MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

MULTI-SCALE FORECASTING OF CITI BIKE DEMAND IN NEW YORK CITY: INCLUDING TIME SERIES, MACHINE LEARNING, AND DEEP LEARNING MODELS

ZOLTAN EDWARD JELOVICH

JANUARY - 2026

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

MULTI-SCALE FORECASTING OF CITI BIKE DEMAND IN NEW YORK CITY: INCLUDING TIME SERIES, MACHINE LEARNING, AND DEEP LEARNING MODELS

ZOLTAN EDWARD JELOVICH

SUPERVISION:

PROFESSOR CARLOS J. COSTA

JANUARY - 2026

Glossary

ACF: Autocorrelation Function.

ACS: American Community Survey.

AI: Artificial Intelligence.

AIC: Akaike Information Criterion.

AQI: Air Quality Index.

ARIMA: Autoregressive Integrated Moving Average.

ARIMAX: Autoregressive Integrated Moving Average with Exogenous Variables.

BIC: Bayesian Information Criterion.

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies.

CH: Calinski–Harabasz.

CNN: Convolutional Neural Network.

CNN-LSTM: Convolutional Neural Network–Long Short-Term Memory.

CRS: Coordinate Reference System.

DB: Davies–Bouldin.

EDA: Exploratory Data Analysis.

ETS: Exponential Smoothing.

GBRT: Gradient Boosted Regression Tree.

GMM: Gaussian Mixture Model.

GNN: Graph Neural Network.

HGB: Histogram-Based Gradient Boosting.

LSTM: Long Short-Term Memory.

MAE: Mean Absolute Error.

MAPE: Mean Absolute Percentage Error.

NOAA: National Oceanic and Atmospheric Administration.

NYC: New York City.

OLS: Ordinary Least Squares.

PLUTO: Primary Land Use Tax Lot Output.

RMSE: Root Mean Squared Error.

SARIMA: Seasonal Autoregressive Integrated Moving Average.

SARIMAX: Seasonal Autoregressive Integrated Moving Average with Exogenous Variables.

SHAP: Shapley Additive Explanations.

STGA-LSTM: Spatial-Temporal Graph Attentional–Long Short-Term Memory.

STGCN: Spatio-Temporal Graph Convolutional Network.

WCSS: Within-Cluster Sum of Squares.

ABSTRACT

A previsão da procura em sistemas de bicicletas partilhadas é fundamental para a eficiência operacional e o planeamento dos sistemas urbanos de micromobilidade, uma vez que a qualidade do serviço depende da capacidade de antecipar quando e onde as bicicletas serão necessárias. Esta dissertação aborda a seguinte questão de investigação: Com que precisão pode ser prevista a procura em sistemas de bicicletas partilhadas e quais os fatores que mais influenciam o desempenho e a interpretabilidade dos modelos? Utilizando os registos completos das viagens do Citi Bike entre 2017 e 2024 na cidade de Nova Iorque, complementados com dados meteorológicos, feriados, qualidade do ar, contagens de mobilidade e características do ambiente construído, o estudo avalia a precisão das previsões em diferentes resoluções espaciais (sistema, agrupamento funcional e estação) e temporais (diária e horária). Recorreu-se a um fluxo de trabalho concebido para evitar *data leakage*, comparando modelos clássicos de séries temporais (ETS e SARIMAX, com e sem variáveis exógenas, e Prophet), métodos de aprendizagem supervisionada baseados em árvores de decisão (Random Forest e XGBoost) e um modelo CNN-LSTM ao nível da estação. Os modelos são avaliados em previsões de um passo em frente, utilizando divisões cronológicas fixas (treino: 2017–2021; validação: 2022; teste: 2023–2024) e métricas de erro padronizadas. Os resultados demonstram que a qualidade das previsões varia em função da resolução temporal e do nível de agregação, não existindo uma única família de modelos que apresente o melhor desempenho em todas as tarefas. Para a procura diária ao nível do sistema, o modelo ETS(A,N,A), com sazonalidade semanal e variáveis exógenas, constitui o modelo de referência com melhor desempenho no período de teste de 2023–2024, sendo seguido de muito perto pelo SARIMAX com o mesmo conjunto de covariáveis, enquanto o Prophet apresenta resultados inferiores. As variáveis meteorológicas são responsáveis pela maior parte dos ganhos incrementais proporcionados pela inclusão de variáveis exógenas. Para a procura horária ao nível do sistema, o Random Forest, com características autorregressivas e variáveis de calendário e meteorologia cuidadosamente construídas, supera o desempenho do XGBoost. A decomposição da procura por agrupamentos funcionais de estações melhora o desempenho relativo dos modelos de aprendizagem supervisionada, tanto em horizontes de previsão horários como diários. Ao nível da estação, o modelo CNN-LSTM reduz o erro absoluto médio (MAE) relativamente a um modelo de referência tabular, mas aumenta a raiz do erro quadrático médio (RMSE), indicando maior sensibilidade a erros de grande magnitude em contextos de elevada esparsidade. As análises de interpretabilidade corroboram estes resultados, identificando a precipitação, os feriados e a temperatura como os principais fatores explicativos da procura, uma vez capturada a sazonalidade semanal.

Keywords: Bike-share demand forecasting; Citi Bike; exponential smoothing and SARIMAX; Random Forest and XGBoost; CNN-LSTM station forecasting; weather and calendar covariates.

JEL Codes: C22; C38; C45; C52; C53; R41.

TABLE OF CONTENTS

TABLE OF FIGURES

Acknowledgments

First, I would like to thank Professor Carlos J. Costa for his guidance, support, and valuable insights throughout this project. His feedback and encouragement were instrumental in shaping this work.

I am also grateful to my family and friends for their constant encouragement, patience, and understanding throughout this process. Their support made the challenges along the way far more manageable.

Finally, I would like to express my thanks to my grandpa. His belief in me, constant support, and motivation made this journey possible, and I would not be where I am today without him.

1. Introduction

Forecasting demand in bike-share systems is essential for sustainable urban transport planning and operations. As cities seek to reduce emissions, alleviate congestion, and support cleaner mobility, reliable estimates of when and where shared bicycles are needed affect service efficiency and user satisfaction. This dissertation addresses the following research question: How accurately can bike-share demand be forecast, and what factors most influence model performance and interpretability?

Bike sharing is a key component of micromobility networks that can connect users to public transport and reduce dependence on private vehicles. Citi Bike, launched in New York City (NYC) in 2013, is North America's largest station-based program and provides a suitable case for demand forecasting. Its high trip volumes, open-access data, and operational structure support the study of short- and longer-term cycling patterns. Accurate demand forecasts at system, cluster, and station levels support rebalancing, infrastructure planning, and resource allocation. Improved predictions can reduce idle capacity, lower operational costs, and increase availability of bicycles and docks at times and locations with the greatest need.

Demand forecasting also has policy relevance. Hamad et al. (2021) observe that imbalances between high- and low-demand stations reduce user satisfaction and create inefficiencies in rebalancing. Forecasts that capture temporal and spatial variation can support proactive redistribution to reduce shortages and surpluses. For policymakers, these insights can inform infrastructure design aligned with equity and environmental objectives. From an academic perspective, the topic contributes to research on urban analytics and sustainable mobility by examining the interface between time series methods and machine learning (Albuquerque et al., 2021). Beyond academia, the work aims to support operator decision-making and provide evidence for infrastructure investment and multimodal integration, including linkages between cycling and public transport (Wei and Zhu, 2023).

Research on bike-share demand forecasting has progressed from classical statistical approaches to machine learning and deep learning. Chapter 2 develops this trajectory in detail and identifies gaps relating to interpretable modeling, multi-scale analysis across system, cluster, and station levels, and the integration of contextual variables. This dissertation builds on those findings by applying and comparing statistical, machine learning, and deep learning methods to Citi Bike data, evaluating performance and interpretability across aggregation levels under a single empirical framework.

The objective of this dissertation is to develop and evaluate a multi-scale forecasting framework for Citi Bike demand in NYC that compares classical time series, machine learning, and deep learning models under a leakage-controlled, time-ordered evaluation, and that assesses how temporal resolution, spatial aggregation, and contextual covariates affect predictive performance and interpretability.

The specific objectives are:

- Construct daily and hourly demand series at system, cluster, and station levels using complete Citi Bike trip records integrated with contextual datasets.
- Implement and compare ETS, Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), and Prophet baselines with and without exogenous regressors where appropriate.
- Implement and compare supervised learning models (Random Forest and XGBoost) using engineered autoregressive, calendar, and contextual features.
- Develop and evaluate a station-level CNN-LSTM model for one-step-ahead hourly demand forecasting and compare it with a tabular baseline.
- Evaluate models using consistent time-ordered splits and standard error metrics, and assess interpretability using model-aligned explainability procedures.

This research adopts an empirical quantitative design emphasizing transparency and reproducibility. Citi Bike trip records for NYC from 2017 to 2024 are integrated with meteorological, calendar, environmental, mobility, and built-environment datasets. Weather data from multiple National Oceanic and Atmospheric Administration (NOAA) stations are standardized into unified explanatory series, while calendar effects are captured through a consistent U.S. federal holiday series and environmental conditions through Air Quality Index (AQI) measures. Mobility context is represented using bicycle and traffic counts aligned to sub-daily demand. Built-environment and neighborhood context are captured through tract-level American Community Survey (ACS) socio-demographics, parcel-level Primary Land Use Tax Lot Output (PLUTO) land use, and cycling network data, spatially linked to stations. All datasets are cleaned, standardized, and merged to support daily and hourly forecasting at the system, cluster, and station levels.

The analytical framework applies complementary approaches aligned with the multi-resolution design. Classical models (ETS, SARIMAX, and Prophet) capture trend and seasonal dependencies and provide interpretable baselines with and without exogenous inputs. Random Forest and XGBoost represent ensemble methods designed to model nonlinear multivariate relationships using engineered temporal and contextual features. A CNN-LSTM model supports station-level short-term forecasting by combining convolutional layers for short-horizon pattern extraction with LSTM units for temporal dependence, while incorporating station identity and static station context within a unified design.

