Transit Equity in Pittsburgh: A data-driven approach

Heinz College Capstone, Spring 2023

Mahnoor Ayub, Cailin De Zeeuw, Yuxuan He,
Matt Lampl, Sara Maillacheruvu,
& Sai Tej Reddy Vulupala

FINAL REPORT

POGOH

Our Team

Mahnoor
Ayub

Cailin
De Zeeuw

Yuxuan
He

Matt
Lampl

Sara
Maillacheruvu

Sai Tej
Reddy Vulupala

Agenda

- Project Motivation and Context
- Data overview
- Our approach and results
- Solutions, Challenges, and Future Work

Project Motivation and Context

Data overview

Our approach and results

Solutions, Challenges, and Future Work

Bike Share Pittsburgh

The non-profit organization behind Pittsburgh's only bikeshare program, **POGOH**

Mission

Provide Pittsburgh with a joyful, sustainable, and affordable mobility service for all residents and visitors

POGOH is Pittsburgh's bike share network

- Launched in May, 2022 (replacing the HealthyRide network)
- 37 existing stations, 360 bikes
- 20 charging stations through partnership with Duquesne Light
- Actively expanding the network

Equity in transportation seeks **fairness in *mobility* and *accessibility*** to meet the needs of all community members. A central goal of transportation is to **facilitate *social* and *economic* opportunities** by providing equitable levels of access to ***affordable* and *reliable* transportation** options **based on the needs of the populations** being served, particularly populations that are traditionally underserved.

US Department of Transportation

Project Motivation

Bike share programs produce social and environmental benefits.

- U.S. households produce 9.5 trips per day, half of which are under 3 miles
- Bike-sharing can reduce congestion upwards of 4% within a neighborhood
- While not climate neutral, bikes use less energy and emit less GHG emissions per person-kilometer over their life cycle
- Transport equity, accessibility, physical health improvements
- Bikeshare ridership concentrated among white, wealthier communities

E-bikes could replace 1 in 5 short trips

Data source: Fan and Harper, 2022

Project Objectives

Provide Bike Share Pittsburgh with a new tool for choosing future station locations to maximize equity.

Through our data-driven approach, we will support POGOH in:

- Expanding the established bike-sharing network into new neighborhoods.
- Increasing ridership particularly among disadvantaged communities.
- Improving service within pre-existing bike-sharing communities with increased station density.

Bike Share Pittsburgh can improve transportation equity by intentionally siting bikeshare stations to maximize measures of accessibility

Project Motivation and Context

Data overview

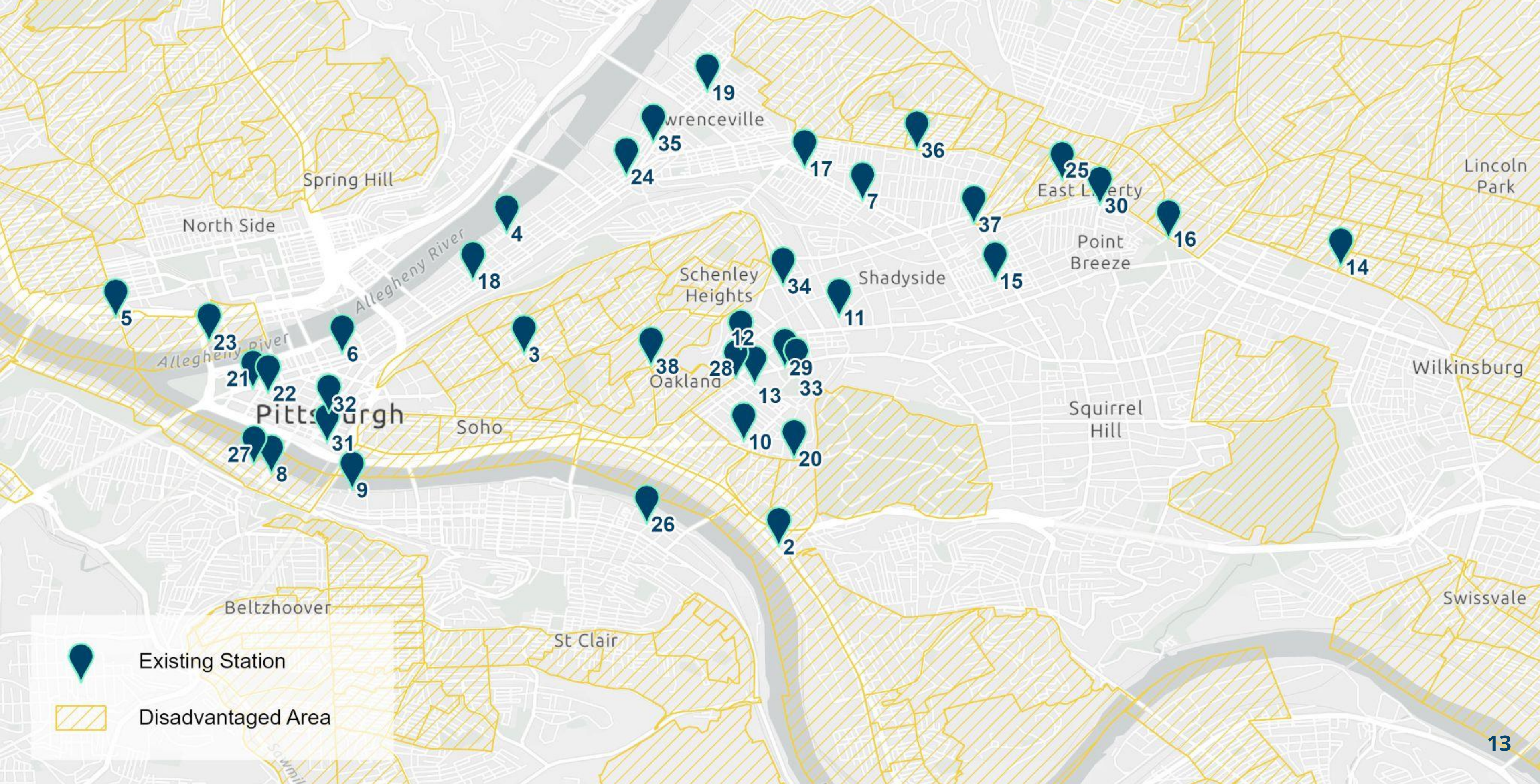
Our approach and results

Solutions, Challenges, and Future Work

Glossary

Trip origins	How many trips **start** at a given station.
Trip destinations	How many trips **end** at a given station.
OD pair	Origin-Destination pair for a given trip. E.g., 21-33 would indicate the trip started at station 21 and ended at station 33.
Disadvantaged area	Block group designated as such if classified as “High Need” or “Very High Need” per Allegheny County’s Community Need Index
Opportunities	Number of hospitals, grocery stores, jobs, etc. someone can access in the Census Block group where a station is located.
Accessibility change	Difference (positive, negative) in opportunities generated by moving from one station to another. Our **proxy** for **transit equity**. E.g., moving from a disadvantaged area to an advantaged area would result in an increased accessibility score.

Current POGOH Stations



Data Sources

POGOH Trip Data

- Number of Docks
- Start Station
- End Station
- Trip Start time
- Trip End time
- RiderID
- BikeModel (Electric or Non-electric)
- Membership status

Station Attributes

- Total Population
- Number of Households
- Percentage of White Population
- Average Age of Population
- Percentage of Young Population (20 - 34)
- Median Income
- Low Vehicle Households
- Labor Force (total population able to work)
- Employment rate
- Bike Path Density
- Park Density
- Number of Transit Stops

Data Cleaning Process

Raw data	Remove grace period trips	Remove trips over 5 hours	Remove trips less than 2 minutes
Row represents **one trip**, made by **one rider**. One rider may take many trips (i.e., they may appear many times in dataset).	***Assumption***: Rider did not actually take trip.	***Assumption***: Rider took the trip for leisure or failed to re-dock.	***Assumption***: Functionally "failed" trips that didn't go anywhere.
Row count: 76,161	***Row count***: 68,829	***Row count***: 68,371	***Final row count:*** 67,780 ***Total rows lost:*** 8,381 ***Percent rows lost:*** 11%

Exploratory Data Analysis Overview

Time:

- Length of trips
- Seasonality of trips
- Time of day in which riders took trips

