Naqsha-e-Safar

Identifying & Classifying Public Transit Deserts
in Lahore Using Graph Neural Networks

Information Technology University
Department of Computer Science

Mujtaba Asim	bscs22012@itu.edu.pk
Muhammad Rehman	bscs22018@itu.edu.pk
Aleena Basit	bscs22005@itu.edu.pk
Abdullah Raffay	bscs22085@itu.edu.pk
Eman Afrooz	bscs22099@itu.edu.pk

December, 2025

Abstract

Public transportation accessibility is fundamental to urban equity and sustainable development. This project addresses the challenge of identifying "transit deserts" in Lahore—areas characterized by high population density yet insufficient access to public transport infrastructure. We constructed a spatial graph of the city by aggregating population estimates from WorldPop, transit route data from CityLines and Zameen, and the road network from OpenStreetMap. By applying a Graph Convolutional Network (GCN), we classified approximately 146,000 points of interest into transit-accessible and transit-deficient regions. Our methodology partitions the city into approximately 152 spatial communities using the Louvain algorithm and employs a heuristic weak supervision strategy to generate training labels from high-confidence examples. The trained GCN successfully identifies underserved areas that warrant prioritization in future transport planning. While acknowledged to be an approximation due to data limitations, this study represents a significant step toward data-driven transit equity analysis in Lahore, with a methodology that is scalable to other developing urban centers.

Introduction

Background and Motivation

Urban transportation systems are the vital backbone of modern cities, enabling economic activity, social connectivity, and access to essential services such as healthcare and education. However, public transit infrastructure is frequently distributed inequitably across urban landscapes. This disparity creates "transit

deserts": areas where residents possess limited access to reliable public transport despite significant need.

Lahore, as Pakistan's second-largest metropolis with a population exceeding 13 million, acutely faces this challenge. Despite recent infrastructural investments—such as the Orange Line Metro Train and the Speedo bus network—substantial portions of the city remain underserved. Transit deserts are not defined solely by Euclidean distance to the nearest stop; rather, they represent the intersection of high population density, limited modal options, and consequent socio-economic exclusion. Residents in these areas often incur higher commute costs and face reduced access to employment opportunities.

For urban planners attempting to allocate resources equitably, accurate identification of these zones is essential. Traditional accessibility metrics—such as simple distance buffers or isochrone analysis—often fail to capture the complexity of urban mobility. They frequently overlook network topology, heterogeneous population distributions, and the spatial dependencies between neighborhoods. This research aims to bridge this gap by proposing a graph-based deep learning approach.

Research Objectives

The primary objective of this study is to develop a data-driven framework for identifying transit deserts in Lahore using Graph Neural Networks (GNNs). Specifically, we aim to:

1. **Construct a Spatial Graph**: Model Lahore as a graph where nodes represent points of interest (POIs) and edges represent road connections, integrating disparate population and transit datasets.
2. **Engineer Features**: Develop node features that encapsulate demographic characteristics, topological properties, and geographic context.
3. **Detect Communities**: Utilize the Louvain algorithm to partition the city into spatial communities, approximating natural urban clusters.
4. **Classify Nodes**: Train a Graph Convolutional Network (GCN) to distinguish between transit-adequate and transit-deficient areas using a weak supervision approach.
5. **Generate Insights**: Produce visualizations and metrics to assist urban planners in evidence-based decision-making.

Contributions

This work offers several key contributions:

- **Novel Application**: We present one of the first applications of GNNs for transit desert identification in the context of a developing country, where data scarcity is a prevalent challenge.
- **Scalable Framework**: The methodology relies entirely on open-source datasets and is modular, allowing for adaptation to other cities with minimal architectural changes.
- **Weak Supervision Strategy**: We introduce a heuristic labeling strategy to generate training data from "obvious" high-confidence examples, bypassing the need for expensive manual annotation.
- **Practical Validation**: We demonstrate the efficacy of the approach on real-world data from Lahore, yielding results that align with intuitive understandings of the city's underserved regions.

Related Work

Transit Accessibility Analysis

Traditional approaches to measuring transit accessibility predominantly utilize distance-based metrics, such as the standard "400m walking distance" buffer around stops, or isochrone maps depicting areas reachable within specific timeframes. Gravity models, where accessibility is weighted by destination attractiveness, remain popular. While straightforward and interpretable, these methods often struggle to account for complex interactions between population density, multi-modal transit options, and the actual topology of road networks. Recent studies have incorporated GIS-based techniques for more nuanced assessments, yet many rely on relatively simple spatial heuristics.