Models are evaluated out of sample using fixed time-ordered splits that preserve temporal ordering and prevent leakage: training (2017–2021), validation (2022), and testing (2023–2024). The primary setting is one-step-ahead forecasting: one day ahead for daily models and one hour ahead for hourly models. Hyperparameters are selected using the fixed validation period under the split design. Accuracy is assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), with R^2 reported where appropriate in comparative summaries. Forecasts are evaluated at system and cluster levels for daily and hourly aggregation, and primarily at station level for hourly forecasting, with percentage-based errors interpreted conservatively where demand can be sparse.

To represent spatial heterogeneity while preserving interpretability, stations are grouped into residential, commercial, and transit-adjacent clusters using k-means clustering based on static socio-demographic and built-environment characteristics linked to station locations. These clusters define cluster-level demand series and support interpretation of how forecasting behavior varies across functional station contexts.

This dissertation contributes methodologically and practically. It evaluates classical time series forecasting and modern machine learning under a single empirical framework using a consistent, leakage-controlled design. By comparing ETS, SARIMAX, Prophet, Random Forest, XGBoost, and CNN-LSTM under shared temporal splits and error metrics, it characterizes trade-offs between interpretability and predictive performance across aggregation levels. The multi-scale approach supports analysis across system, cluster, and station levels and evaluates how predictability changes with spatial aggregation. The study also integrates weather, holiday structure, environmental conditions, mobility context measures, and tract-based socio-demographic and land-use context within a unified data preparation and feature engineering workflow.

The dissertation advances explainable machine learning through model-aligned interpretability. SHAP is used to attribute predictions in tree-based ensembles to feature inputs, while station-level deep learning interpretability is assessed using structured feature-group sensitivity analysis. Practically, the study provides evidence relevant to Citi Bike operations and urban planning: more accurate short-term forecasts can support redistribution and fleet management, while daily forecasts and cross-model comparisons can inform planning and evaluation of operational strategies (Mohseni Hosseinabadi et al., 2024). The findings also have relevance for smart city policy by illustrating how data-driven analytics can support sustainability, efficiency, and multimodal integration (Wei and Zhu, 2023; Gong et al., 2024).

The dissertation is structured as follows. Chapter 2 reviews peer-reviewed studies organized around time series modeling, machine learning and deep learning forecasting, and contextual analyses of bike-share systems. Chapter 3 details the methodology, including data sources, preprocessing, feature engineering, model configurations, evaluation design, and explainability procedures. Chapter 4 reports and discusses empirical findings across models and aggregation levels, interpreting key drivers and robustness limitations evidenced by results. Chapter 5 concludes by synthesizing findings and presenting operational and policy implications, limitations, and evidence-grounded directions for future work. Due to page-length constraints, all appendices (Appendices A–F), including supplementary figures, tables, code, and extended methodological details, are provided in an external repository. The materials are publicly available at: https://github.com/zoltanjelovich/citi-bike-demand-forecasting. All in-text references to appendices refer to this repository.

2. Literature Review

Bike-share systems have become an essential component of modern urban mobility, improving accessibility and supporting sustainability goals across many cities worldwide (Torres et al., 2024). As these systems expand, including large-scale operations such as Citi Bike in NYC, researchers increasingly focus on understanding and forecasting demand to improve resource allocation, operational planning, and user satisfaction (Wei and Zhu, 2023). Forecasting is complex because ridership reflects interacting spatial and temporal influences, including weather, temporal rhythms, and demographic characteristics (Gong et al., 2024; Adrianzen, 2025). This chapter reviews analytical approaches for predicting bike-share demand, tracing the shift from classical statistical methods to machine learning and hybrid frameworks, and identifies trends and limitations that inform the conceptual model developed in this study.

Early studies relied on traditional time series and econometric models, such as ARIMA and ETS, to represent cyclical patterns and forecast short-term demand (Wang, 2016). These methods offered interpretability and a clear statistical framework but often struggled to capture the non-linear and multidimensional interactions typical of large urban systems (Chuwang and Chen, 2022). Accuracy decreased under abrupt weather shifts and station-level fluctuations common in networks such as Citi Bike (Hamad et al., 2021). Researchers incorporated external regressors, including temperature, precipitation, and public holidays, through ARIMAX and SARIMA specifications to better reflect environmental and temporal dynamics (Wang, 2016; Chuwang and Chen, 2022). While these refinements improved short-term forecasts, linear assumptions continued to limit representation of behavioral variability (Torres et al., 2024).

As data availability and computational tools improved, researchers increasingly used machine learning to capture non-linear relationships and complex feature interactions. Comparative analyses reported that ensemble methods such as Random Forest and Gradient Boosted Regression Tree (GBRT), as well as artificial neural networks, outperformed traditional statistical approaches (Torres et al., 2024; Mitchell, 2018). Torres et al. (2024) evaluated these algorithms using Citi Bike data and found that, while accuracy was comparable across models, Random Forest required less calibration effort and produced interpretable results suitable for operational forecasting. Wang (2016) similarly applied tree-based methods with iterative feature enhancement and reported improved accuracy and robustness. Together, these studies indicate that ensemble learning can better accommodate non-linearity and variability in bike-share demand.

Deep learning extended this work by explicitly integrating spatial and temporal information. Ma et al. (2022) proposed a STGA-LSTM network that captured temporal dependencies and spatial correlations between stations. By combining historical ridership, weather, and land-use data, the model achieved significantly higher accuracy than standard regression or univariate time series models. Zhou et al. (2024) introduced a STGCN that encoded inter-station relationships as a connected graph and improved short-term forecasting while revealing shared behavior across clusters of stations. These approaches highlight the value of representing structural and temporal complexity in urban mobility systems.

Subsequent refinements incorporate attention mechanisms, hybrid architectures, and domain-specific variables. Van der Wielen (2024) implemented LSTM and CNN-LSTM models for station-level forecasting using Citi Bike data and incorporated station capacity and embeddings to strengthen spatial representation. These architectures outperformed ARIMA and Random Forest benchmarks for one-hour-ahead predictions. Reinforcement learning methods also link prediction to operations. Zhang et al. (2024) developed a two-layer scheduling framework using deep reinforcement learning to connect demand prediction with bike rebalancing and optimize operational decisions in real time. Collectively, these studies position forecasting within integrated decision-support settings rather than as a standalone predictive task.

Parallel research emphasizes data quality and feature engineering. Porat, Ben-Elia and Fire (2025) proposed a micromobility demand forecasting framework that integrated temporal, spatial, and network variables and reported accuracy improvements of almost half compared with single-dimensional baselines, underscoring interdependence between spatial structure and temporal behavior. Wilkesmann (2022) similarly identified precipitation, temperature, and proximity to public transport nodes as key determinants in a station-based system. Using clustering analysis of Helsinki’s HSL City Bikes, Adrianzen (2025) identified commuter-driven usage patterns and weather sensitivity aligned with temporal and environmental cycles. These findings reinforce the need to model weather and temporal effects within spatial context when forecasting demand.

Graph-based methods further enhance the representation of spatial complexity. Betkier, Oszczypała and Dawid (2025) applied GNNs with spatial aggregation by grouping stations into kilometer-grid cells to reduce noise while improving performance, and their models achieved high explanatory power. Mohseni Hosseinabadi, Nourinejad and Park (2024) extended graph representations to network expansion planning by predicting ridership between station pairs and treating trips as weighted links in a directed graph. Their approach identified demand corridors and supported station siting decisions, demonstrating how network-based modeling can inform infrastructure design and policy.

Across methods, external determinants are consistently examined. Weather is repeatedly identified as a dominant influence on trip frequency and duration (Adrianzen, 2025; Ma et al., 2022). Moderate temperatures encourage cycling, whereas extreme conditions, such as heavy rain or excessive heat, reduce ridership (Gong et al., 2024). Precipitation remains the most disruptive variable, while wind speed tends to exert a weaker but measurable influence (Adrianzen, 2025). Seasonal variation is widely documented, with summer peaks and winter declines across northern hemisphere cities (Wilkesmann, 2022). Time-related factors, including weekdays, weekends, and public holidays, improve model performance when included as categorical predictors (Wang, 2016; Torres et al., 2024).

The built environment and urban form shape where and when demand occurs. Wei and Zhu (2023) identified correlations between land-use mix, density, and proximity to public transport in shaping Citi Bike usage and distinguished complementary and competitive relationships between bike sharing and public transit. They reported substitution in central areas and greater synergy in outer districts. Gong et al. (2024) incorporated natural environmental attributes such as greenery, air quality, and visual openness and found that aesthetic and sensory qualities influence ridership distribution. These studies indicate that demand patterns reflect both functional urban form and user sensitivity to environmental comfort and safety.

Socio-demographic context further explains spatial heterogeneity. Population density, employment distribution, and income contribute to variation, with employment zones generating morning peaks and residential areas driving evening returns (Hamad et al., 2021). In Shenzhen, Xing, Wan and Luo (2025) showed that jointly modeling demographic concentration and metro connectivity enhances predictive accuracy, underscoring the importance of multimodal integration. Together, these results support modeling socio-economic and spatial context alongside temporal and environmental factors to explain heterogeneous demand.

Recent research combines statistical time series methods with machine learning in hybrid frameworks. Ma et al. (2022) and Zhou et al. (2024) showed that hybrid neural architectures can capture both periodic and random components of demand and improve short-term accuracy without sacrificing interpretability. Prophet models, introduced in transportation forecasting by Chuwang and Chen (2022) and applied to bike sharing by Wilkesmann (2022), provide trend and seasonality decomposition that supports operational insight while maintaining automation. Explainable AI methods, including SHAP values, have been incorporated to identify influential predictors and improve transparency (Porat, Ben-Elia and Fire, 2025). These developments address interpretability concerns commonly associated with deep learning in operational and policy contexts.

Forecasting is also linked to operational and sustainability outcomes. Zhang et al. (2024) demonstrated that improving prediction precision reduces the cost and environmental impact of redistribution operations. Torres et al. (2024) argued that predictive optimization contributes to economic and environmental sustainability by minimizing unnecessary vehicle movements and emissions. Gong et al. (2024) linked bike usage patterns to public health and environmental quality, framing forecasting as a tool for efficiency as well as urban livability and resilience.