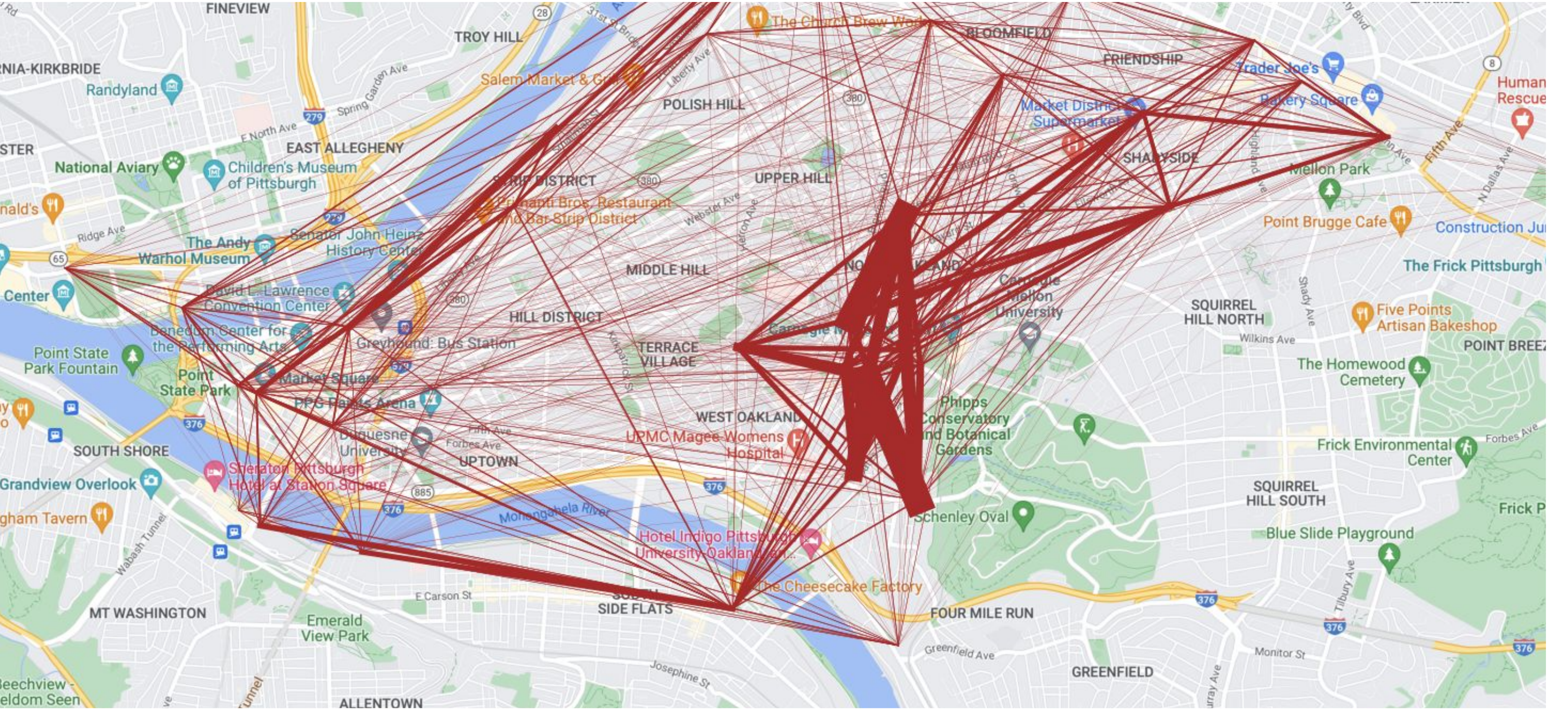
Ridership:

- Trip breakdowns by:
  - POGOH members
  - Casual riders
  - Mobility Justice Membership (MJM) riders

Geography:

- Routes preferred by:
  - MJM riders
  - Member riders
  - All riders
- Trip origins and destination by
  - Disadvantaged block groups
  - Non-disadvantaged block groups

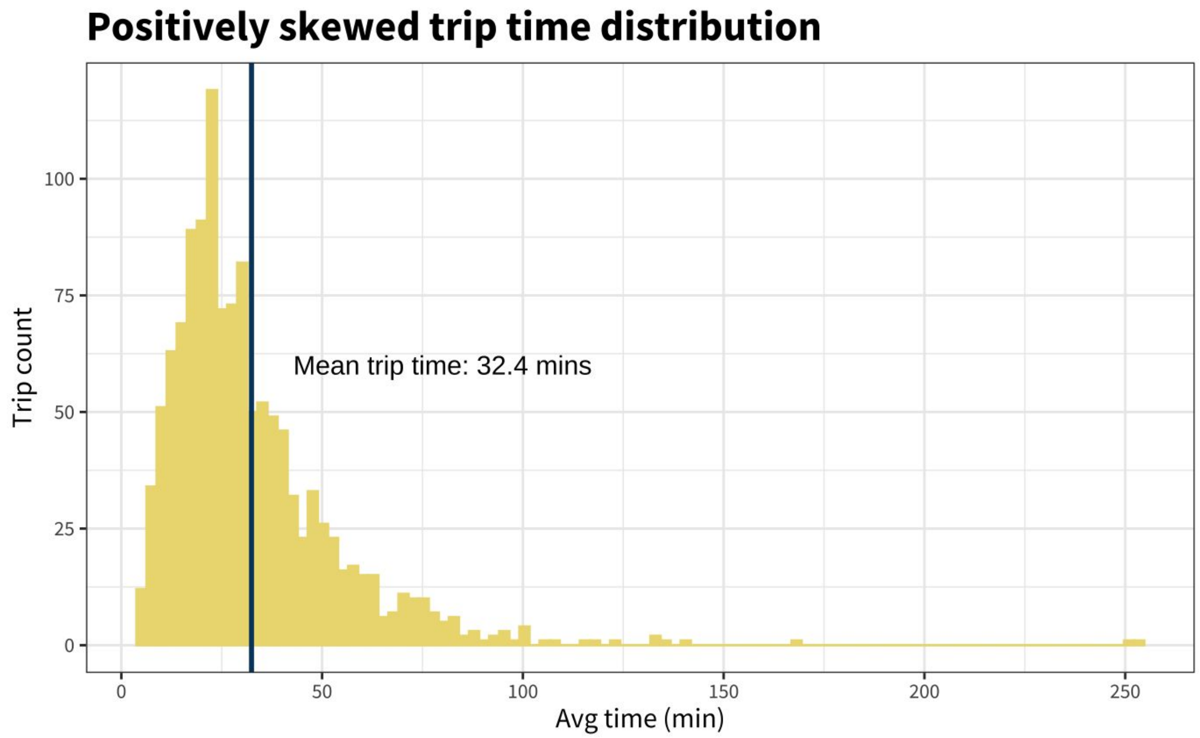
Most Popular Routes



Fewer Trips More Trips

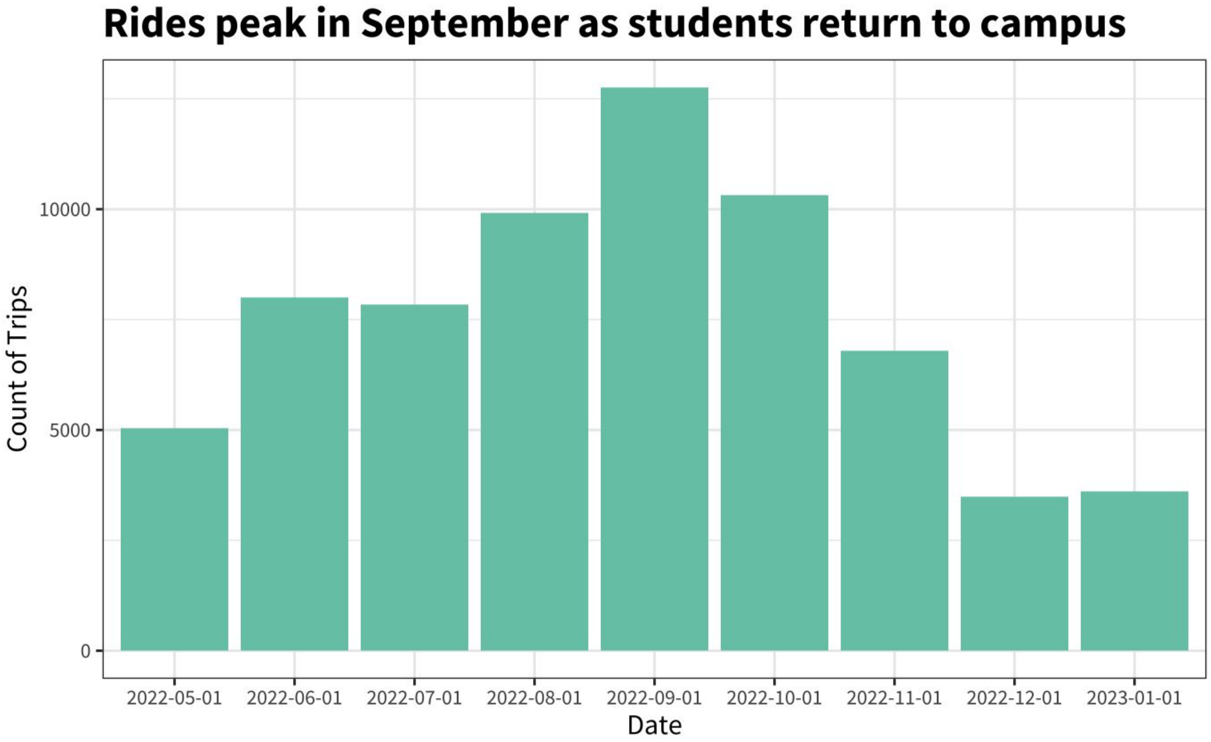
Trips are concentrated in Oakland and downtown.

