained complete for modeling purposes, minimizing the risk of bias due to missing information.

3. **Creation of Lag Features:** Lagged features were generated for key variables, including traffic volume and speed. For example:

   - *Sum_V_EA_lag1* represents the traffic volume for the early morning of the previous time step.
   - Lag periods of 1, 2, and 3 time steps were applied to capture temporal dependencies and enhance model predictions.

4. **Adding Time-Based Variables:** New features such as *hour of the day, day of the week*, and *is_weekend* were created from datetime information. These variables provide contextual understanding of temporal patterns in traffic behavior.

5. **Normalization of Data:** All numerical variables were normalized using MinMaxScaler to ensure consistent scaling across features. This step prevents model bias towards variables with larger scales.

5.2 Importance of Feature Engineering in Scenarios

- **Scenario Relevance:** The feature engineering process enables the models to leverage lagged and time-based variables, which are critical in scenarios focusing on temporal patterns (e.g., *Scenario 2: Enhanced Feature Engineering* and *Scenario 3: Advanced Feature Engineering*).

- **Model Accuracy:** Lag features enhance the ability of models such as LSTM to capture sequential dependencies, while Random Forest benefits from the added explanatory variables for non-linear relationships.

- **Temporal and Spatial Insights:** Time-based features provide the models with additional context to understand variations across hours, days, and weekends, which is crucial for scenarios with dynamic traffic patterns.

- **Improved Forecasting:** By normalizing the data and incorporating comprehensive temporal features, the models are better equipped to provide accurate forecasts for 2024, leveraging trends and patterns from the historical ABM outputs.

5.3 Outcome

The final feature-engineered dataset integrates historical traffic and socio-demographic data with lagged and time-based variables, creating a robust foundation for predictive modeling. This ensures that all four scenarios leverage the full potential of the ABM outputs, enabling precise traffic volume forecasts.

6 Bayesian Modeling

This section provides an in-depth explanation of the three Bayesian modeling strategies applied to our traffic volume dataset: the pooled model, unpooled model, and hierarchical model. Each method is described in terms of its structure, assumptions, strengths, and potential limitations with respect to our data. The results for each model are also presented in tabular form for comparison.

6.1 Pooled Model

In the pooled model, we assume a single set of parameters shared across all groups (MTAZ zones). This model treats the data as if it comes from one homogeneous population and does not distinguish between individual spatial zones.

Model Structure:

$$Y_i \sim \mathcal{N}(\alpha + X_i \cdot \beta, \sigma)$$

Advantages:

- Simple and computationally efficient.
- Suitable when data is truly homogeneous or group differences are minimal.

Limitations:

- Ignores heterogeneity between MTAZs.
- May lead to biased or oversimplified inferences.

Results:

Table 1: Pooled Model Parameters

Parameter	Value
α	-0.000
$\beta_{\text{Avg\_SPEED}}$	0.000
$\beta_{\text{Pop\_Acres}}$	0.000
$\beta_{\text{Sum\_POP}}$	-0.005
$\beta_{\text{Sum\_V\_AM}}$	0.286
$\beta_{\text{Sum\_V\_PM}}$	0.274
$\beta_{\text{Sum\_V\_EV}}$	0.184
$\beta_{\text{Sum\_V\_MD}}$	0.258

6.2 Unpooled Model

The unpooled model fits a separate set of coefficients for each group (MTAZ), without sharing information across groups.

Model Structure:

$$Y_{ij} \sim \mathcal{N}(\alpha_j + X_{ij} \cdot \beta_j, \sigma)$$

Advantages:

- Captures individual group-level dynamics.
- More flexible for modeling heterogeneous regions.

Limitations:

- High variance in parameter estimates, especially with small sample sizes per group.
- Overfitting risk.