Graph Neural Networks in Urban Computing

GNNs have gained significant traction for modeling relational data in urban environments. They have been successfully applied to traffic prediction, air quality forecasting, and urban land use classification. The primary advantage of GNNs lies in their ability to aggregate information from neighboring nodes via message passing, thereby capturing the spatial autocorrelation inherent in urban data. Despite this, their application to transit equity analysis remains limited, particularly in the Global South where urban challenges differ distinctively from developed nations.

Community Detection in Urban Graphs

Community detection algorithms identify clusters of densely connected nodes within graphs. The Louvain method is widely utilized due to its computational efficiency on large networks. In urban contexts, it has been used to delineate neighborhoods based on mobility patterns or functional structures. While algorithms like DBSCAN are also common, graph-based partitioning provides a natural method to group fine-grained spatial data into regions relevant for planning.

Methodology

Data Collection and Preprocessing

Data was aggregated from multiple open-source repositories to construct a comprehensive view of Lahore's urban environment.

Population Density Data

Population estimates were sourced from the WorldPop project, which provides gridded data at a 1km $\times$ 1km resolution derived from census records and satellite imagery. A significant challenge involved integrating this raster data with our vector graph representation. Initial attempts to assign population based on nearest grid cells proved ineffective, as they artificially inflated density in areas where multiple nodes clustered near a single high-population cell. To rectify this, we distributed each population reading evenly among all nodes within a 1km radius. For instance, if a cell containing 10,000 people encompassed 50 nodes, each node was assigned a value of 200. This approach yields a more realistic representation of population distribution across the graph.

Transit Infrastructure Data

Transit route data was derived from two primary sources:

- **CityLines**: Used to extract the geometry of the Orange Line Metro Train.
- **Zameen**: Provided route information for the Speedo bus system.

We manually cleaned the Speedo dataset, geocoding addresses and validating stop locations. The final dataset includes 519 transit nodes derived from CityLines data (representing the Metro and Orange Line infrastructure) and 70 unique Speedo bus stops. It is acknowledged that this dataset is not exhaustive, particularly regarding informal transport (e.g., Qingqis), but it represents the most reliable publicly available data.

Road Network and Points of Interest

We utilized OSMnx to retrieve Lahore's road network and points of interest (POIs) from OpenStreetMap (OSM). Filtering for relevant locations (schools, hospitals, markets, etc.) resulted in a graph comprising approximately 146,000 nodes and roughly 200,000 edges.

Graph Construction

The city is modeled as a spatial graph $G = (V, E)$ where V represents POIs and E represents physical road segments.

Node Insertion and Connectivity

The transit stops were inserted as additional nodes. Crucially, to prevent these stops from forming isolated components, artificial edges were created connecting each stop to its nearest existing POI or road intersection. This step ensures that the transit nodes are fully integrated into the graph topology, allowing the community detection and message-passing algorithms to function correctly.

Feature Engineering

Each node $v \in V$ is assigned a feature vector capturing its urban context:

- **Geometric**: Latitude and longitude (x, y).
- **Demographic**: Population density p_v (distributed from WorldPop).
- **Topological**: Node degree, indicating road connectivity.
- **Accessibility**: Network distance to the nearest transit stop d_v, computed along the road network rather than via Euclidean distance.
- **Community**: One-hot encoded membership derived from the Louvain algorithm.

Community Detection

We employed the Louvain method to partition the graph. It is important to note that Louvain optimizes for **modularity**—the density of links inside communities compared to links between communities—rather than spatial compactness or diameter. For our graph, this resulted in approximately 152 communities. These clusters represent mathematical partitionings of the graph based on connectivity and do not necessarily align with administrative "societies" or neighborhoods, a limitation we account for in our analysis.

We briefly explored DBSCAN, but it produced ambiguous results with our initial parameters. Given the robustness of Louvain for large graphs, it was selected as the primary method.