Length of trips average around 30 minutes



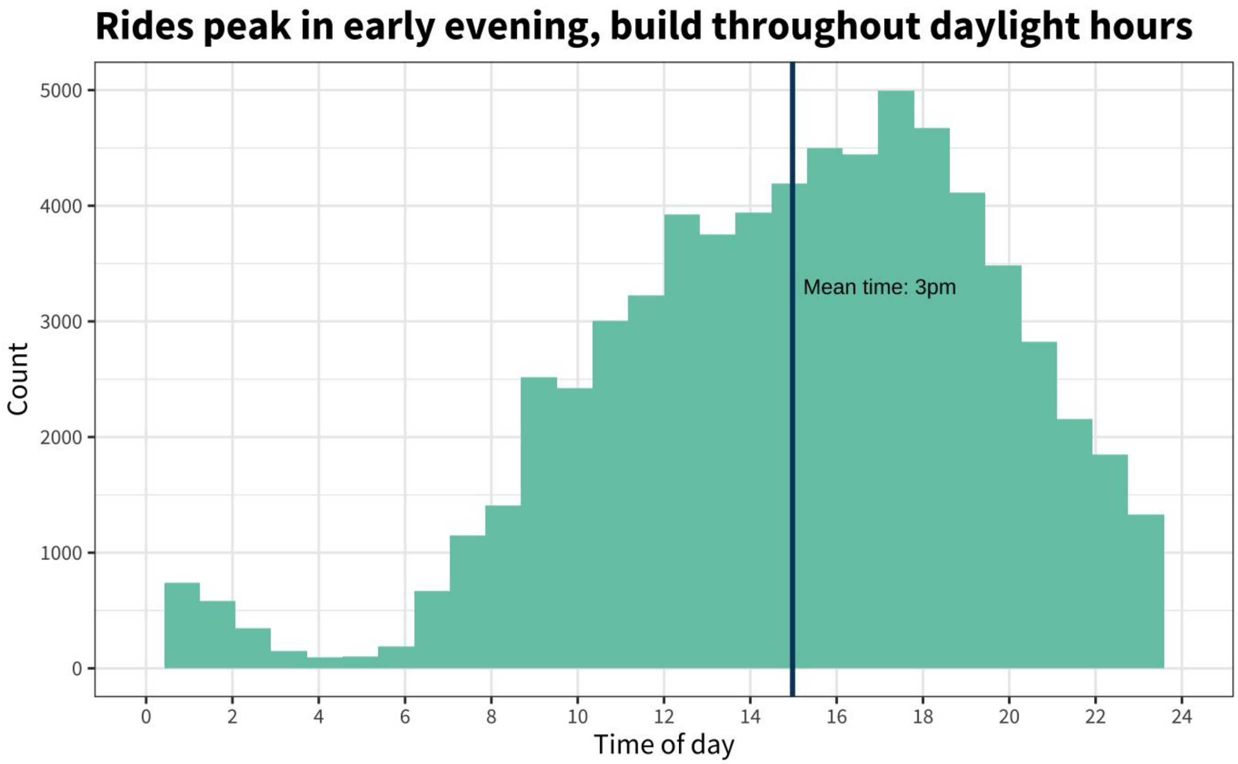
Trip time distribution skews positive, with the mean trip time just above the 30-minute limit.

Time of year affects trip counts



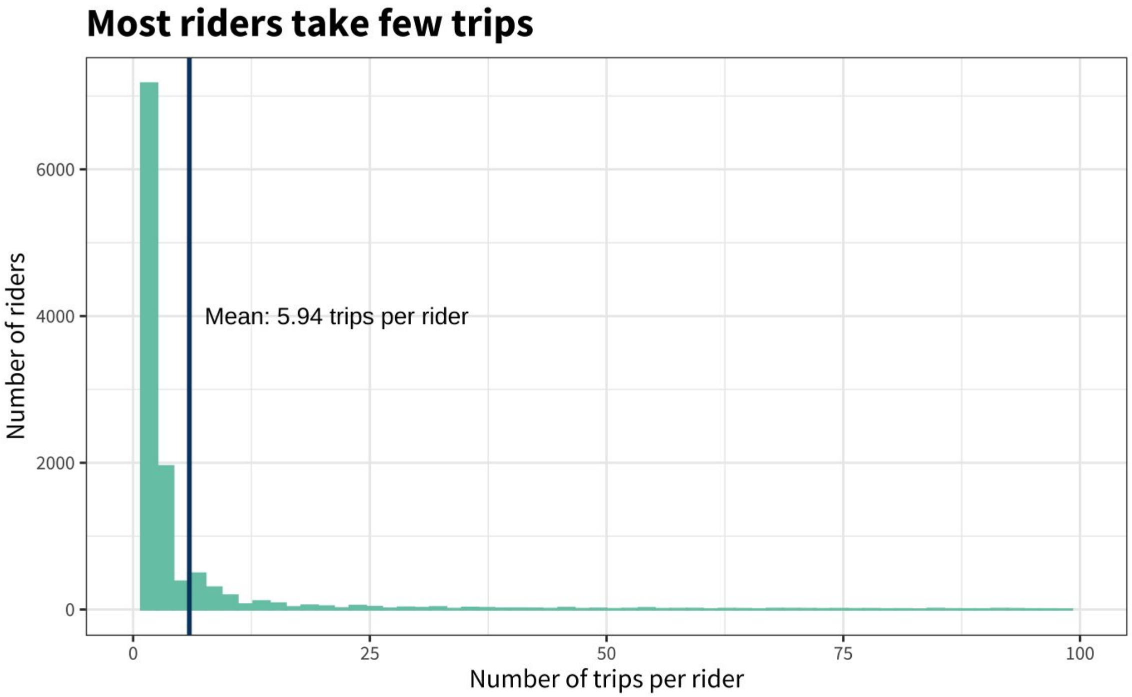
Rides build throughout the warmer months, peaking when college students return to campus.

Trips accrue throughout daytime



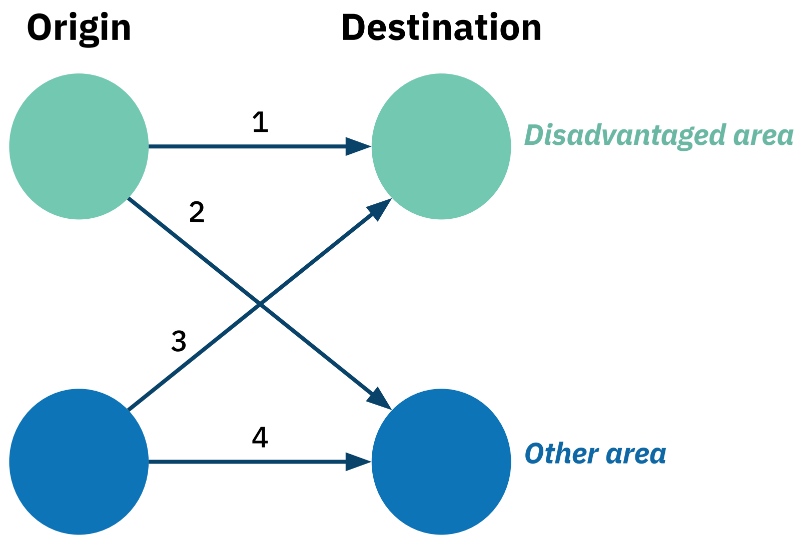
Ridership builds throughout the day, peaking in the mid-afternoon.

Low mean trip count per rider



Trip count per rider is rightward skewed, though the mean is low. Nearly 1% of riders take over 100 trips, and one rider took 717 trips!

Statistics on Routes in Disadvantaged Areas



TRIP COUNT

Origin\Destination	Others	Disadvantaged
Others	47,985 trips	2,805 trips
Disadvantaged	4,340 trips	263 trips

AVERAGE TIME

Origin\Destination	Others	Disadvantaged
Others	23 min	33 min
Disadvantaged	22 min	28 min

Ridership dominated by members, MJM riders take disproportionate number of trips

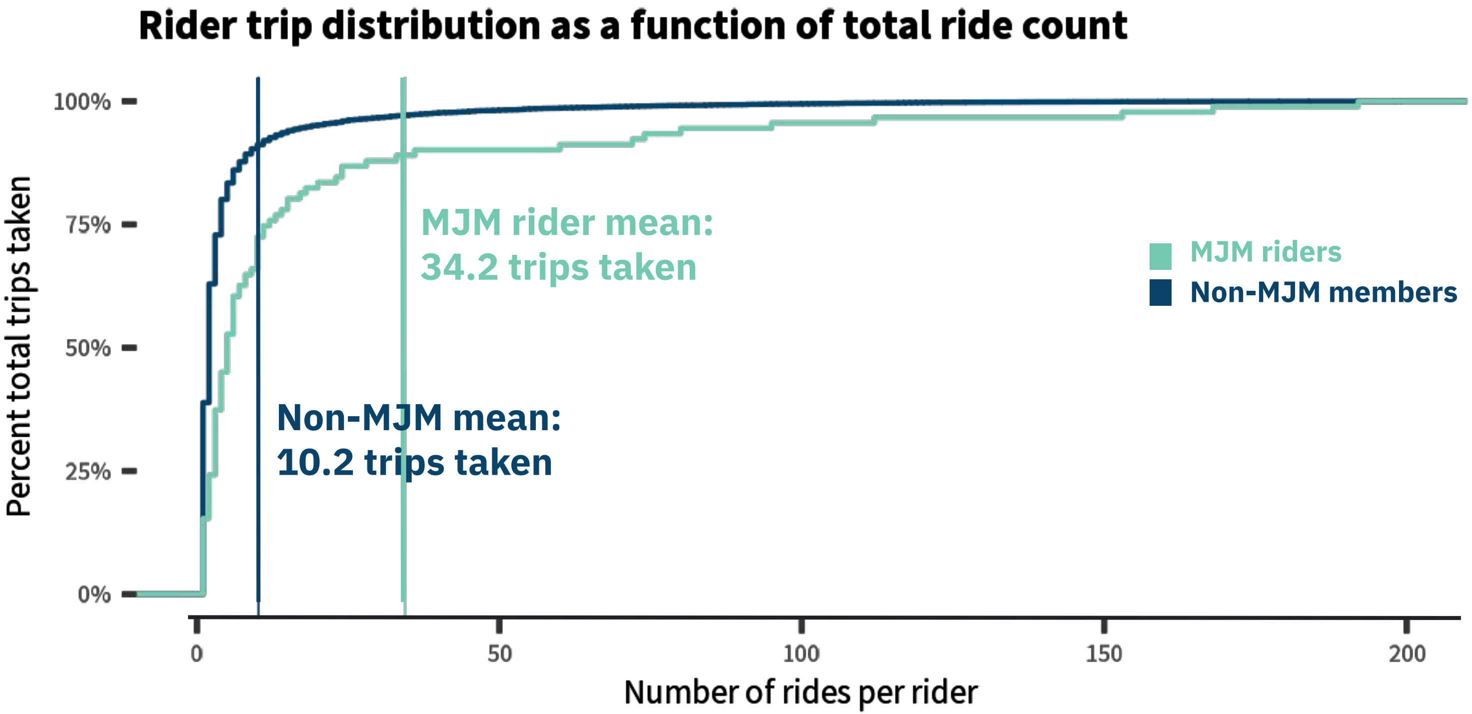
Rider Type	Ride count	Percent of rides
Non-member	18,523	27.3%
Member	49,257	72.7%

While MJM riders only account for 1.1% of POGOH ridership, they take nearly 5% of total trips.

