# Predicting Bike Demand at Indego Stations in Philadelphia

**Rui Jiang**

## Introduction

Bike-sharing systems rely on accurate short-term demand forecasting to support service reliability and efficient rebalancing. Predicting station-level demand is challenging because usage varies with time of day, weather, land-use context, and neighborhood characteristics.

This study focuses on forecasting the next-hour demand at each Indego station in Philadelphia. We develop a spatio-temporal deep learning model that integrates the station connectivity graph with time-varying node features to improve prediction accuracy.

## Data & Preprocessing

We construct a multi-year station–hour dataset combining:

**Indego trip data (2023–2025):** hourly demand per station.

**Weather:** temperature, precipitation, wind (ASOS hourly).

**Spatial features:** station coordinates and pairwise distances.

**Amenities:** counts and distances of bus shelters, trolley stops, regional rail stations, schools, hospitals, and parks within 500 m.

**Temporal indicators:** hour, weekday, month.

**Socioeconomic:** median household income per station catchment (ACS).

Preprocessing includes cleaning missing coordinates, clipping amenities to the city boundary, merging data sources, and standardizing numerical features.

## Problem Formulation

Each station is represented by a time-varying feature vector combining historical demand, weather, temporal indicators, socioeconomic attributes, and amenity context.

A 24-hour sequence of these node features is used as input:

**Input** $X$ **:** a sequence of graph-structured node feature matrices

**Output** $y$ **:** next-hour demand for all stations

The goal is to learn a function

$$f(X_{t-23},\dots,X_t, A) \rightarrow y_{t+1}$$

that produces accurate next-hour station-level demand forecasts.

## References


Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, *54*, 101882. https://doi.org/10.1016/j.scs.2019.101882

Faghih-Imani, A., Eluru, N., & Paleti, R. (2017). How bicycling sharing system usage is affected by land use and urban form: Analysis from system and user perspectives. European Journal of Transport and Infrastructure Research, 17(3). https://doi.org/10.18757/ejtir.2017.17.3.3205

Wong, V. W. H. (2024). Enhancing Data-Driven Predictive Modeling of Pedestrian Crowd Flow with Spatial Priors – Case Studies with Post-Event Crowd Data on a University Campus. *2024 IEEE International Conference on Big Data (BigData)*, 6795–6804. https://doi.org/10.1109/BigData62323.2024.10825781


## Model Architecture (GNN + GRU)

Our architecture captures both spatial dependencies and temporal dynamics.

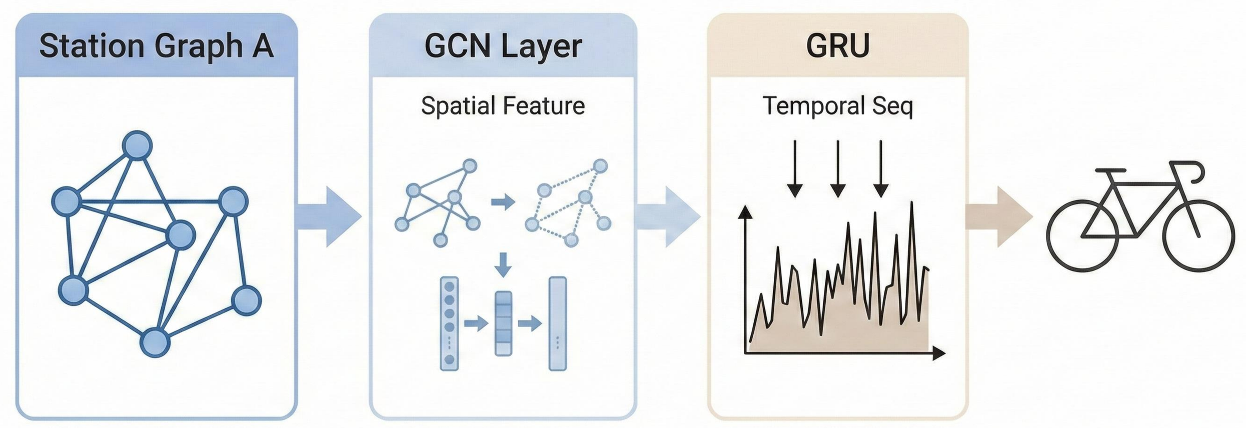
Graph Construction:
Stations are nodes; edges reflect geographic proximity based on pairwise distance.

