Predicting Bicycle Demand by Incorporating Station-Specific Spatial Features in a Graph Convolustion Networks Model

First Author[1], Second Author[1], Third Author[1] and Third Author[2]

[1] Institution/Department, Affiliation, City, Country, {test1-1, test1-2}@ email.com
[2] Institution/Department, Affiliation, City, Country, test2@email.com

Abstract

This research aims to accurately predict bicycle rental volumes at each station to address supply-demand imbalances. However, previous studies on bicycle demand prediction in Seoul did not consider the relationships between the city's various bicycle stations, nor did they take into account the rental volume and weather in conjunction with the specific spatial characteristics of each station. This paper employs a Graph Convolutional Network (GCN) based model to reflect the relationships between bicycle stations by designating each station in Seoul as a node and setting the rental volume between stations as the weight. Additionally, to consider the previously overlooked rental volume and weather characteristics alongside spatial features, we constructed a GCN model with a feature matrix combining spatial information and rental data, and then proposed a new method that calculates the daily demand at each station using a Fully Connected Network (FCN) model inclusive of weather variables. Furthermore, this approach generally achieved better performance compared to other models.

Keywords— *Deep Learning, Demand Forecasting*

I. Introduction

The burgeoning global interest in eco-friendly and healthy lifestyles has spurred a rise in demand for shared bicycle systems worldwide, highlighting the critical issue of supply-demand imbalances. Traditional manual rebalancing methods, reliant on labor-intensive truck deployments, underscore the need for accurate real-time demand prediction to mitigate these disparities. In Seoul, the rapid expansion of the public bicycle system, 'Ttareungi', without proper demand analysis, has led to increased user dissatisfaction due to supply-demand mismatches, necessitating precise demand forecasting for effective bicycle station placement and public bicycle distribution.

The complexity of predicting bicycle demand stems from the non-uniformity of rental and return locations, requiring consideration of the interconnectivity between stations, especially when forecasting across all of Seoul's rental stations. Additionally, spatial factors such as nearby bicycle paths, transportation facilities, and residential areas must be integrated with demand and weather variables to accurately influence rental volumes.

In our study, we construct a graph-based model where Seoul's bicycle stations are nodes connected by edges representing rental volumes, capturing the complex inter-station relationships. This model, a Graph Convolutional Network (GCN), integrates spatial and temporal data, including weather conditions, into its input matrix for a holistic spatio-temporal analysis. Subsequently, a Fully Connected Network (FCN) leverages this integrated data to forecast daily rental demand with enhanced accuracy over traditional models.

II. Data Collection and Preprocessing

A. *Bicycle rental history data*

We downloaded the daily usage information data for Seoul public bike rental stations from "Seoul Open Data Plaza." This dataset contains rental information for each station, and we conducted preprocessing to group this information by rental station, calculating the daily rental counts for each station.

B. *Spatial data*

We utilized spatial data to extract information about facilities within a 500-meter radius of Seoul bicycle stations. Data downloaded from "Seoul Open Data Plaza" and "Smart Seoul Map" included information about bus stops, subway stations, schools, universities, apartments, cultural facilities, and large academies. To determine the count of these facilities around each bicycle station, we performed spatial data preprocessing. We excluded buildings that closed or were demolished before 2022 and those that were constructed or opened after 2022.

Regarding the processing of bicycle roads in the spatial data, we connected bending points to form roads with a radius of 500 meters centered around bicycle stations. We addressed two scenarios: in the first scenario, we counted roads if at least one bending point was within a 500-meter radius to prevent duplicates. In the second scenario, when all bending points were outside the 500-meter radius, we used linear equations and the distance from the bicycle station to the intersection point to count roads that were less than 500 meters. These classified road variables were divided into A and B types, separate and unseparated, and categorized according to road characteristics as shared roads, dedicated roads, priority roads, and car-only lanes, considering the significance of bicycle priority roads from previous research.

Through this preprocessing, we generated spatial data consisting of 15 variables within a 500-meter radius for each bicycle rental station.

C. *Time series data*

We downloaded ASOS data from the "Korea Meteorological Administration" for Seoul to obtain nine weather variables for demand prediction. Although there were many missing values in the precipitation variable, we replaced NA with 0, considering it as an absence of rain.

Through this preprocessing, we obtained a time-series format of nine weather variables from January 1, 2022, to December 31, 2022.