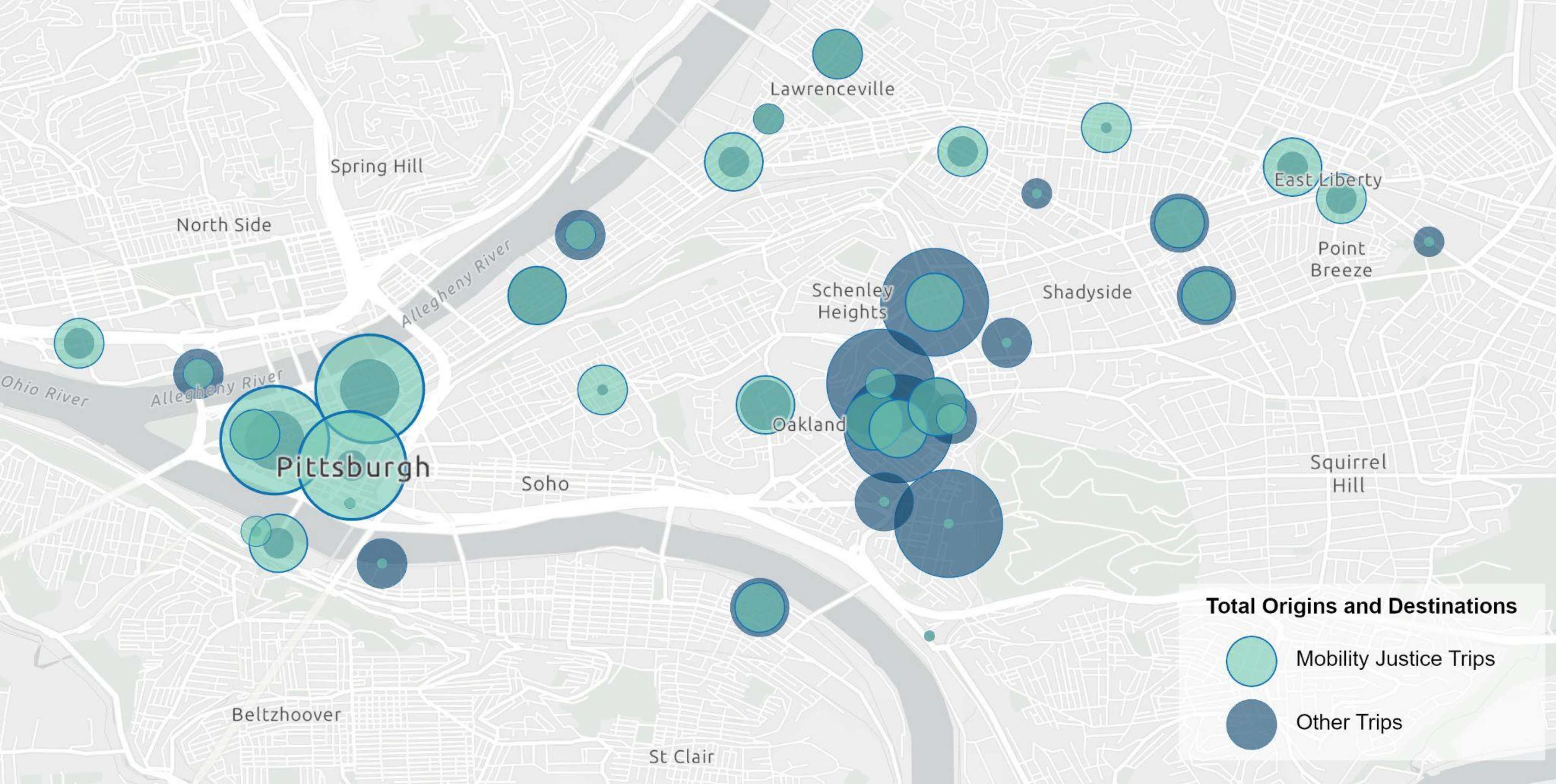
Rider Type	Rider count	Percent of riders
MJM riders	128	**1.1%**
Non-MJM riders	11,281	98.9%

Rider Type	Ride count	Percent of rides
MJM riders	3,338	**4.9%**
Non-MJM riders	64,442	95.1%

MJM riders take three times more trips than other members, on average



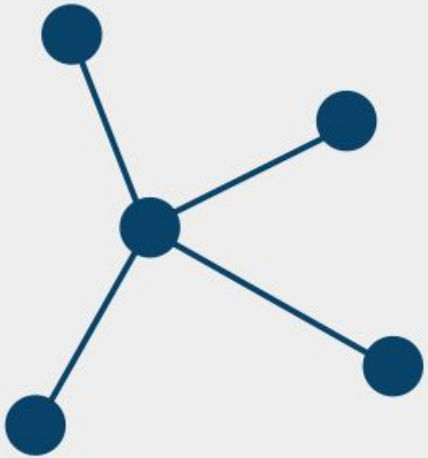
MJM Trips vs. Other Trips



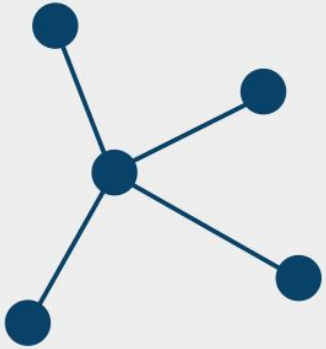
- Project Motivation and Context
- Data overview
- Our approach and results
- Solutions, Challenges, and Future Work

Trip Origins & Destinations

Network Flow

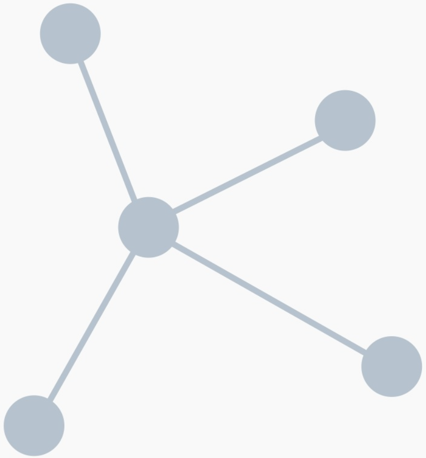


Optimizing Station Network



Trip Origins & Destinations

Network Flow

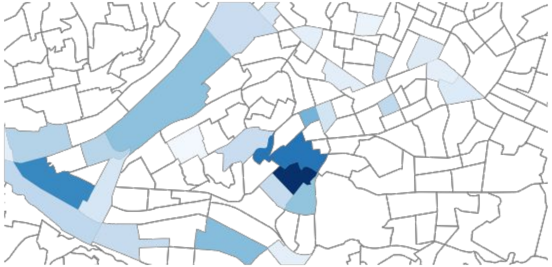


Optimizing Station Network

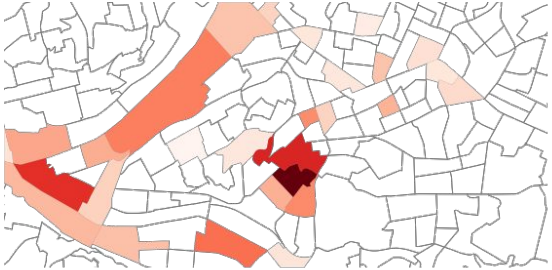
Phase I: Station Origins & Destinations

- **Task**: Estimate ridership in a block group, given its neighborhood characteristics
- **Method:** Negative Binomial Regression Model
  - Good for count data (i.e., number of starts and ends of trips in a block group)
  - Predictor Variables: Characteristics of a block group
  - Response Variables: Origins (Start Counts) and Destinations (End Counts)
- **Model outputs**: Number of trips originating and ending in each block group