Despite these advances, limitations remain. Many studies rely on large, high-quality datasets that may not be available across cities, restricting the scalability and transferability of complex deep learning approaches. Models are frequently optimized for short-term forecasting, leaving limited evidence on medium- and long-term trends needed for strategic planning. Variation in data granularity complicates comparisons between system-level and station-level analyses (Betkier, Oszczypała and Dawid, 2025). Finally, interpretability and transparency remain challenging when deep learning models are used for policy-oriented decision-making (Porat, Ben-Elia and Fire, 2025).

Overall, the literature indicates a methodological evolution. Linear statistical models established a foundation for temporal analysis, while ensemble and neural approaches improved performance for non-linear, multivariate demand. Graph-based and hybrid methods offer a stronger balance among accuracy, interpretability, and operational relevance. Persistent issues related to transparency, generalizability, and integration of heterogeneous predictors justify a modeling framework that combines statistical and machine learning methods to forecast Citi Bike demand more accurately and comprehensively while remaining suitable for policy and operational use.

2.1. Conceptual Framework

This section organizes the determinants of bike-share demand identified in the preceding literature into a layered conceptual structure. Figure 1 summarizes this structure: environmental conditions, temporal patterns, and socio-demographic structures act as successive layers that together shape observed demand. The remainder of this section describes each layer and the interactions among them.

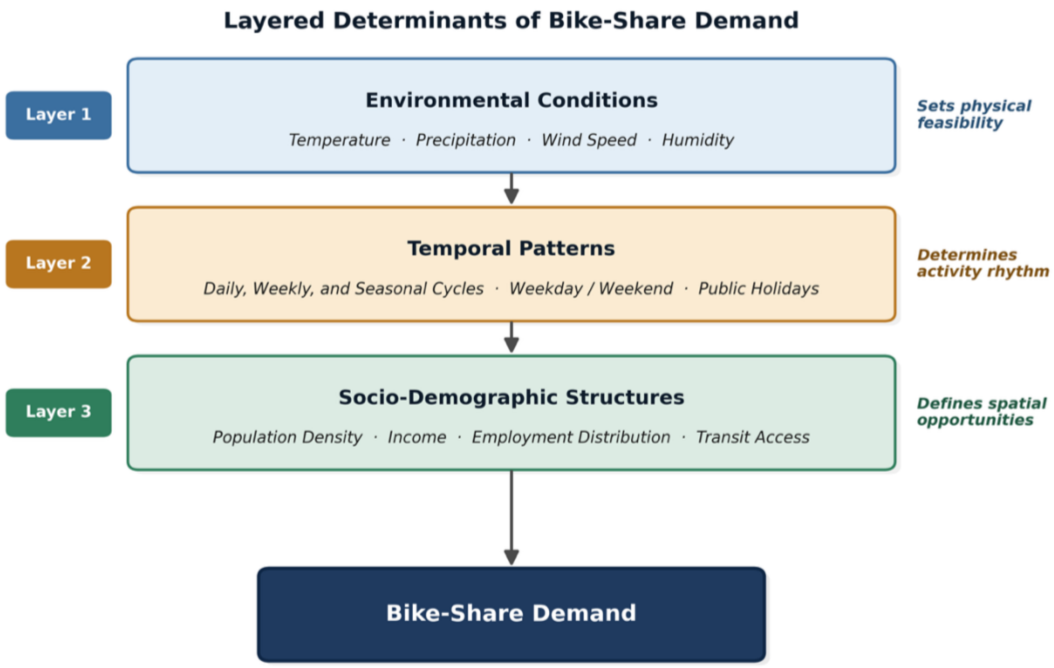

Figure 1 – Conceptual framework of layered determinants of bike-share demand.

Weather conditions are among the most significant external determinants of bike-share demand. Studies show that favorable temperatures promote ridership, while rainfall, snowfall, and extreme heat discourage it (Adrianzen, 2025; Ma et al., 2022). Temperature exerts a non-linear influence, where moderate conditions maximize cycling activity and deviations on either side of this range result in decline (Gong et al., 2024). Precipitation creates immediate deterrence because users are unlikely to rent bikes during rain or snow. Wind speed and humidity further shape comfort levels and perceived risk, moderating the effect of temperature. In New York City, where seasonal variation is pronounced, these weather dynamics generate recurring fluctuations in daily and monthly ridership patterns (Wang, 2016). In the conceptual model, weather acts as an external forcing variable that directly influences demand volume while also interacting with calendar variables such as season and day type. As shown in Figure 1, this constitutes the first layer of the framework, setting the boundary of physical feasibility for cycling.

Temporal structure, reflected in daily, weekly, and seasonal cycles, captures behavioral regularities linked to work, leisure, and commuting patterns. Studies show that weekday mornings and evenings correspond to commuting peaks, while weekends exhibit more leisure-oriented usage (Torres et al., 2024; Wilkesmann, 2022). Public holidays and special events create further deviations from routine patterns. Calendar effects therefore serve as temporal indicators of human activity and help explain why similar weather conditions can yield different demand levels depending on whether they occur on a weekday or weekend. This second layer of Figure 1 determines the rhythm of daily and weekly activity within the feasibility boundary set by environmental conditions.

Demographic and socio-economic characteristics provide the contextual layer through which weather and temporal effects are expressed spatially. Population density, income, employment distribution, and access to public transit influence the intensity and distribution of bike-share demand (Wei and Zhu, 2023; Xing, Wan and Luo, 2025). High-density employment areas generate concentrated morning departures and evening returns, while residential areas tend to produce the opposite flow. Integration with subway stations and other transit modes facilitates multimodal trips, expanding the catchment area of the bike-share network (Xing, Wan and Luo, 2025). These socio-spatial factors determine baseline propensity for bike-share use, while temporal and environmental variables modulate this baseline over time. This third layer of Figure 1 defines the spatial opportunities and constraints within which the preceding layers operate.

In this framework, predictors interact dynamically rather than independently. Machine learning models can capture non-linear effects and higher-order interactions without requiring a pre-specified linear or additive functional form (Hastie, Tibshirani and Friedman, 2009; James et al., 2021). Statistical time series components complement this by explicitly modeling autocorrelation and seasonal structure, ensuring that recurring temporal patterns are represented (Hyndman and Athanasopoulos, 2021). Integrating these techniques allows forecasting to account for long-term regularities and short-term fluctuations that characterize urban mobility demand.

Conceptually, and as summarized in Figure 1, the relationship between these determinants and bike-share demand can be described as a layered process. Environmental conditions set the boundary of physical feasibility for cycling. Temporal patterns determine the rhythm of daily and weekly activity. Socio-demographic structures define spatial opportunities and constraints, shaping who uses the system and where. The forecasting models developed in this study aim to produce forecasts that are accurate, interpretable, and aligned with how factors jointly shape urban bike-share demand.

3. METHODOLOGY

This dissertation examines bike-share demand in the Citi Bike system in NYC using an empirical, quantitative workflow based in the Cross-Industry Standard Process for Data Mining (CRISP-DM) and designed for reproducible forecasting at system and station levels. CRISP-DM is a widely adopted, industry-independent process model for structuring data mining projects and remains the de-facto standard for such work two decades after its release (Schröer, Kruse and Gómez, 2021). The framework organizes a project into six iterative phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Business understanding defines the project objectives and translates them into a data mining problem, while data understanding focuses on initial data collection, familiarization, and quality assessment. Data preparation covers the construction of the final dataset from raw inputs, including cleaning, integration, and feature engineering. Modeling involves selecting and applying analytical techniques and calibrating their parameters, and evaluation assesses whether the resulting models meet the original objectives. Deployment concerns the organization and presentation of results so that they can support decision-making (Schröer, Kruse and Gómez, 2021).

The present study adopts CRISP-DM because it provides a structured yet flexible framework suited to multi-source, multi-resolution forecasting work, and because its phased structure aligns with the dissertation's emphasis on transparency and reproducibility. The business understanding phase is reflected in the research question and objectives set out in Chapter 1, which frame demand forecasting as an operational and planning task for Citi Bike. Data understanding and data preparation are addressed in Sections 3.1 through 3.3, which describe the integration of trip records with weather, calendar, environmental, mobility, and built-environment datasets, alongside the exploratory analysis used to validate temporal and spatial structure. Modeling is covered in Section 3.4, which specifies the classical, supervised learning, and deep learning configurations applied at system, cluster, and station levels. Evaluation is addressed in Section 3.5 through fixed time-ordered splits and standard error metrics, and in Section 3.6 through explainability and robustness procedures. Deployment is treated in a limited operational sense through the discussion of policy and operational implications in Chapter

5, consistent with the observation that many academic applications of CRISP-DM emphasize the analytical phases more than full operational deployment (Schröer, Kruse and Gómez, 2021).

The primary data are complete Citi Bike trip records from 1 January 2017 to 31 December 2024, stored as monthly Parquet files. The period spans major operational growth and reporting changes, so harmonization and temporal consistency are prerequisites for modeling.

To support forecasting with contextual inputs, the trip data were augmented with several external datasets. These included daily NOAA weather from four stations (Central Park, LaGuardia, JFK, and Newark), a standardized U.S. federal holiday series, and daily AQI at both citywide and county levels. Additional inputs consisted of automated bicycle counts and traffic volume counts, tract-level ACS socio-demographics, and parcel-level PLUTO land use data prepared for spatial integration. NYC bicycle route network data were also transformed into an annual panel and incorporated into the analysis. Metropolitan Transit Authority subway turnstile usage was converted and inspected, but excluded due to formatting inconsistency and the assessment that it would not materially improve explanatory value relative to retained inputs. Event data were available but were not used, as the methodology emphasizes federal holidays as the calendar augmentation described here.

3.1. Data Preparation and Integration

Data preparation followed a reproducible sequence to ensure stable schemas and consistent temporal and spatial alignment. Across auxiliary sources, inputs were standardized (Parquet conversion where applicable, explicit typing, lowercase snake case). Spatial datasets were validated and assigned Coordinate Reference System (CRS) metadata prior to joins.

Citi Bike trips were cleaned to address schema drift and inconsistent semantics across years. A global scan identified the union of historical fields (19 variables), and a canonical schema was imposed on all monthly files. Because ride identifiers restarted within files, a new globally unique trip identifier replaced legacy ride identification values. Both *started_at* and *ended_at* were parsed as timestamps, and trip duration was recalculated in integer seconds, with records containing missing or negative durations set to missing and then filtered to remove trips shorter than one minute and longer than 24 hours. Membership labels were standardized to the two-level scheme (“member”, “casual”). Rider attributes were standardized by enforcing a birth-year range of 1900–2021 with missingness preserved and by recoding and labeling gender, and *bike_id* was removed after confirming extensive missingness beyond the early years.