Label Generation Strategy

Lacking ground-truth labels for "transit deserts," we employed a weak supervision strategy to generate labels for training. We defined heuristic rules to identify "obvious" cases:

Let μ_p be the median population density. We assigned labels $y_v \in \{0, 1\}$ as follows:

$$y_v = \begin{cases} 1 \text{ if } p_v > \mu_p \text{ and } d_v < 1500\text{m} \\ 0 \text{ if } p_v < \mu_p \text{ and } d_v > 4500\text{m} \\ \text{unlabeled otherwise} \end{cases}$$

We utilized strict thresholds (1500m for accessibility, 4500m for lack thereof) to ensure that only high-confidence nodes were used for training. Nodes with $y_v = 1$ represent acces-

sible, high-density areas, while $y_v = 0$ represents low-density areas far from transit. The "unlabeled" nodes—comprising the vast majority of the graph—are those where the status is ambiguous, which the model aims to classify.

Graph Neural Network Architecture

We implemented a Graph Convolutional Network (GCN) for node classification. GCNs were chosen for their efficiency in aggregating features from local neighborhoods. The architecture consists of two GCN layers, each with 64 hidden units.

Mathematically, the layer-wise propagation rule of the GCN is defined as:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

Where:

- $\tilde{A} = A + I_N$ is the adjacency matrix of the graph with added self-loops.
- $\tilde{D}$ is the degree matrix of $\tilde{A}$.
- $W^{(l)}$ is the layer-specific trainable weight matrix.
- $H^{(l)}$ is the matrix of activations in the l-th layer, with $H^{(0)}$ being the input feature matrix.
- σ is the activation function, for which we utilized ReLU.

We initially attempted to use Graph Attention Networks (GATs) to better capture the varying importance of specific road connections. However, this approach was discarded due to prohibitive computational costs (training times exceeding 1.5 hours per epoch) and stability issues. The GCN model provided a balance of performance and efficiency, training in approximately 15 minutes.

Community-Level Aggregation

Post-classification, node-level predictions were aggregated to the community level. For each community C_i, we computed a transit adequacy ratio. Communities with a low ratio of adequately served nodes were flagged as priority transit deserts. This aggregation helps smooth out noise at the individual node level, providing actionable insights for district-level planning.

Results

Model Performance

The GCN demonstrated strong spatial coherence in its predictions on the test set. While the absence of ground-truth labels precludes traditional accuracy metrics, the model effectively generalized from the heuristic "obvious" examples to the ambiguous regions, clustering transit-adequate nodes along major corridors.

Analysis of Critical Underserved Clusters

The quantitative assessment of the accessibility index highlights three distinct clusters with the lowest scores. Manual verification of these coordinates indicates that these regions are not homogenous; rather, they represent three specific typologies of transit exclusion: peri-urban disconnection, density-induced inaccessibility, and industrial corridor gaps.

1. North-East Fringe: Peripheral Agricultural Settlements

This cluster encompasses the settlements of **Taqi Pur**, **Chhapa**, and **Nathoki**. Morphologically, this region is characterized by low-density, agrarian land use that technically falls within the district boundaries but functions as a series of satellite villages. The extremely low accessibility scores here reflect a classic "cov-

erage gap," where the formal transit network terminates at the urban periphery, failing to extend into the rural hinterland. Residents in these areas face a "spatial mismatch," likely necessitating the use of intermediate informal paratransit to access even the outermost nodes of the formal bus network.

2. Central-North: Infrastructurally Isolated Urban Pockets

Geographically central regions, identified as **Harbanspura** and **Aziz Bhatti Town**, emerged as significant transit deserts. This result highlights the issue of "network penetration" rather than mere distance. Despite the proximity of high-capacity transit corridors (such as the Orange Line Metro), the internal depth and density of these mixed-use neighborhoods create a barrier to access. The walking distance from the centroid of these residential zones to the nearest arterial transit stop exceeds the standard 1 km catchment threshold. Consequently, these areas suffer from a "last-mile" connectivity failure, where high density precludes the entry of standard buses, rendering the core transit network effectively inaccessible without feeder modes.

3. South-West Corridor: The Industrial Access Gap

The cluster spanning **Mohlanwal**, **Chung**, and **Sundar** represents a critical misalignment between economic zones and transit planning. This corridor hosts the Sundar Industrial Estate and associated worker colonies. The low index scores here illustrate the "termination effect," where the formal Metro and Speedo bus networks effectively cease operations at Thokar Niaz Baig. This forces a reliance on unregulated transport for the final leg of the commute to major employment centers. Unlike the rural fringe, this region demonstrates high potential demand (commuter flow) that is currently unmet by the formal network's existing operational limits.