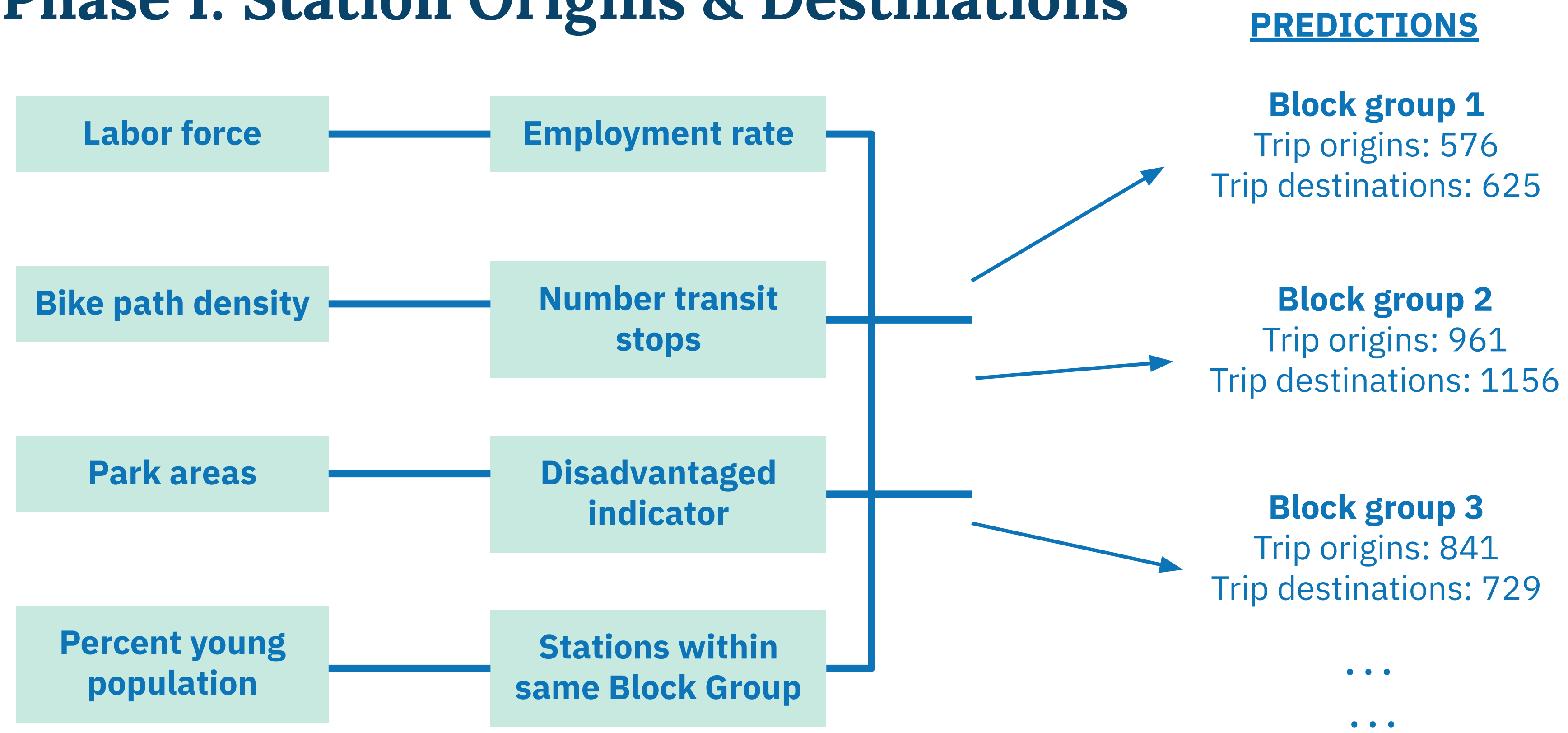
Origins (Start Counts)



Destinations (End Counts)

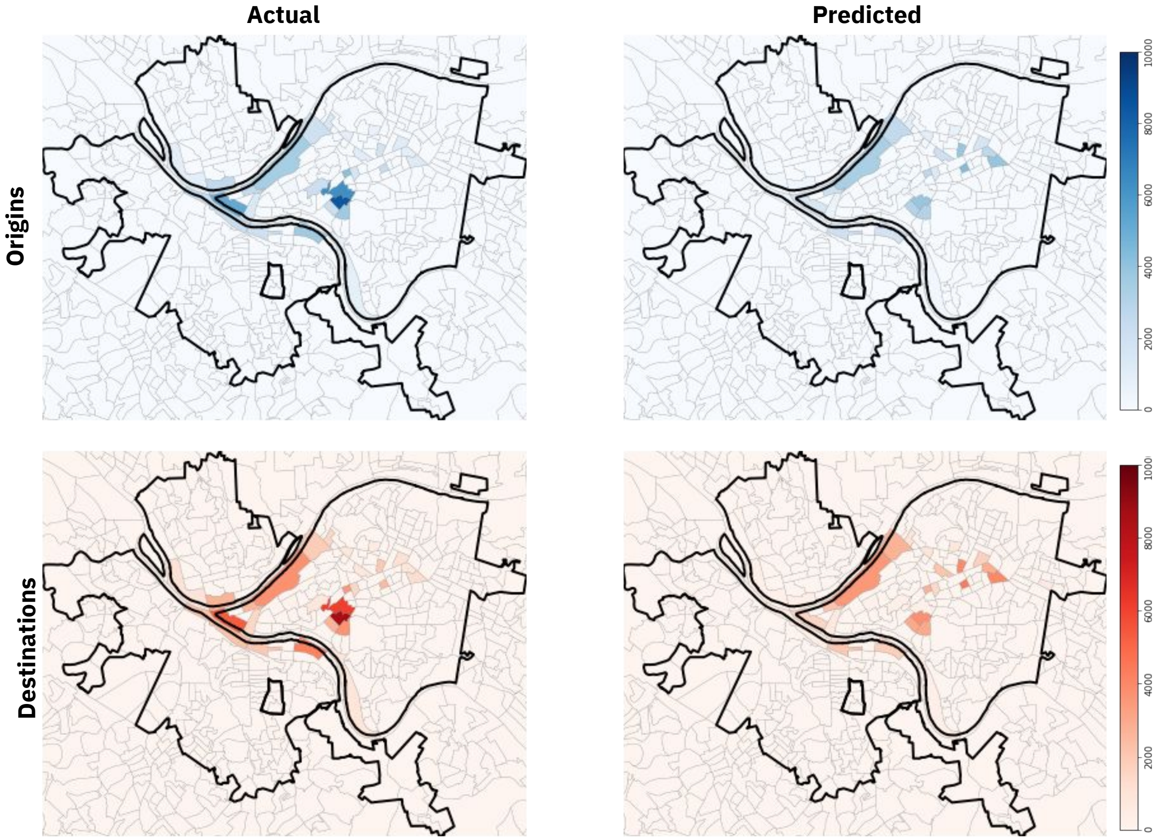


Phase I: Station Origins & Destinations



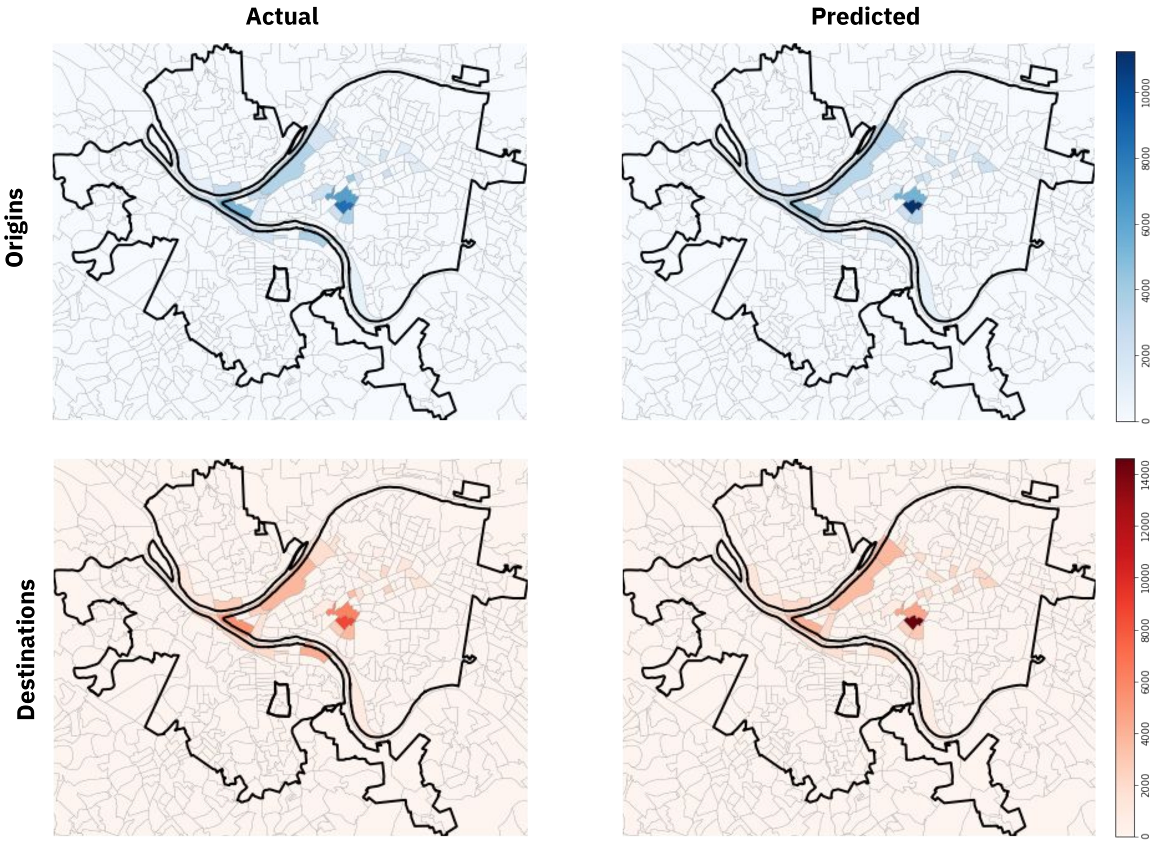
The Chicago Model Scaled

Comparing Actual and Predicted Number of ***Originations*** *and* ***Destinations***