Station metadata were harmonized through a station master derived from all trip files, providing a unified activation span, stable name, and corrected coordinates per station. An *is_real_station* flag identified internal operational locations (e.g., warehouses, quality control, valet depots) for exclusion. Real stations were assigned borough, census tract, and neighborhood tabulation area using contemporary boundaries, with ambiguous cases such as Roosevelt Island resolved by nearest-neighbor assignment while respecting administrative designations. The master was enriched with tract-level ACS attributes and tract-aggregated PLUTO measures, and corrected station metadata replaced drifting station fields in the trip files. After validation, business-key deduplication identified a duplicate cluster confined to 2017, and 2,349 trips were removed, yielding 210,269,083 unique trips.

Auxiliary datasets were prepared for integration. NOAA weather tables were harmonized across stations (including *wt01–wt09*). Missing average temperature was reconstructed as the mean of the minimum and maximum temperature. Average wind was linearly interpolated when partially missing, while precipitation and snow variables were not interpolated, except for limited forward-filling only when the prior day recorded zero precipitation or snow. Weather-type indicators were recoded to binary variables, sparse variables were dropped, and daily indicators were engineered. A blended NYC daily weather series was constructed as the primary input using a weighted average (Central Park 0.4, LaGuardia 0.4, JFK 0.1, Newark 0.1), with mean- and median-based alternatives also produced. Cleaned station-level weather tables were retained for sensitivity analysis.

Federal holidays were standardized to a canonical datetime join key, with normalized holiday names and stable identifiers (short codes and *holiday_id*). AQI series were typed and standardized, with citywide AQI renamed *aqi_citywide*, county metadata fields removed, and county FIPS constructed. Traffic and bicycle count data were cleaned, timestamped, converted to spatial geometries where applicable, and aggregated to hourly and daily resolutions consistent with demand targets. ACS tract data were standardized, geometries validated, and derived indicators computed. PLUTO parcels were standardized and aggregated to the tract level to produce land-use and built-form summaries, including floor area ratio. Bicycle route data were converted to spatial geometries, facility types were harmonized, segment lengths were computed, and installation and retirement dates were cleaned to produce a yearly segment panel from 1980 to 2025.

3.2. Target Variable and Feature Engineering

Targets were constructed at system, station, and cluster resolutions under hourly and daily granularities using standardized Parquet inputs, DuckDB aggregation and joins, and pandas validation.

System targets were trip counts by trip start time: hourly counts by hour-truncated start timestamp and daily counts by calendar date, built as complete sequences over 2017–2024. Station targets were constructed from station-coded events: hourly departures, arrivals, net flow (arrivals minus departures), and daily departures. Because station activity varies over time, station targets were produced as sparse panels with observed station–time combinations and dense panels with complete station–time grids and explicit zeros.

Sparse station targets exclude records with missing station identifiers, restrict to real stations, cast station identifiers to string type, and limit observations to each station's verified operational window. Dense hourly and daily targets were generated by constructing a full station–time cartesian product within each year, restricted to station active periods. Sparse observations were left joined, missing values filled with zeros, and checks verified grid sizes, type stability, and equality of aggregate flows between representations.

Stations were clustered into functional groups using static socio-demographic and built-environment features from the integrated ACS plus PLUTO tract dataset linked through the station master (including land-use shares, transit commute prevalence, population density, and built-form intensity such as floor area ratio and building density). Candidate k-means solutions ($k \in \{2,...,10\}$) were evaluated using Within-Cluster Sum of Squares (WCSS), silhouette, Calinski–Harabasz (CH), and Davies–Bouldin (DB). Figure

A1 (see Appendix A) visualizes these diagnostics, and Table I reports the values for k = 2–5. Cluster structure was strongest at small values of k. Although k = 2 produced the highest silhouette and CH values, it was considered overly coarse, while k = 4 minimized the DB index but introduced groupings that lacked conceptual coherence. A three-cluster solution was selected as a balance between quantitative support and interpretability and labeled residential, commercial, and transit-adjacent based on underlying feature characteristics. A complete station-to-cluster mapping was validated, and any remaining unclustered edge cases were assigned by the nearest clustered station. Cluster composition is reported in Appendix A (Table AI), and cluster totals and distributions are reported in Appendix A (Tables AII and AIII). Cluster-level hourly and daily targets were generated by aggregating station flows within clusters (hourly departures, arrivals, net flow, and daily departures), and differences versus system totals reflect exclusion of trips with missing station identifiers.

TABLE I

CLUSTER VALIDITY METRICS FOR K-MEANS CANDIDATE SOLUTIONS

K	WCSS	Silhouette	Calinski–Harabasz	Davies–Bouldin
2	13,982	0.382	1,520	1.301
3	10,816	0.349	1,476	1.181
4	8,765	0.345	1,477	1.093
5	7,755	0.251	1,362	1.142

Author's computations based on standardized station-linked ACS and PLUTO features.

Feature engineering used a unified, leakage-controlled pipeline shared across system, station, and cluster tables. Weather features were joined by date (hourly panels inherit daily values) and include temperature, precipitation and snow measures, wind speed, binary event and extreme indicators, and rolling precipitation windows. Calendar features were centralized in canonical daily and hourly dimensions with consistent temporal descriptors, holiday and weekend flags, pandemic-regime labels, and (hourly) sinusoidal hour-of-day encodings. AQI enrichment uses citywide AQI as the uninterrupted backbone and substitutes citywide AQI when county AQI is missing, with threshold indicators derived and station-hour tables replicating station-day AQI values across 24 hours.

Mobility features map stations to the nearest bicycle counter and nearest traffic segment, and aggregate traffic within 1,000 m and 2,000 m radii. Sensor data were standardized to unique sensor-hour observations, with duplicates aggregated, and missingness arising from sensor sparsity was retained. Autoregressive features were computed within each series using strictly backward-looking windows, including daily lags (1, 7) and rolling means (7, 30), as well as hourly lags (1, 24, 48, 168) and rolling means (6, 24, 72). Static station attributes from the enriched station master were joined onto station-time panels and aggregated to clusters as needed, with join constraints and row-count checks. After feature construction, only provably identical redundant columns were removed. Integration steps used left joins on stable keys (date, hour timestamp, station, cluster), with checks for key uniqueness, row-count preservation, schema and type stability, and consistency between sparse and dense aggregates.

3.3. Exploratory Data Analysis Guiding Modeling Choices

Exploratory Data Analysis (EDA) was used to validate assumptions about temporal structure, exogenous sensitivity, and spatial heterogeneity prior to model specification, and to confirm that the engineered feature sets reflect observable structure.

System-level daily demand shows strong seasonality and regime-dependent behavior. Figure 2 displays daily trip volumes with pandemic regime labels, showing stable seasonality prior to 2020, disruption during COVID, and a recovery phase with different levels and variability. This motivates explicit regime indicators and evaluation designs that respect temporal ordering and regime boundaries.

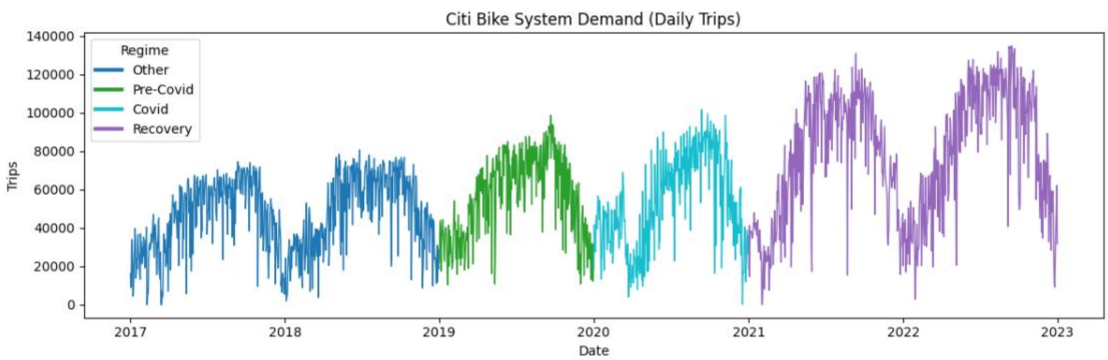

Figure 2 – Citi Bike system demand (daily trips) with pandemic regime labels.

Hourly system demand exhibits strong cyclicality and weekly repetition. Figure 3 presents the Autocorrelation Function (ACF) for system hourly trips over lags up to 14 days, showing repeated peaks consistent with daily and weekly periodicity. This supports explicit daily and weekly seasonal structure in classical models, and cyclic encodings and lag features (e.g., 24-hour and 168-hour lags) in supervised learning.

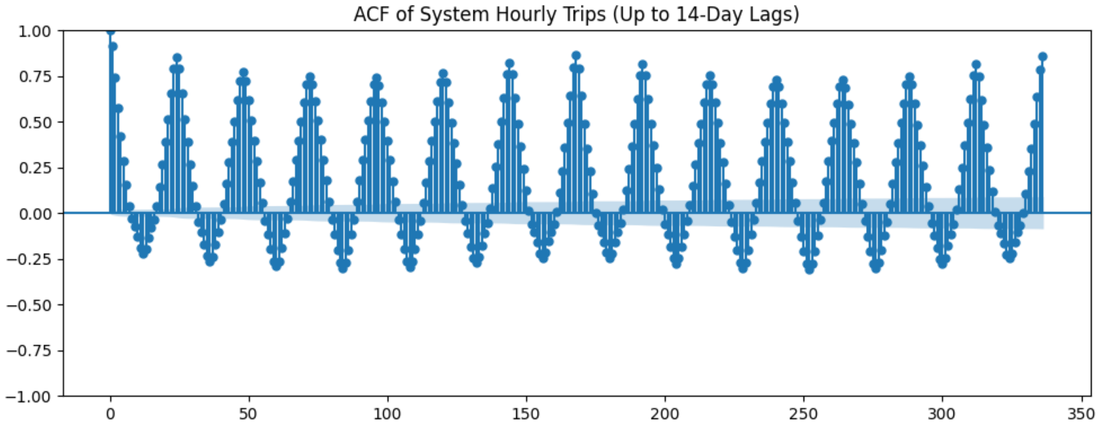

Figure 3 – ACF for system hourly trips (lags up to 14 days).