III. Methodology

A. *Framework*

The general framework of Graph Convolutional Network (GCN) models typically comprises nodes and edges representing relationships between nodes. In this study, bicycle stations are designated as nodes, while the relationships between stations are defined as edges. Notably, the relationships between stations are structured using historical data on bicycle rentals and returns. The proposed GCN model in our paper incorporates novel spatial characteristics specific to individual stations, a feature absent in prior literature. It predicts bicycle demand by integrating station-to-station relationships, past bicycle rental records, and contextual spatial data such as proximity to subway stations, bus stops, and residential areas.

We constructed matrices to encompass the spatial information of nearby subway stations, bus stops, residential facilities, among others, incorporating data for bicycle rentals in the days immediately preceding the prediction date, as well as over several weeks. These matrices were concatenated to form a feature matrix. Consequently, the model was designed to learn the spatial characteristics of stations and historical bicycle usage patterns. Weather variables were included, and to facilitate smoother operations between the adjacency matrix and the feature matrix, their outputs were transformed into one-dimensional format. At the conclusion of this transformation, weather variables for the prediction date, such as 'precipitation' and 'temperature', were allocated. These components were integrated into a Fully Connected Network (FCN) model, resulting in the calculation of the anticipated demand for each station on the predicted date.

B. *Graph convolution network*

The Graph Convolution Network (GCN) represents an advancement within the Graph Neural Network framework, addressing certain limitations of GNN models. GNNs propagate information through several layers; however, this process fails to adequately consider interactions among nodes located at considerable distances within the graph. This limitation results in the inefficiency of information processing due to the substantial computational load involved.

In contrast, GCN has emerged as a model that overcomes these specific drawbacks. It effectively computes interactions among bicycle stations while concurrently capturing the distinct characteristics of individual stations. By doing so, GCN resolves the inadequacy in considering interactions among distant nodes within the graph and efficiently manages the computational complexity involved in the process.

Through its refined architecture, GCN offers a more efficient and comprehensive approach to modeling interactions between bicycle stations, providing a promising solution to the challenges previously faced by conventional GNNs.

C. *Adjacency Matrix: Interaction among stations*

The study represents the interactions between bicycle stations using total historical records of bicycle rentals and returns. The rows and columns in the adjacency matrix correspond to the station numbers with identical station numbers. Rows signify the rental station numbers, while columns denote the return station numbers. The adjacency matrix is constructed by summing the occurrences of rental and return station number pairs across all historical rental and return records:

$$A_{ij} = \sum RR_{ij}$$

Here, A_{ij} represents the adjacency matrix containing information on rentals from station i and returns to station j, and $\sum RR_{ij}$ signifies the total count of bicycles moved from station $itostation$j.

$$A_{ij}^{\text{scaled}} = \frac{(A_{ij} - \text{Min})}{\text{Max} - \text{Min}}$$

Here, A_{ij}^{scaled} is the scaled matrix, where Max and Min represent the maximum and minimum values within matrix A_{ij}. The scaled matrix, denoted as $\tilde{A}$(assumed from A_{ij}^{scaled}), is further utilized in forming the normalized symmetric Laplacian matrix:

$$L^{\text{sym}} = \tilde{D}^{(-1/2)} \tilde{A} \tilde{D}^{(-1/2)}$$

Here, $\tilde{D}$ is the diagonal matrix derived from matrix $\tilde{A}$ and L^{sym} represents the normalized symmetric Laplacian Matrix obtained by multiplying both sides of the scaled adjacency matrix with the power of -1/2 of the diagonal matrix. Constructing a symmetric Laplacian Matrix serves the purpose of efficient information management when performing operations with the feature matrix in eigenvalue-eigenvector forms.

D. Feature Matrix 1: spatial feature by station level

This research introduces a novel approach to constructing spatial matrices that have previously not been explored. It includes the spatial characteristics specific to individual bicycle stations. The matrix encapsulates parameters such as the number of transportation facilities, residential households in nearby apartments, educational infrastructure, and other related features within a 500-meter vicinity of each station.

The rows of the matrix correspond to a total of N unique bicycle stations, while the columns represent 15 distinct spatial characteristics.

$$F_{\text{spatial}}^{\text{scaled}} = \frac{F_{\text{spatial}} - \text{Min}}{\text{Max} - \text{Min}}$$

IV. RESULT

In this study, we compared the performance of the GCN-pro model with other models. We investigated the performance differences between models that utilize only weather data, specifically time series data, such as LSTM, ARIMA, and XG-Boost, and the GCN-pro model, which incorporates spatial data. Below are the RMSE values for each model.

model	RMSE
CGN-pro	15.48
GCN-up	00.0
LSTM	16.9
ARIMA	00.0
XG-Boost	00.0

Table 1. RMSE of models

REFERENCES