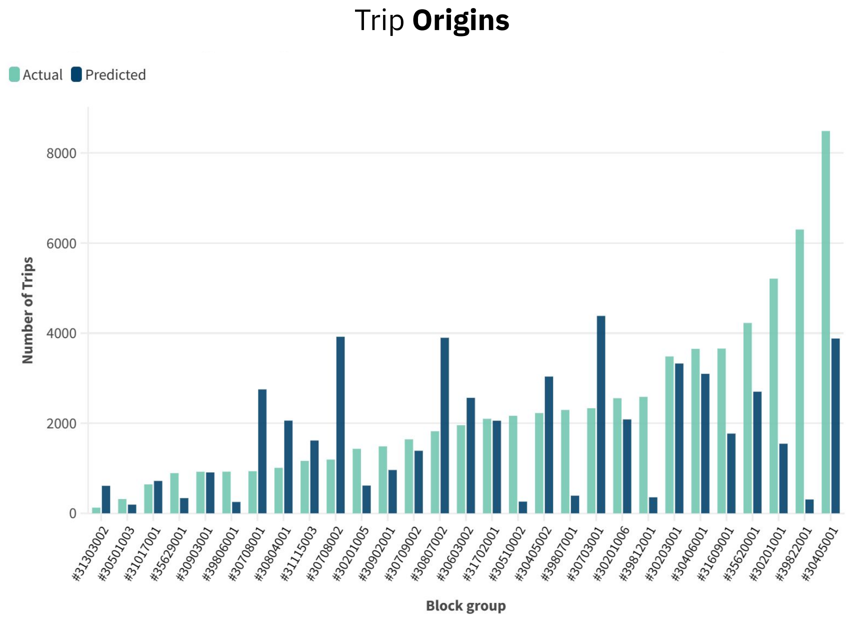


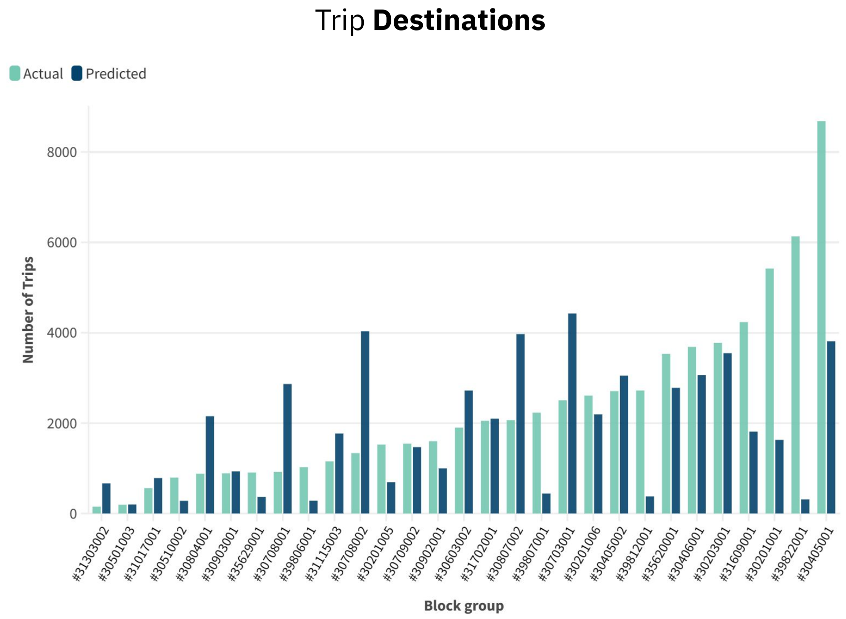
The Pittsburgh Model

*Comparing Actual and Predicted Number of **Originations** and **Destinations***



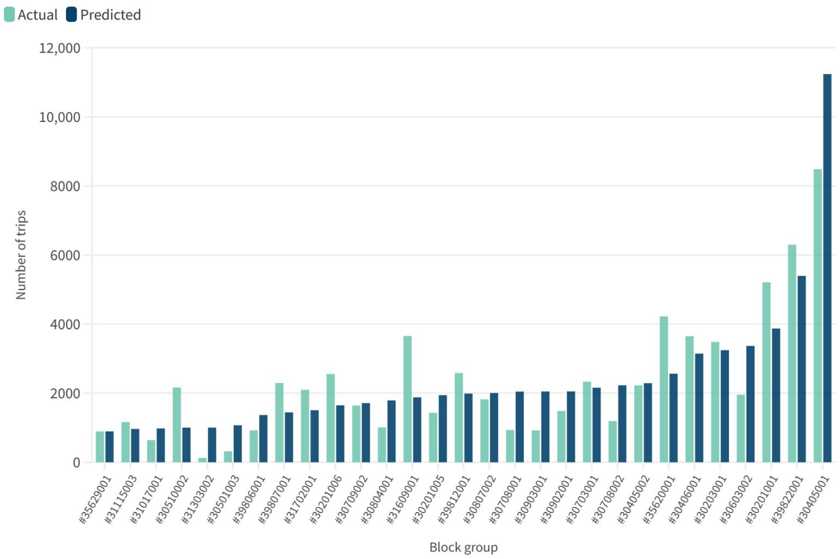
Results from Scaled Chicago Model



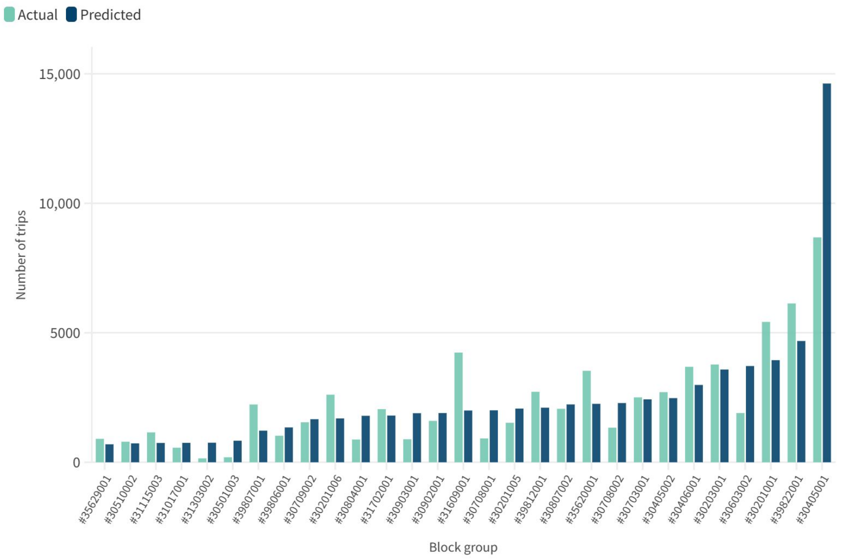


Results from Pittsburgh Model

Trip **Origins**

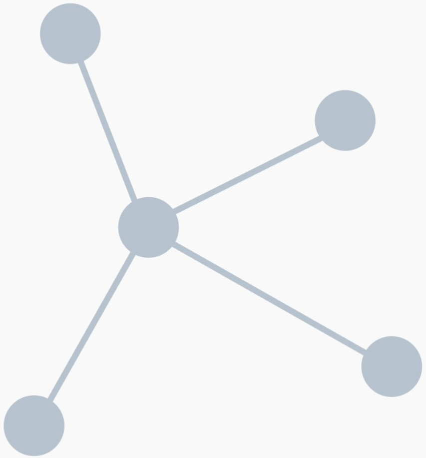


Trip **Destinations**



Trip Origins & Destinations

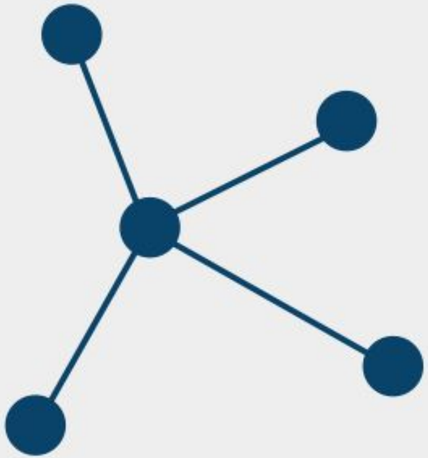
Network Flow



Optimizing Station Network

Trip Origins & Destinations

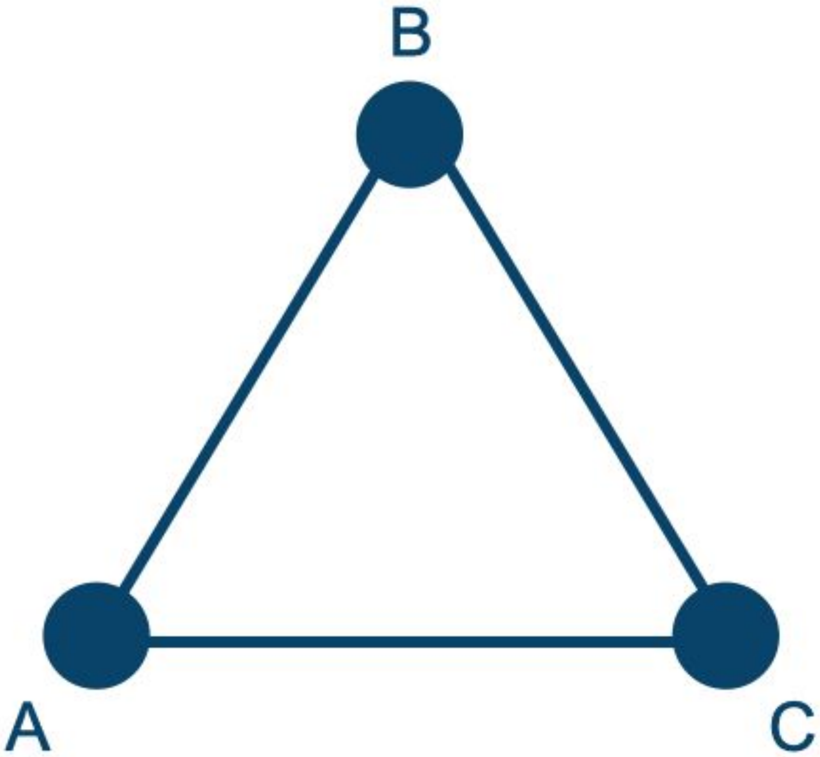
Network Flow



Optimizing Station Network

Phase II: Network Flow

- Take trip count predictions from Phase I
- Determine how those trips flow through network
    - From A to B
    - From A to C
    - From B to A
    - From B to C
    - From C to A
    - From C to B