Station demand is heterogeneous, and static covariates explain only part of the cross-sectional variation. A station-level Ordinary Least Squares (OLS) model explains approximately 27% of variation in average daily departures ($R^2 = 0.268$) and is jointly statistically significant ($p < 0.001$). Associations are consistent with higher demand in denser and higher-employment environments and lower demand where stations are farther from cycling infrastructure or where transit commuting dominates. However, the moderate explanatory power, non-normal residual behavior, and multicollinearity, reflected in a large condition number, imply that a single global linear specification is incomplete and that coefficients should not be interpreted causally. Residual behavior also differs by station type (Figure 4), supporting cluster aggregation and models capable of nonlinearities and interactions between static context and dynamic drivers. Additional EDA outputs are provided in Appendix B.

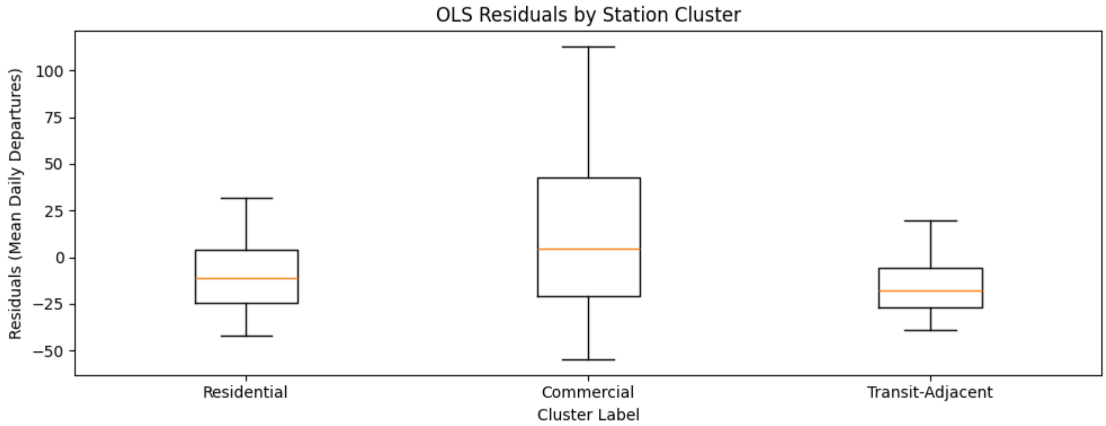

Figure 4 – OLS residuals by station cluster (mean daily departures).

3.4. Forecasting Models and Configurations

This section summarizes the forecasting models implemented at system, cluster, and station levels and the configurations used to ensure comparability. The strategy combines multiple forecasting approaches, including ARIMA, ETS, and Prophet configurations (each evaluated with and without seasonal and exogenous components), alongside supervised learning models (Random Forest and XGBoost) and a station-level CNN-LSTM model. Configurations enforce strict temporal ordering, document feature inputs, and use fixed specifications and selection criteria.

Daily system models for 2017–2022 were informed by unit-root diagnostics and inspection of the ACF and PACF. Stationarity testing indicated non-stationarity in the raw daily series. First differencing and weekly seasonal differencing each achieved stationarity (Appendix C, Table CI). The primary baseline uses weekly seasonal differencing with $s = 7, D = 1$ and no additional non-seasonal difference ($d = 0$). Candidate SARIMA orders were restricted to low-order dynamics and selected based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) criteria together with Ljung–Box diagnostics at weekly multiples (Appendix C, Table CII), with residual tests used to assess relative adequacy.

System-level SARIMAX extends the seasonal baseline with standardized exogenous regressors: *is_holiday*, *is_weekend*, pandemic regime indicators (*regime_pre_covid*, *regime_covid*, *regime_recovery*), and weather variables (*temp_avg_f*, *precip_inches*). Table II reports the comparison assessing the incremental contribution of weather covariates.

TABLE II

COMPARISON OF SARIMAX MODELS

Model	AIC	BIC	LB_p (7)	LB_p (14)	LB_p (21)
SARIMAX + Calendar + COVID	47,853	47,904	9.42e-15	1.03e-12	1.01e-10
SARIMAX + Calendar + COVID + Weather	46,478	46,540	3.73e-09	1.08e-08	4.13e-08

Author’s computations based on the clean Citi Bike system-level daily demand series.

ETS benchmarks were implemented with weekly seasonality fixed at s = 7, and the retained baseline is ETS(A,N,A) (Appendix C, Table CIII). ETS models augmented with the same exogenous regressors were also estimated. Prophet was configured with additive seasonality, weekly and yearly components enabled, daily seasonality disabled, and a piecewise linear trend with automatic changepoint selection. For comparability, Prophet included *temp_avg_f*, *precip_inches*, *is_holiday*, and regime indicators.

To assess disaggregation, daily baselines were also estimated for cluster aggregates and a subset of high-demand stations (selected by mean daily demand over 2017–2022). Weekly seasonality (s = 7) was retained, and orders kept simple. Station baselines do not impose weather regressors by default, reserving weather inclusion for targeted sensitivity checks.

Random Forest and XGBoost were implemented for system and cluster forecasting at daily and hourly resolutions using the engineered feature matrices (calendar, lags and rolling terms, weather, AQI, traffic, bicycle counts). One-step-ahead prediction aligns features at time t with demand at t + 1, and split membership is defined by prediction timestamp to prevent leakage. Hyperparameters were tuned on the validation period using fixed time-ordered splits, with models refit on combined training and validation data prior to test evaluation.

Station-level short-term forecasting used a global CNN-LSTM trained across stations with a learnable station embedding. Inputs are sequences over lookback window T (24 or 48 hours) containing lagged demand and time-varying covariates (weather and calendar). Static station attributes (built environment, socio-demographics, borough, cluster) enter via embeddings and concatenated numeric context. The architecture applies onedimensional convolution over time, followed by LSTM layers and produces a one-step-ahead forecast for the next hour. Training uses early stopping on validation loss. For memory-safe training, station-hour inputs were exported as partitioned Parquet files and converted to chunked NPZ tensors. Station coverage was held fixed across splits. A Histogram-Based Gradient Boosting (HGB) model was trained on the same station subset and splits as a fast station-level tabular comparator.

3.5. Evaluation Design and Metrics

All evaluations use time-ordered splits with no shuffling: training 2017–2021, validation 2022, testing 2023–2024. The same split boundaries are applied at daily and hourly resolutions and, where applicable, at system, cluster, and station levels.

For supervised learning, one-step-ahead forecasting is implemented by shifting the target forward one time unit and assigning split membership by the prediction timestamp. For classical models and Prophet, specification selection uses the training and validation split, and models are re-estimated on training and validation when required for test evaluation.

Models are evaluated at the system level using NYC aggregate series, at the cluster level using cluster-level aggregates, and at the station level using station-hour forecasts for a fixed subset of stations for the CNN-LSTM and aligned tabular baselines. Station results are summarized primarily by macro-averaging (equal weight per station), with volume-weighted summaries reserved for secondary interpretation.

Forecast accuracy is assessed using RMSE, MAE, and MAPE. MAPE is interpreted conservatively at the station level because actual demand is frequently near zero. RMSE and MAE are treated as primary station metrics. Classical model selection also uses AIC, BIC, and Ljung–Box diagnostics at weekly multiples, with supporting tables in Appendix C (Tables CI–CIII). Machine learning models select hyperparameters on validation performance, and refit on training and validation. CNN-LSTM uses early stopping against validation loss, and selection procedures avoid any use of test-period information during specification, tuning, or feature construction.

3.6. Explainability and Robustness

Explainability is treated as a design requirement: models are interpreted in terms of their information sets and structural roles rather than directional claims about fitted parameters. Classical models provide structure-based explanations by incorporating seasonal differencing and seasonal terms, short-run dependence, and exogenous regressors. Tree ensembles are explained using SHAP global and local attributions to summarize reliance on feature families and to check local logic. CNN-LSTM explainability uses feature-group ablation to assess relative dependence on calendar encodings, recent demand history, external forcing variables, and station context while preserving marginal distributions.

Robustness checks evaluate whether these explanation patterns remain coherent across pandemic regimes and across residential, commercial, and transit-adjacent clusters. Feature-group removal experiments test whether behavior aligns with intended feature roles (external forcing, calendar structure, context), and explanations are compared across system, cluster, and station aggregation levels to assess consistency under spatial disaggregation.

4. RESULTS AND DISCUSSION

This chapter synthesizes Citi Bike forecasting results across system, cluster, and station levels at daily and hourly resolutions, linking performance patterns to interpretability and robustness evidence. It begins with daily, system-level baseline one-day-ahead forecasting, comparing ETS(A,N,A) with and without exogenous information under a weekly seasonal period ($s = 7$), SARIMAX$(1,0,1)\times(0,1,1)_7$, and Prophet variants. Consistent with Chapter 3, performance is evaluated using rolling one-day-ahead expanding-window forecasts with 2022 as validation and 2023–2024 as the test period. Errors are summarized using RMSE, MAE, and MAPE, while exogenous variables include temperature, precipitation, holiday and weekend indicators, and regime indicators.

Across the 2022 validation year, ETS with exogenous variables and SARIMAX are effectively tied (Table III). The difference between these two specifications is small relative to the gap between models with and without covariates (Appendix D, Table DI), indicating that exogenous information drives the main aggregate accuracy gains. Prophet remains weaker than both state-space baselines even with regressors, and the endogenous Prophet specification performs worst.