Spatial Patterns

- Transit adequacy is strongly correlated with the Orange Line corridor and primary Speedo routes.
- Accessibility drops precipitously beyond a 2-3 km radius from these corridors.
- POI density is notably higher in the city center, likely introducing a bias in the GCN's attention toward well-mapped central areas.

Limitations and Caveats

It is crucial to acknowledge several limitations:

- **Incomplete Data**: The dataset includes only 70 Speedo stops and the Metro/Orange Line. Informal transport (Qingqis, vans), which serves a massive portion of Lahore, is not modeled.
- **Community Skew**: The "communities" detected by Louvain are graph-theoretic clusters. They serve as a proxy for neighborhoods but do not perfectly map to real-world residential societies.
- **OSM Bias**: The completeness of OpenStreetMap data varies by neighborhood wealth, potentially skewing density estimates.

Discussion

Methodological Insights

The use of GNNs allows for the integration of topological, demographic, and geometric features into a unified framework. The "artificial edges" created during graph construction proved critical; without them, transit stops would have remained isolated, rendering the

message-passing mechanism ineffective for those nodes.

The weak supervision strategy demonstrated that useful classifiers can be trained without manual annotation, provided the heuristic anchors are sufficiently strict. The choice of 1500m and 4500m thresholds was deliberate to avoid noise in the training set.

Ethical Considerations and Data Bias

A critical consideration in this analysis is the "Digital Divide" inherent in crowdsourced mapping platforms like OpenStreetMap (OSM). In Lahore, affluent neighborhoods such as DHA and Gulberg typically exhibit higher fidelity in OSM data regarding road networks and points of interest (POIs) compared to lower-income areas or informal settlements (katchi abadis). This discrepancy introduces a potential bias in our model: areas with sparse data may be algorithmically misinterpreted as "low density" rather than "unmapped." Consequently, the model might under-prioritize legitimate transit deserts in marginalized communities simply because they lack digital visibility. Future iterations of this work must integrate ground-truth validation or supplementary datasets to mitigate this socio-technical bias.

Implementation Challenges

We encountered significant implementation hurdles regarding Coordinate Reference Systems (CRS). Discrepancies between datasets (EPSG:4326 vs. EPSG:32643) initially led to erroneous distance calculations, requiring a rigorous standardization of the pipeline. Furthermore, managing memory overhead within the interactive development environment required aggressive garbage collection and state management, particularly during the handling of large graph objects.

Practical Implications

For urban planning, our results suggest:

- **Priority Corridors**: Radial routes connecting peripheral deserts to the city center should be prioritized.
- **Last-Mile Connectivity**: Investments in pedestrian infrastructure around existing stops could effectively shrink "interstitial" deserts without requiring new vehicle fleets.
- **Demand-Responsive Transit**: Low-density peripheral areas may be better served by flexible micro-transit rather than fixed-route buses.

Scalability

A significant strength of this framework is its scalability. By relying on global datasets like WorldPop and OpenStreetMap, the pipeline can be adapted to other cities—such as Karachi or Rawalpindi—with minimal modification. The modular design allows components (e.g., the community detection algorithm or GNN architecture) to be swapped as improved methods or data become available.

Conclusion

This study demonstrates that Graph Neural Networks offer a viable, scalable pathway for identifying transit deserts in resource-constrained environments. By integrating diverse spatial datasets into a unified graph model, we successfully approximated transit accessibility across Lahore. While the results are an approximation limited by data completeness, they provide a data-driven foundation for transit equity analysis. Future work should focus on incorporating informal transit networks and conducting sensitivity analyses on labeling

thresholds. Ultimately, this project represents a step toward shifting urban planning from ad-hoc decision-making to systematic, equity-centered methodologies.

Future Work

- **Data Enhancement**: Integrating informal transport data (Qingqis, rickshaws) is the most critical next step to reflect ground realities.
- **Temporal Dynamics**: Moving from static analysis to time-varying models that account for peak/off-peak frequency.
- **Equity Overlays**: Incorporating socioeconomic data to identify where transit deserts overlap with vulnerable populations.
- **Validation**: Conducting household travel surveys to ground-truth the model's predictions against resident experiences.