Phase II: Network Flow

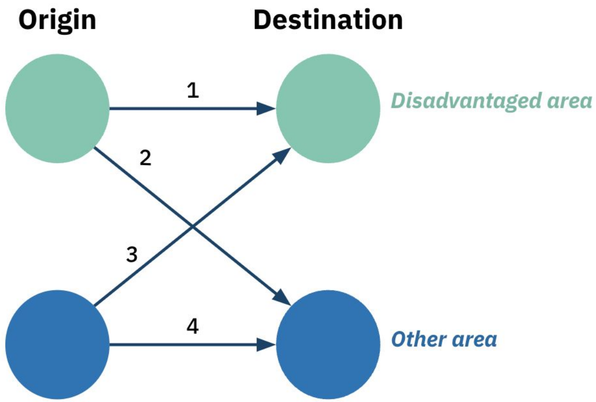
- **Task:** Understand trip distribution
- **Method:** Competing Destination Model (CDM)
- **Model inputs:**
  - Predicted trip origins and destinations from Negative Binomial model
  - Distance/time matrix between stations
  - Accessibility improvement matrix
- **Model outputs:**
  - Rho (ρ): Measure of improved accessibility
  - Beta (β): Willingness of people to travel at a distance

Historical OD Pair trip data + **Travel benefits:** Changes in opportunities − **Travel Decay:** Trip features (Distance, time)

Phase II: CDM Results

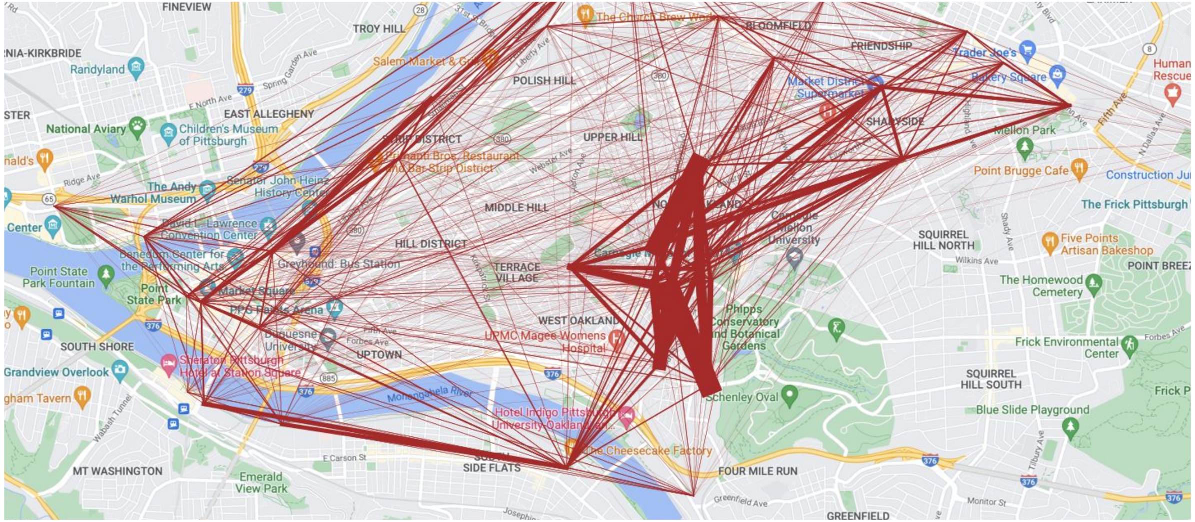
Parameters

OD Pair Type	OD Pair Description	ρ casual	β casual	ρ member	β member
1	Disadvantaged - Disadvantaged	0.02	-0.76	0.11	-0.48
2	Disadvantaged - Other	0.01	-0.64	0.09	-0.52
3	Other - Disadvantaged	0.02	-0.65	0.10	-0.49
4	Other - Other	0.02	-0.68	0.09	-0.50

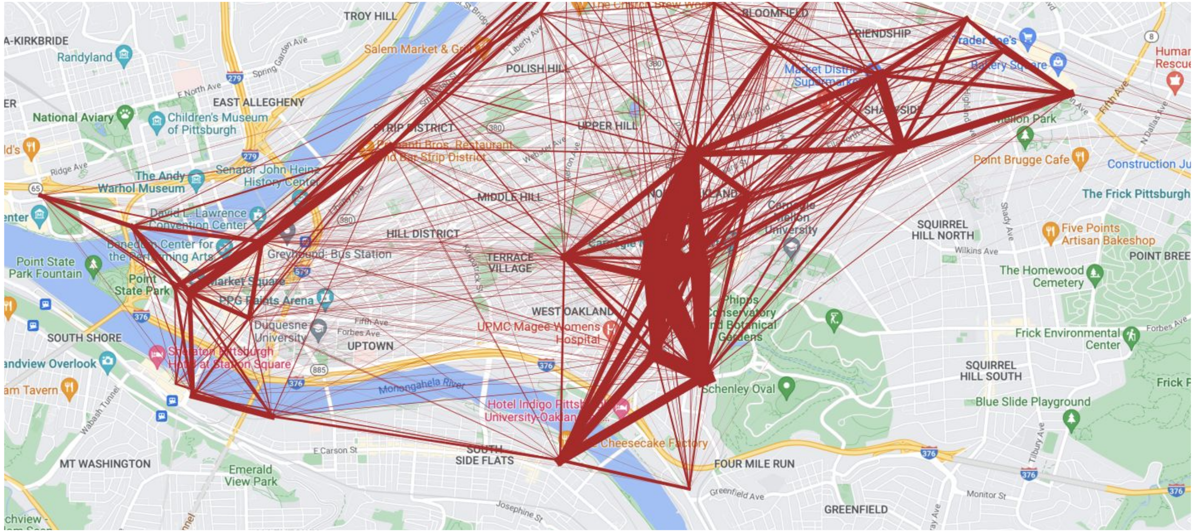


Phase II: CDM Predictions

Actual
Trip Distribution



Predicted
Trip Distribution



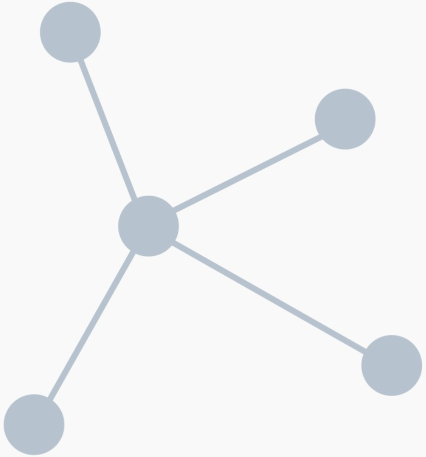
Trip Origins & Destinations

Network Flow

Optimizing Station Network

Trip Origins & Destinations

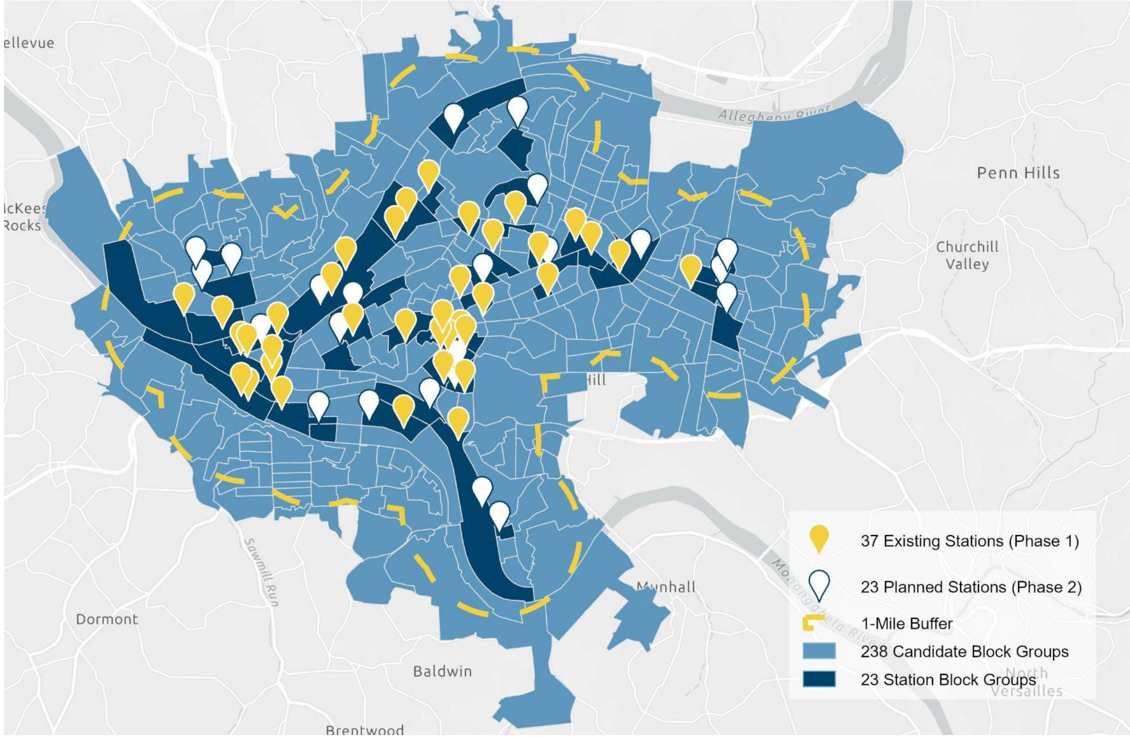
Network Flow



Optimizing Station Network

Phase III: Optimizing Station Network

- **Task:** Maximize accessibility by optimally placing new stations
- **Method:** Optimization with genetic and greedy algorithm
- **Model inputs:**
  - Output from Phase II
  - List of candidate stations
- **Model outputs:**
  - List of optimal stations