TABLE III

SYSTEM-LEVEL BASELINE PERFORMANCE FOR DAILY ONE-DAY AHEAD FORECASTS

Model	Split	RMSE	MAE	MAPE (%)
ETS(A,N,A), $s = 7$ + Exog	Test (2023–2024)	15,181	10,895	14.00
SARIMAX$(1,0,1)\times(0,1,1)_7$	Test (2023–2024)	15,476	11,095	14.22
Prophet + Exog	Test (2023–2024)	19,601	15,271	18.54
Prophet (Weekly + Yearly)	Test (2023–2024)	24,123	18,742	24.44
ETS(A,N,A), $s = 7$ + Exog	Validation (2022)	12,171	9,183	16.80
SARIMAX$(1,0,1)\times(0,1,1)_7$	Validation (2022)	12,235	8,964	16.67
Prophet + Exog	Validation (2022)	15,087	12,086	26.53
Prophet (Weekly + Yearly)	Validation (2022)	19,770	15,615	38.39

Author's computations based on Citi Bike system-level demand forecasts.

On the 2023–2024 test period, the ordering is unchanged: ETS with exogenous variables is strongest, followed closely by SARIMAX, with Prophet variants weaker

(Table III). Appendix D, Table DI shows that ETS(A,N,A), $s = 7$ without regressors is substantially worse than ETS with regressors in both evaluation periods (e.g., test RMSE 20,598 versus 15,181). Moreover, removing weather regressors from ETS with exogenous variables yields performance close to the endogenous ETS specification, indicating that weather is the primary driver of the ETS with exogenous improvement in this setup. The dominance of weather covariates is consistent with findings that precipitation and temperature are among the most influential determinants of bike-share ridership (Adrianzen, 2025; Gong et al., 2024; Ma et al., 2022).

Error behavior across periods is consistent with demand scaling. For models with exogenous variables, MAPE decreases from validation (2022) to test (2023–2024) while RMSE and MAE increase modestly (Table III), implying larger absolute errors at higher demand levels but smaller proportional errors. Because system-level demand does not approach zero, MAPE remains interpretable here. This becomes more restrictive in sparse station-level settings.

To complement aggregate metrics, rolling one-day-ahead forecasts over the final 90 days of the test period were examined for ETS with exogenous variables, SARIMAX, and Prophet with exogenous variables. In Figure 5, ETS and SARIMAX track weekly structure closely through October and early November. From mid-November through December, observed demand includes abrupt drops and rebounds. Both baselines follow level and seasonality but understate the sharpest declines, while Prophet departs more and shows more persistent misalignment during late-period shocks. This behavior is consistent with reports that predictive accuracy decreases under abrupt shifts in demand conditions and weather-driven variability in Citi Bike and related systems (Hamad et al., 2021).

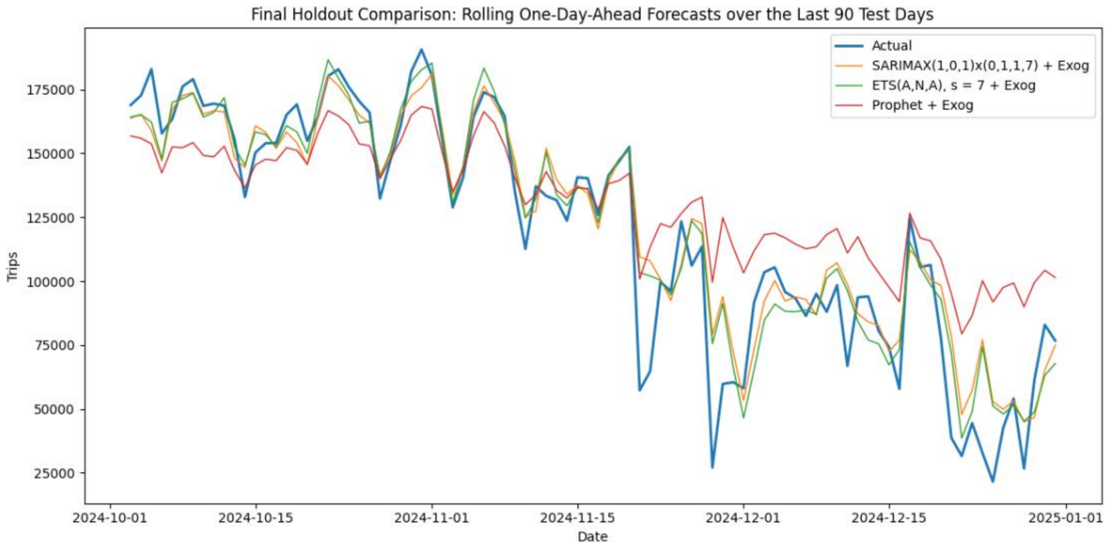

Figure 5 – Observed system-level demand versus rolling one-day-ahead forecasts over the final 90 days of the test period.

Figure 6 summarizes the same window using smoothed absolute error (7-day rolling mean). ETS with exogenous variables maintains the lowest smoothed error for most of the holdout period, with SARIMAX typically slightly higher, while Prophet with exogenous variables is highest throughout and rises sharply in late December. The corresponding unsmoothed daily absolute errors are provided in Appendix D (Figure D3) and show error spikes for all models during abrupt shocks, with larger and more frequent spikes under Prophet than under ETS or SARIMAX.

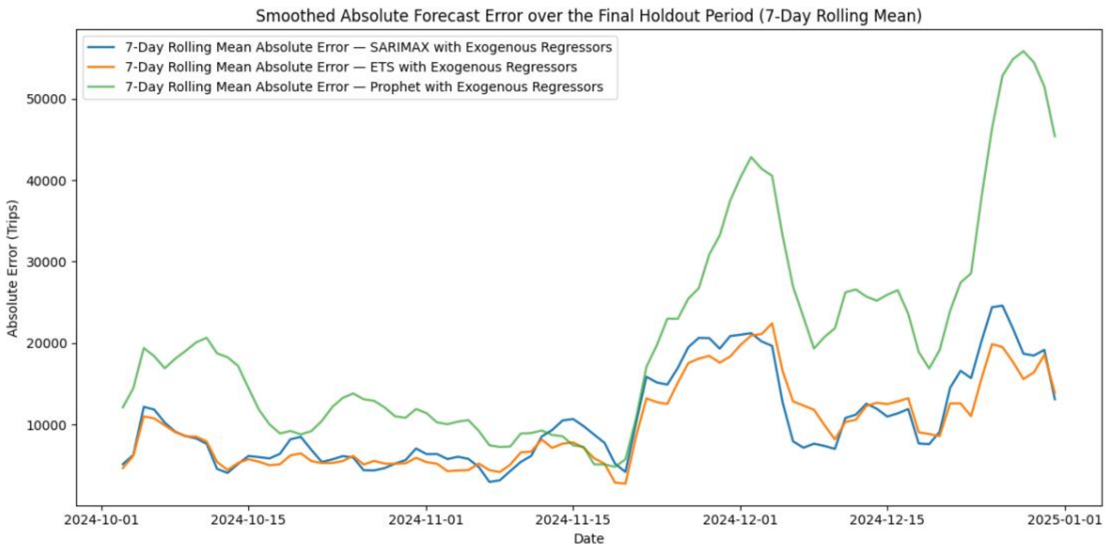

Figure 6 – Smoothed absolute error (7-day-rolling mean) over the final 90-day holdout window for rolling one-day-ahead forecasts.

Baseline interpretability is assessed primarily through SARIMAX and Prophet regressors. For SARIMAX$(1,0,1)\times(0,1,1)_7$, temperature, precipitation, and holidays are statistically significant (positive, strongly negative, and negative, respectively), while the weekend indicator is effectively zero and statistically non-significant (Appendix D, Table DII). Regime indicators are not statistically significant in this fitted baseline. These directions align with prior findings that temperature and precipitation are dominant external determinants and that calendar indicators can improve demand prediction (Wang, 2016; Torres et al., 2024; Adrianzen, 2025).

Prophet's approximations of regressor effect sizes offer a complementary ranking. Appendix D (Table DIII) reports scale-restored approximations intended for directional and relative comparison rather than precise marginal effects. Holidays, precipitation, and weekends are the largest negative contributors, while temperature and regime indicators are smaller after Prophet's seasonal structure is accounted for.

Residual diagnostics indicate remaining structure in errors for both state-space baselines. SARIMAX diagnostics reject residual independence, normality, and homoskedasticity (Appendix D, Figure D1), and ETS with exogenous variables shows similar behavior with significant Ljung–Box and Jarque–Bera statistics and evidence of heteroskedasticity (Appendix D, Figure D2). These findings align with error spikes in Appendix D (Figure D3) and motivate the robustness discussion in Section 4.3.

Overall, two findings carry forward. First, endogenous-only baselines (ETS or Prophet) are inadequate for system-level Citi Bike demand. Second, once weather and calendar covariates are included, ETS(A,N,A), $s = 7$ is a strong baseline, matching SARIMAX in validation and marginally outperforming it in the test period (Table III), while residual non-normality and heteroskedasticity remain evident limitations. The reliance on exogenous information at the system level is consistent with prior Citi Bike analyses emphasizing weather and calendar forcing in demand variation (Wang, 2016; Hamad et al., 2021).

4.1. Comparative Performance Across Models and Levels

This section synthesizes out-of-sample performance across model families and aggregation levels under the evaluation design in Chapter 3. Because targets differ in scale across system, cluster, and station levels and across hourly and daily temporal resolutions, comparisons focus on within-level rankings, with R^2 used to contextualize predictability across aggregation choices.

At the system level with hourly resolution, tree-based models perform strongly when supplied with autoregressive features and calendar–weather covariates. Random Forest outperforms XGBoost on the test set and explains more variance ($R^2 = 0.90$ versus 0.86; Table IV). This aligns with prior Citi Bike studies reporting strong performance for tree-based ensembles when supplied with autoregressive structure and weather–calendar features (Wang, 2016; Torres et al., 2024).

TABLE IV

SYSTEM-LEVEL HOURLY FORECASTING PERFORMANCE ON THE TEST SET

Model	RMSE	MAE	MAPE (%)	R^2
Random Forest	1,181	688	23.46	0.899
XGBoost	1,398	815	25.12	0.859

Author's computations based on system-level Citi Bike demand forecasts.

At the system level with daily resolution, classical baselines dominate and daily tree-based models degrade materially. ETS with exogenous variables and SARIMAX substantially outperform daily Random Forest and daily XGBoost (Table V), while daily tree models explain limited variance ($R^2 = 0.42$ and 0.20). These findings are aligned with early work emphasizing the value of classical time series structure (Wang, 2016; Chuwang and Chen, 2022), but they are not fully aligned with comparative studies reporting that ensemble methods often outperform traditional statistical approaches in multivariate settings (Torres et al., 2024; Mitchell, 2018).