GCN Layer:
Extracts spatially smoothed node embeddings by aggregating neighbor information using the station graph.

GRU Encoder:
Models short-term temporal patterns over the past 24 hours.

Prediction Layer:
Maps the final GRU hidden states to next-hour demand for each station.

This hybrid design leverages spatial priors and temporal continuity simultaneously.




## Results & Error Analysis

The GNN+GRU model achieves **lower MAE** than the baseline on both validation and test sets, indicating improved accuracy for typical demand ranges. RMSE remains similar because both models struggle with irregular spikes.

The architecture produces smoother and more realistic predictions during morning and evening peaks, and spatial features help reduce errors at high-volume stations in Center City and University City.

Error analysis shows:

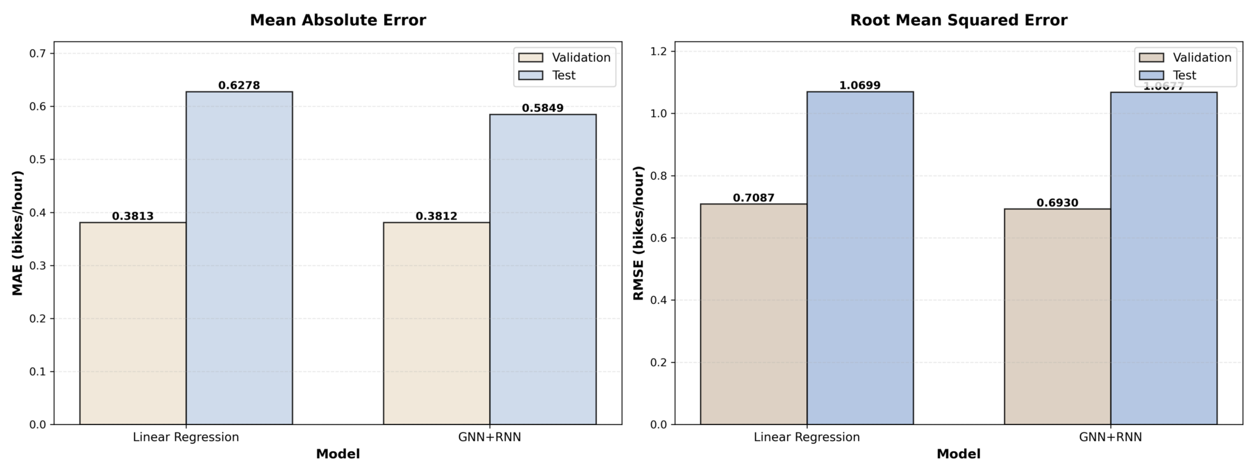
**Good performance** at stations with consistent temporal patterns and strong spatial connectivity.

**Underestimation** of high-demand spikes (4–10+ bikes), visible as compressed predictions in the scatter plots.

**Higher errors** at low-volume or event-driven stations (universities, parks).

**Peripheral stations** show correlated residuals, suggesting limitations in using static geographic distance for graph construction.

These findings highlight the benefit of joint spatial–temporal modeling while revealing challenges in capturing rare peaks and irregular usage.




## Baselines & Evaluation Metrics

We use a feature-based linear regression model as the baseline.

The model takes as input lagged demand, hourly weather conditions, temporal indicators (hour, weekday, month), median household income, and amenity-based spatial context (counts and distances of nearby transit stops, schools, hospitals, and parks).

This baseline captures global linear trends but cannot model nonlinear relationships, spatial dependencies between stations, or temporal dynamics beyond a single-step lag.

Evaluation Metrics.

Performance is evaluated using:

- MAE – average absolute prediction error
- RMSE – penalizes large errors and reflects performance during demand spikes

We report test-set MAE and RMSE for both the baseline and the proposed GNN+GRU model.

## Conclusion

We present a spatio-temporal deep learning framework for station-level bike demand forecasting in Philadelphia's Indego system.

The GNN+GRU model provides **measurable improvements in MAE** and produces more stable predictions in regular demand ranges. Despite limited GPU capacity preventing full convergence, the model demonstrates the usefulness of incorporating spatial priors into temporal forecasting.

**Future extensions include:**

Dynamic graph edges driven by real-time trip flows.

Integration of event, holiday, and land-use activity data.

Attention-based models (GAT, transformers).

Multi-step forecasting and uncertainty quantification.

Accurate demand prediction supports system rebalancing and operational planning for bike-share systems.