Results:

Table 2: Unpooled Model Parameters

Groups	α	$\beta_{\text{Avg\_SPEED}}$	$\beta_{\text{Pop\_Acres}}$	$\beta_{\text{Sum\_POP}}$	$\beta_{\text{Sum\_V\_AM}}$	$\beta_{\text{Sum\_V\_PM}}$	$\beta_{\text{Sum\_V\_EV}}$	$\beta_{\text{Sum\_V\_MD}}$
0	0.285	-0.405	-0.097	-0.235	0.645	0.006	0.295	0.228
1	-0.317	-0.070	0.555	0.872	0.268	0.112	0.227	0.279
2	0.443	0.346	-0.546	0.032	0.487	0.988	-0.012	-1.400
3	0.011	-0.396	0.348	0.192	-0.033	0.567	0.072	0.122
4	0.224	0.000	-0.224	-0.935	0.058	-0.748	0.220	0.321
5	-0.510	-0.060	0.570	-0.181	-0.427	0.657	0.317	0.530
6	0.086	0.000	-0.086	-0.715	-0.223	-0.173	-0.516	-0.178
7	-0.350	-0.000	0.350	0.309	0.575	-1.294	0.148	-0.084
8	-0.430	0.686	0.016	-0.599	0.658	-0.371	0.311	-0.110
9	-0.253	0.000	0.253	1.141	0.210	-0.027	-0.339	-1.916

6.3 Hierarchical (Multilevel) Model

The hierarchical model combines the strengths of the previous two models by allowing partial pooling. Parameters are allowed to vary by group, but are also regularized by hyperpriors, which borrow strength from the entire dataset.

Model Structure:

$$\begin{aligned}\alpha_j &\sim \mathcal{N}(\mu_\alpha, \sigma_\alpha)\\ \beta_j &\sim \mathcal{N}(\mu_\beta, \sigma_\beta)\\ Y_{ij} &\sim \mathcal{N}(\alpha_j + X_{ij} \cdot \beta_j, \sigma)\end{aligned}$$

Advantages:

- Balances flexibility with regularization.
- Reduces overfitting while modeling heterogeneity.

- Ideal for grouped data with varying sample sizes.

Limitations:

- Computationally intensive.
- Requires careful prior specification.

Results:

Table 3: Hierarchical Model Summary

Parameter	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
mu_alpha	0.040	0.291	-0.525	0.550	0.010	0.007	883.0	788.0	NaN
mu_beta[Avg_SPEED]	0.043	0.318	-0.582	0.576	0.010	0.007	1041.0	916.0	NaN
mu_beta[Pop_Acres]	0.023	0.325	-0.620	0.605	0.010	0.007	962.0	983.0	NaN
mu_beta[Sum_POP]	0.036	0.356	-0.644	0.670	0.012	0.008	934.0	875.0	NaN
mu_beta[Sum_V_AM]	0.040	0.364	-0.637	0.705	0.012	0.009	897.0	869.0	NaN
mu_beta[Sum_V_PM]	0.052	0.352	-0.612	0.722	0.012	0.009	888.0	868.0	NaN
mu_beta[Sum_V_EV]	0.047	0.351	-0.591	0.717	0.011	0.007	1035.0	1023.0	NaN
mu_beta[Sum_V_MD]	0.040	0.358	-0.620	0.673	0.012	0.008	915.0	868.0	NaN
sigma	1.033	0.223	0.659	1.448	0.007	0.006	967.0	741.0	NaN