TABLE V

SYSTEM-LEVEL DAILY FORECASTING PERFORMANCE ON THE TEST SET

Model	RMSE	MAE	MAPE (%)	R^2
ETS + Exog	15,181	10,895	14.00	—
SARIMAX	15,476	11,095	14.22	—
Random Forest	31,044	24,631	25.56	0.422
XGBoost	36,618	29,242	27.81	0.195

Author's computations based on system-level Citi Bike demand forecasts.

Moving from system-level to cluster-level forecasting improves predictability for tree-based models. At the cluster-hourly level, both models achieve very high explanatory power, with Random Forest stronger on the test set (R^2 = 0.95 versus 0.93; Table VI). At the cluster-daily level, Random Forest again yields lower RMSE and MAE and higher R^2 (0.86) than XGBoost (0.79), while MAPE is similar (Table VII). The improvement from spatial aggregation is consistent with studies showing that incorporating spatial context or aggregating stations can reduce noise and improve predictive performance (Betkier, Oszczypała and Dawid, 2025; Porat, Ben-Elia and Fire, 2025).

TABLE VI

CLUSTER-LEVEL HOURLY FORECASTING PERFORMANCE ON THE TEST SET

Model	RMSE	MAE	MAPE (%)	R^2
Random Forest	409	205	19.78	0.952
XGBoost	495	238	20.12	0.929

Author's computations based on cluster-level Citi Bike demand forecasts.

TABLE VII

CLUSTER-LEVEL DAILY FORECASTING PERFORMANCE ON THE TEST SET

Model	RMSE	MAE	MAPE (%)	R^2
Random Forest	11,198	6,871	22.73	0.862
XGBoost	13,726	7,868	22.64	0.792

Author's computations based on cluster-level Citi Bike demand forecasts.

At the station level with hourly resolution, evaluation is performed on a matched subset of stations with consistent splits. A fast tabular baseline (HGB) achieves validation RMSE 1.556 and test RMSE 1.697, with test MAE 1.032 and R^2 0.561. The CNN-LSTM improves MAE relative to HGB (0.935 versus 1.032) but yields higher RMSE (1.928 versus 1.556), indicating greater sensitivity to large errors. Table VIII reports checkpoint, macro-averaged, and volume-weighted summaries. This mixed result is not fully aligned with station-level Citi Bike evidence reporting CNN-LSTM improvements over ARIMA and Random Forest for one-hour-ahead prediction (Van der Wielen, 2024) and is more consistent with evidence that deep learning gains are strongest when architectures explicitly encode spatial correlations (Ma et al., 2022; Zhou et al., 2024).

TABLE VIII

STATION-LEVEL HOURLY FORECASTING PERFORMANCE ON THE TEST SET FOR CNN-LSTM

Metric Type	RMSE	MAE	MAPE (%)
Checkpoint Test	1.928	0.935	41.00
Macro Average (150 Stations)	1.527	0.935	—
Volume-Weighted	2.846	1.854	—

Author's computations based on station-level Citi Bike demand forecasts.

MAPE is intentionally de-emphasized at the station level because near-zero denominators make percentage error unstable. Appendix E, Table EI reports extremely large station-level MAPE values driven by stations with extended zero or near-zero demand. Accordingly, RMSE and MAE are treated as primary metrics, while MAPE is retained only where numerically stable. This evaluation constraint is consistent with broader concerns that data granularity can complicate model comparison and metric interpretation in station-level settings (Betkier, Oszczypała and Dawid, 2025).

Across levels, test results support three comparative conclusions. First, the best model class depends on aggregation level and temporal resolution: system-level daily forecasting favors ETS with exogenous variables (SARIMAX is close), while system-level hourly forecasting favors Random Forest among evaluated tree-based models. Second, spatial aggregation to clusters improves predictability for tree-based models at both hourly and daily resolutions. Third, at the station level, performance depends on how stations are weighted, and metric stability becomes a binding constraint for percentage errors. Overall, the dependence of rankings on temporal resolution and aggregation level is consistent with prior observations that predictability changes with granularity across bike-share forecasting settings (Betkier, Oszczypała and Dawid, 2025).

4.2. Interpretability and Drivers of Demand

Interpretability is assessed using model-aligned approaches: regression coefficients and stochastic terms for parametric time series models, SHAP for tree-based models, and feature-group ablation for the station-level CNN-LSTM. The use of SHAP aligns with explainable AI approaches applied to micromobility demand forecasting (Porat, Ben-Elia and Fire, 2025).

For SARIMAX$(1,0,1)\times(0,1,1)_7$, estimated stochastic terms indicate strong weekly dependence after seasonal differencing. The AR(1) coefficient is 0.957, the MA(1) coefficient is −0.675, and the seasonal MA coefficient at lag 7 is −0.904 (Appendix D, Table DII).

SARIMAX exogenous coefficients identify the strongest observable short-term drivers of daily system demand: temperature is positive (*temp_avg_f* = 819), precipitation is strongly negative (*precip_inches* = −20,900), and holidays are negative (*is_holiday* = −11,265). The weekend indicator is effectively zero (*is_weekend* = 0), and regime indicators are not statistically distinguishable from zero (Appendix D, Table DII). This pattern aligns with studies identifying temperature and precipitation as dominant determinants, with precipitation consistently disruptive to demand (Adrianzen, 2025; Gong et al., 2024; Ma et al., 2022; Wilkesmann, 2022).

Interpretation of ETS(A,N,A), $s = 7$ differs because the ETS with exogenous variables construction used here does not embed regressors as state-space inputs in the same manner as SARIMAX. Exogenous effects are handled outside ETS, and ETS components reflect remaining temporal structure after accounting for regressors.

Prophet is included for transparent additive decomposition and, secondarily, regressor effects. The fitted Prophet model with regressors is provided in Appendix F (Figure F1). Approximate scale-restored effects summarized in Appendix F (Figure F2) indicate that holidays and precipitation correspond to the largest negative adjustments, temperature is smaller, and regime indicators are smaller and less stable. This use is consistent with applications of Prophet in transportation and bike-share contexts as a decomposable trend and seasonality model intended to support operational insight (Chuwang and Chen, 2022; Wilkesmann, 2022).

For the system-level hourly Random Forest, SHAP indicates that predictions are driven primarily by intraday timing and demand-history persistence. The highest-ranked feature is *hour_cos*, followed by lagged demand at one hour, one day, and one week (*trips_lag1h*, *trips_lag24h*, *trips_lag168h*) and rolling means (*roll6h*, *roll24h*) (Appendix F, Table FI). Weather variables rank lower, with precipitation exhibiting a larger mean absolute SHAP magnitude than temperature. Representative local explanations are shown in Appendix F, Figure F3 and Figure F4. The prominence of intraday timing and lag structure is consistent with documented commuting cycles and weekly regularities in bike-share demand (Torres et al., 2024; Wilkesmann, 2022; Hamad et al., 2021).

At the station level, SHAP results are similar. Rolling and lagged departures (*departures_roll24h*, *departures_lag168h*, *departures_lag24h*) and hour-of-day encodings (*hour_cos*, *hour_sin*) dominate for a station-level hourly Random Forest (Appendix F, Table FII).

For the station-level CNN-LSTM, feature-group ablation provides model-consistent interpretability. Baseline performance on the fixed subset is RMSE = 2.468, MAE = 1.314, and R^2 = 0.500. Permuting the calendar and time group produces the largest degradation (RMSE 3.204; MAE 1.864; R^2 0.157), while permuting lagged and rolling demand features yields a smaller but non-zero degradation (RMSE 2.567; MAE 1.397; R^2 0.459) (Appendix F, Table FIII). The dominant contribution of the calendar and time group aligns with the conceptual emphasis on temporal rhythm as a primary driver of demand variability (Torres et al., 2024; Wilkesmann, 2022).

4.3. Robustness and Limitations Evidenced by Results

Robustness is evaluated via regime-stratified test performance, spatial-context stratification, feature-group sensitivity, and diagnostic evidence from residual behavior and metric stability.

For system-level hourly forecasting, regime robustness is assessed by splitting the test period into a late-pandemic *other* regime and a *recovery* regime (Appendix F, Table FIV). Both Random Forest and XGBoost perform worse in the *other* regime and improve sharply in *recovery* ($R^2 > 0.91$), with relative ordering stable: Random Forest outperforms XGBoost in both regimes. A limitation remains that the chronological split precludes direct held-out evaluation on the acute COVID period. This regime sensitivity is consistent with evidence that forecasting performance can deteriorate under abrupt demand shifts and atypical operating conditions (Hamad et al., 2021).

At the system-level daily aggregation, regime-stratified results show substantially weaker robustness for daily tree-based models: in the *other* regime, daily Random Forest and daily XGBoost have negative R^2 values (−0.13 and −0.41) and improve in *recovery* but remain modest (Appendix F, Table FV). Classical daily models remain strong in both regimes, with SARIMAX yielding lower RMSE and MAE than ETS within each regime slice (Appendix F, Table FVI). Consistent with the baseline discussion and Table III, ETS remains the stronger baseline in overall evaluation across the full test period, while SARIMAX is stronger when evaluated within regime slices.

Spatial-context robustness is assessed using cluster-stratified daily forecasting for commercial, residential, and transit-adjacent clusters (Appendix F, Table FVII). Absolute errors vary substantially by cluster type, with transit-adjacent clusters lowest and commercial clusters highest. R^2 values are negative across cluster types, indicating volatility sufficient for a mean baseline to outperform in variance-explained terms even when RMSE and MAE improve. In this setting, RMSE and MAE remain the primary indicators of practical performance. The variation by cluster type is consistent with evidence that built-environment and socio-demographic context shape ridership intensity and distribution across urban areas (Wei and Zhu, 2023; Gong et al., 2024).

Feature-group sensitivity results clarify where driver inferences are stable. For cluster-daily ETS, removing calendar variables produces relatively small changes, while removing weather variables yields large degradations, especially in commercial and residential clusters (Appendix F, Table FVIII). For SARIMAX, the pattern differs: across cluster types, the specification without weather yields lower RMSE and MAE than the full specification in the 2023–2024 test window (Appendix F, Table FVIII). The ETS sensitivity to weather aligns with literature emphasizing precipitation and temperature effects (Adrianzen, 2025; Ma et al., 2022), whereas the SARIMAX result in this window is not fully aligned with those expectations and indicates specification sensitivity under the fixed evaluation design (Wang, 2016; Hamad et al., 2021).

Limitations are also evidenced by diagnostics and metric instability. For system-level daily baselines, residual diagnostics reject normality and homoskedasticity and indicate remaining autocorrelation for SARIMAX and ETS, consistent with error spikes in Appendix D (Figure D3). At the station level, sparsity and near-zero demand make MAPE unstable (Appendix E, Table EI), motivating reliance on RMSE and MAE. These results echo broader concerns about regime sensitivity, granularity constraints, and stability of inference in operational forecasting settings.

5. CONCLUSION

This dissertation evaluated Citi Bike demand forecasting across spatial system, cluster, and station levels and temporal daily and hourly resolutions using time-ordered splits and leakage controls. No single approach dominated across tasks.

For system-level daily demand in 2023–2024, ETS(A,N,A), s = 7 with exogenous variables delivered the strongest baseline performance: RMSE 15,181, MAE 10,895, and MAPE 14.00%. SARIMAX$(1,0,1)\times(0,1,1)_7$ was comparable but slightly less accurate (RMSE 15,476, MAE 11,095, MAPE 14.22%). Overall, Prophet performed the worst. Among the exogenous inputs, weather variables accounted for most of the incremental improvement, whereas calendar and regime indicators provided limited additional value beyond what was already captured by weekly seasonality.

For system-level hourly demand, Random Forest achieved RMSE 1,181, MAE 688, and R^2 0.899, whereas XGBoost achieved RMSE 1,398, MAE 815, and R^2 0.859. At the cluster level, the hourly results were RMSE 409, MAE 205, and R^2 0.952 for Random Forest, compared with RMSE 495, MAE 238, and R^2 0.929 for XGBoost. The daily cluster-level results were RMSE 11,198, MAE 6,871, and R^2 0.862 for Random Forest, versus RMSE 13,726, MAE 7,868, and R^2 0.792 for XGBoost. At the station level, the CNN-LSTM reduced MAE relative to HGB (0.935 vs. 1.032) but produced a higher RMSE (1.928 vs. 1.556). Across 150 stations, macro-averaged performance was RMSE 1.527 and MAE 0.935, while volume-weighted performance was RMSE 2.846 and MAE 1.854.

Interpretability analyses supported these patterns. SARIMAX coefficients indicated temperature as positive and precipitation and holidays as strongly negative, while weekend and regime indicators were not statistically distinguishable from zero after accounting for weekly seasonality. Random Forest SHAP emphasized intraday timing and autoregressive persistence, with precipitation more influential than temperature. CNN-LSTM ablation showed calendar and time encodings were dominant, lagged and rolling demand added signal, and weather and contextual station features contributed smaller but consistent value.

5.1. Operational and Policy Implications

The results support matching forecasting methods to decision cadence and spatial scale. For daily system-wide planning, ETS(A,N,A), $s = 7$ with exogenous variables provides a strong one-day-ahead baseline (marginally outperforming SARIMAX in 2023–2024) that can inform staffing, fleet movement, and service monitoring. Given that residuals indicate heavy-tailed errors and heteroskedasticity, and that errors increase during shocks, point forecasts should be used with conservative buffers when volatility rises.

For hourly operations, Random Forest is the preferred baseline among evaluated supervised learning models at both system and cluster levels, with stronger generalization than XGBoost in the held-out years. One-hour-ahead forecasts can support tactical rebalancing and intraday resource allocation, but departures from regular temporal rhythms should be expected to increase error.

Cluster-level forecasting provides a practical intermediate layer to guide redistribution priorities across functional areas (residential, commercial, transit-adjacent), while station-level forecasting is best reserved for targeted corridors or high-impact stations.

At the station level, evaluation should match the operational objective (macro-averaged performance for typical-station service quality versus volume-weighted performance for system impact). Given MAPE instability, RMSE and MAE should remain the primary metrics, with percentage-based errors used only where demand is sufficiently above zero to be interpretable.

From a policy perspective, the primary implication is methodological rather than causal: leakage-controlled, reproducible pipelines using open data can yield reliable short-horizon forecasts and interpretable drivers. The results do not establish downstream outcomes (e.g., cost, emissions, equity) without operational experiments or policy evaluations beyond forecasting, so recommendations are best framed as enabling more proactive and transparent decision processes.

5.2. Limitations and Future Work

Several limitations are evidenced directly by the workflow and results. First, daily system-level ETS and SARIMAX residual diagnostics indicated non-normality, heteroskedasticity, and remaining autocorrelation, and all baselines showed higher errors during abrupt shocks. Second, the chronological split prevented leakage but limited direct held-out testing for the most disrupted acute COVID period. Robustness is therefore strongest for late-pandemic and recovery holdouts. Third, station-level forecasting was constrained by sparsity, heterogeneous station activation windows, and metric instability (notably MAPE). Fourth, contextual datasets had limited coverage or were excluded due to quality constraints, including early gaps in traffic volume coverage and subway turnstile data excluded for formatting issues.

Future work should extend evaluation while retaining accuracy, interpretability, and reproducibility constraints. First, evaluate multi-step-ahead forecasting by quantifying error accumulation and stability across horizons at each aggregation level. Second, expand station-level coverage under consistent splits and investigate why CNN-LSTM increased RMSE despite improving MAE, particularly for high-volume stations, while retaining macro versus volume-weighted reporting. Third, explore alternative spatial groupings and the stability of cluster assignments over time while keeping functional labels interpretable for decision-makers. Fourth, evaluate spatially structured models that capture inter-station dependence while maintaining transparent feature-availability rules, documented split logic, and interpretable attribution or sensitivity analyses.

Overall, the dissertation’s empirical evidence supports a cautious but actionable conclusion: forecasting performance and interpretability for Citi Bike demand improve when model choice is matched to aggregation level and time scale, exogenous weather information is incorporated in a leakage-controlled manner, and evaluation is aligned with the statistical realities of volatility and sparsity at fine spatial resolutions.

References

Adrianzen, R. (2025) *Analysing usage patterns in a bike-sharing system by leveraging clustering algorithms: Case HSL City Bikes*. Thesis. Aalto University.

Albuquerque, V., Sales Dias, M. and Bacao, F. (2021) 'Machine learning approaches to bike-sharing systems: A systematic literature review', *ISPRS International Journal of Geo-Information*, 10(2), pp. 1–25.

Betkier, I., Oszczypała, M. and Dawid, W. (2025) 'Bike-sharing systems for sustainable cities: Demand prediction using graph neural networks and spatial aggregation', *Sustainable Cities and Society*, 128, 106479.

Chuwang, D.D. and Chen, W. (2022) 'Forecasting daily and weekly passenger demand for urban rail transit stations based on a time series model approach', *Forecasting*, 4(4), pp. 904–924.

Gong, W., Rui, J. and Li, T. (2024) 'Deciphering urban bike-sharing patterns: An in-depth analysis of natural environment and visual quality in New York's Citi Bike system', *Journal of Transport Geography*, 115, 103799.

Hamad, S.Y.Y., Ma, T. and Antoniou, C. (2021) 'Analysis and prediction of bike-sharing traffic flow – Citi Bike, New York'. IEEE.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning*. 2nd edn. Springer.

Hyndman, R.J. and Athanasopoulos, G. (2021) *Forecasting: Principles and practice*. 3rd edn. OTexts.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) *An introduction to statistical learning*. 2nd edn. Springer.

Ma, X., Yin, Y., Jin, Y., He, M. and Zhu, M. (2022) 'Short-term prediction of bike-sharing demand using multi-source data: A spatial-temporal graph attentional LSTM approach', *Applied Sciences*, 12(3), 1161.

Mitchell, P.P. (2018) *Predicting bike-sharing traffic flow using machine learning*. Thesis. Norwegian University of Science and Technology.

Mohseni Hosseinabadi, G., Nourinejad, M. and Park, P.Y. (2024) ‘Bike-share ridership prediction for network expansion using graph neural networks’. Preprint. York University.

Porat, O., Ben-Elia, E. and Fire, M. (2025) ‘A comprehensive machine learning framework for micromobility demand prediction’, *arXiv*, 2507.02715.

Schröer, C., Kruse, F. and Gómez, J.M. (2021) 'A systematic literature review on applying CRISP-DM process model', Procedia Computer Science, 181, pp. 526–534.

Torres, J., Jiménez-Meroño, E. and Soriguera, F. (2024) ‘Forecasting the usage of bike-sharing systems through machine learning techniques to foster sustainable urban mobility’, *Sustainability*, 16(16), 6910.

Van der Wielen, R. (2024) *Forecasting short-term bike-sharing demand at station level using deep neural networks*. Thesis. Tilburg University.

Wang, W. (2016) *Forecasting bike rental demand using New York Citi Bike data*. Technological University Dublin.

Wei, B. and Zhu, L. (2023) ‘Exploring the impact of built environment factors on the relationships between bike sharing and public transportation: A case study of New York’, *ISPRS International Journal of Geo-Information*, 12(7), 293.

Wilkesmann, F.L. (2022) *Forecast of short-term demand for train station-based round-trip bikesharing using identified determinants*. Thesis. Delft University of Technology.

Xing, X., Wan, L. and Luo, F. (2025) ‘Demand prediction for shared bicycles around metro stations incorporating STAGCN’, *PLOS One*, 20(7), e0328452.

Zhang, C., Wu, F., Wang, H., Zhang, H., Ma, H. and Liu, Y. (2024) ‘A deep reinforcement learning model for a two-layer scheduling policy in urban public resources’, *IEEE Internet of Things Journal*, 11(2), pp. 2712–2728.

Zhou, C., Hu, J., Zhang, X., Li, Z. and Yang, K. (2024) ‘A bike-sharing demand prediction model based on spatio-temporal graph convolutional networks’, *PeerJ Computer Science*, 10, e2391.

APPENDICES

All appendices referenced in this dissertation are available externally at: https://github.com/zoltanjelovich/citi-bike-demand-forecasting.