6.4 Model Comparison (ELPD)

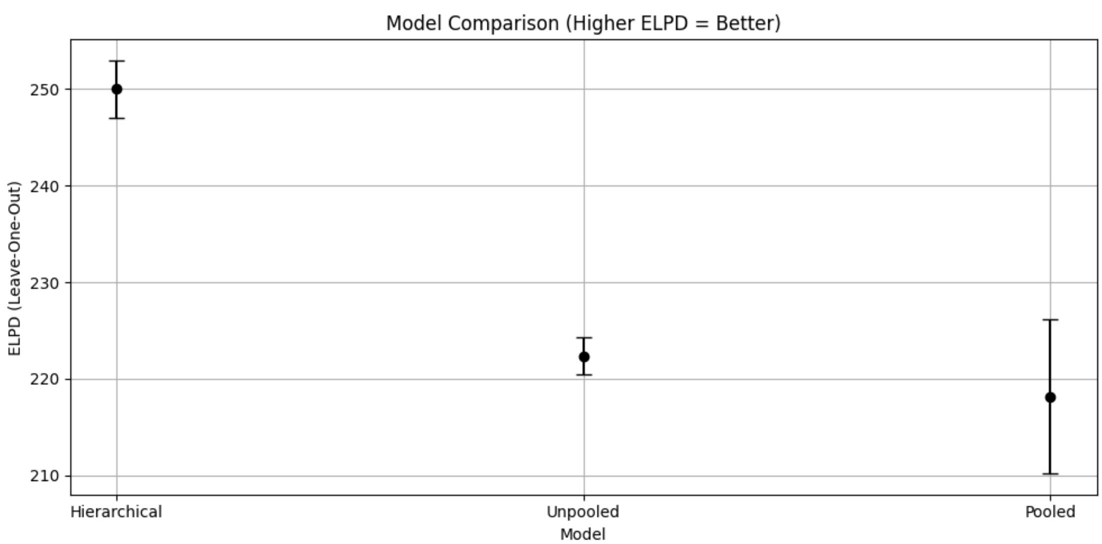

Figure 8: ELPD-Based Model Comparison: Higher ELPD Indicates Better Predictive Performance

Figure 8 visualizes the comparative ELPD values across the three models:

- **Hierarchical Model** achieved the highest ELPD, indicating the best overall generalization. It effectively captures both within-group (TAZ-level) and between-group variability.
- **Unpooled Model** performed better than the pooled model by allowing group-specific parameters, but showed signs of overfitting due to lack of regularization across TAZs.

- **Pooled Model** yielded the lowest ELPD, reflecting its inability to account for group-level differences, which is a limitation when modeling data with spatial segmentation like TAZs.

Table 4: Leave-One-Out Cross-Validation Comparison

Model	`elpd_loo`	`p_loo`	`elpd_diff`	`se`	`dse`	**Rank**
Hierarchical	250.00	35.88	0.00	2.99	0.00	1
Unpooled	222.34	26.19	27.66	1.93	2.05	2
Pooled	218.13	13.51	31.87	7.97	8.16	3

The differences in ELPD between models are further quantified in Table 4, where the `elpd_diff` values represent the expected difference in predictive accuracy compared to the top model. The small standard error (`se`) for the hierarchical model confirms its robust performance.

These results confirm that the hierarchical model provides the best trade-off between model complexity and generalization, while the pooled model oversimplifies and the unpooled model risks overfitting.

6.5 Discussion

The comparison of three Bayesian regression strategies—Pooled, Unpooled, and Hierarchical—demonstrates the benefits of incorporating group-level structure in traffic prediction tasks.

The **Pooled model** is the simplest, estimating a single set of coefficients for all TAZs. While easy to interpret, this model assumes homogeneity across groups, leading to poor performance (lowest ELPD score). It fails to account for structural differences in traffic volume determinants across TAZs.

The **Unpooled model** improves upon this by allowing each TAZ to have its own regression parameters. However, this approach can lead to overfitting, especially in groups with fewer data points. The moderately improved ELPD score reflects this tradeoff—more flexibility, but less regularization.

The **Hierarchical model**, which allows for partial pooling via shared hyperpriors, balances flexibility and regularization. It permits group-level variations while borrowing strength across all TAZs, leading to the highest ELPD and the most reliable performance. This is particularly important for traffic datasets, where some TAZs have sparse data.

Despite the hierarchical model's strengths, it is computationally more intensive. However, given its superior generalization, the trade-off is justified for policy and planning applications that demand robust inference.

7 Conclusion

This analysis demonstrates that Bayesian hierarchical modeling offers significant advantages for forecasting traffic volumes in spatially segmented datasets. The Hierarchical model outperforms both the Pooled and Unpooled models based on ELPD, indicating superior predictive accuracy and generalizability.

While the Pooled model is computationally simple, its assumption of homogeneity across zones limits its applicability. The Unpooled model captures group-specific behavior but lacks the regularization needed to generalize well, particularly in data-scarce groups.

The Hierarchical model offers the best of both worlds by enabling group-specific coefficients while maintaining stability through shared hyperpriors. It is particularly suited to transportation planning problems, where regional heterogeneity exists but data across groups may be imbalanced.

In practice, this model supports more nuanced and confident decision-making across geographic regions. Future work could extend this framework to dynamic models incorporating time trends or integrating additional socio-economic covariates.

Overall, the hierarchical Bayesian approach stands out as a robust, interpretable, and high-performing tool for modeling real-world traffic data.