Phase III: Optimization Methods

Greedy Algorithm

- Selects the single optimal node to add to the network at every iteration using an exhaustive search
- Deterministic, no changes if re-run
- Disadvantage: Does not guarantee an optimal solution
- Advantage: Faster run time

Genetic Algorithm

- Optimizes for the entire network at once
- Stochastic, could change if re-run
- Disadvantage: Complexity, long run time increases
- Advantage: It can identify the optimal network solution and outperform greedy

Phase III: Genetic Algorithm

- **Fitness function**: Maximize accessibility

$$Maximize: \sum_{i=1}^{238}\sum_{j=1}^{238} Accessibility\ change_{i,j} * Number\ of\ trips_{ij}$$

- Uses “evolutionary” techniques
  - **Selection**: Select fittest *parents*
  - **Crossover**: Merge traits (or *genes*) of parents node to create *child*
  - **Mutation**: Insert random changes (or *mutations)* to introduce novel genetic combinations

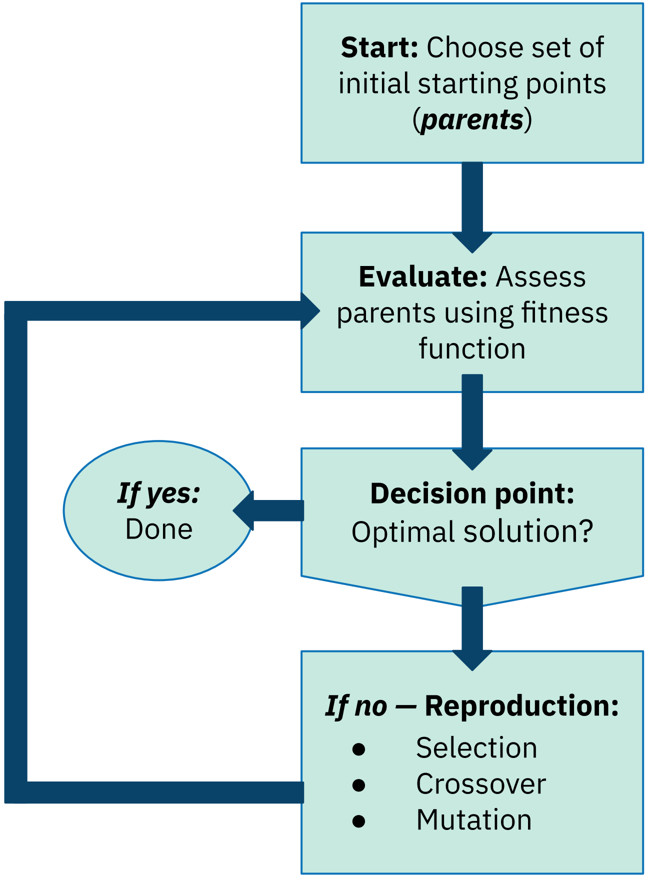


Diagram adapted from this video.

Phase III: Genetic Algorithm

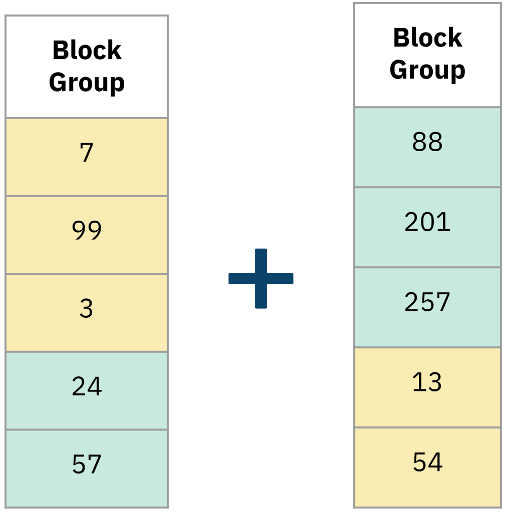
- **Genes:** A list of stations that the fitness function has selected as producing the highest accessibility gain.
- **Parent Solutions:** A list of the block groups that the fitness function has selected to host new stations. Parents are selected because they maximize accessibility.
- **Children:** The new solution values that are created by crossing over the lists of block groups from the two parents.
- **Mutations:** Randomly removing genes from the pool
- **K**: The number of stations to consider adding in a given run. E.g., if K = 5, we would select the 5 best stations to increase network accessibility.

Parent Solutions

Block Group	Block Group
7	88
99	201
3	257
24	13
57	54

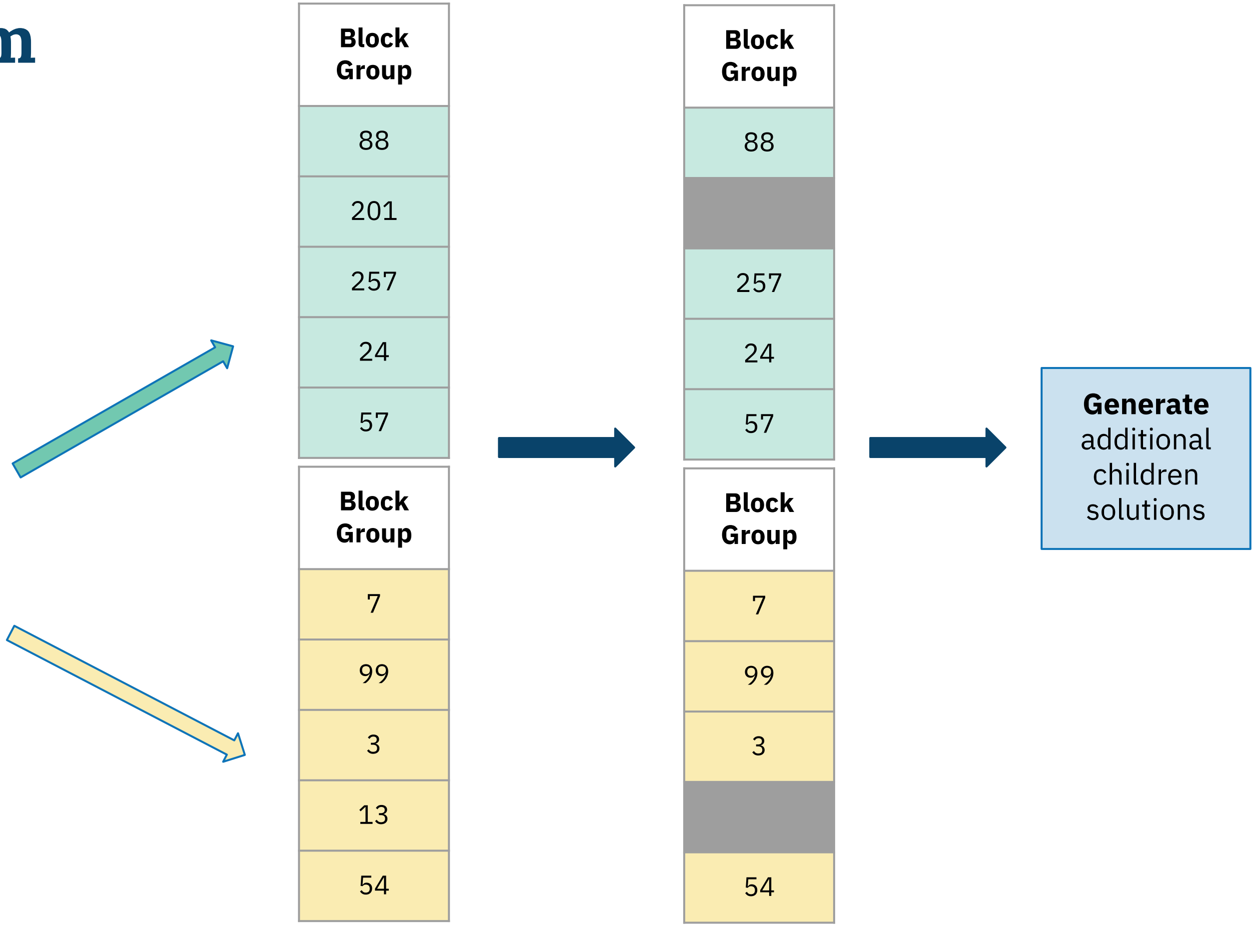
Genetic algorithm overview

Selection: These are our two best parent solution lists.

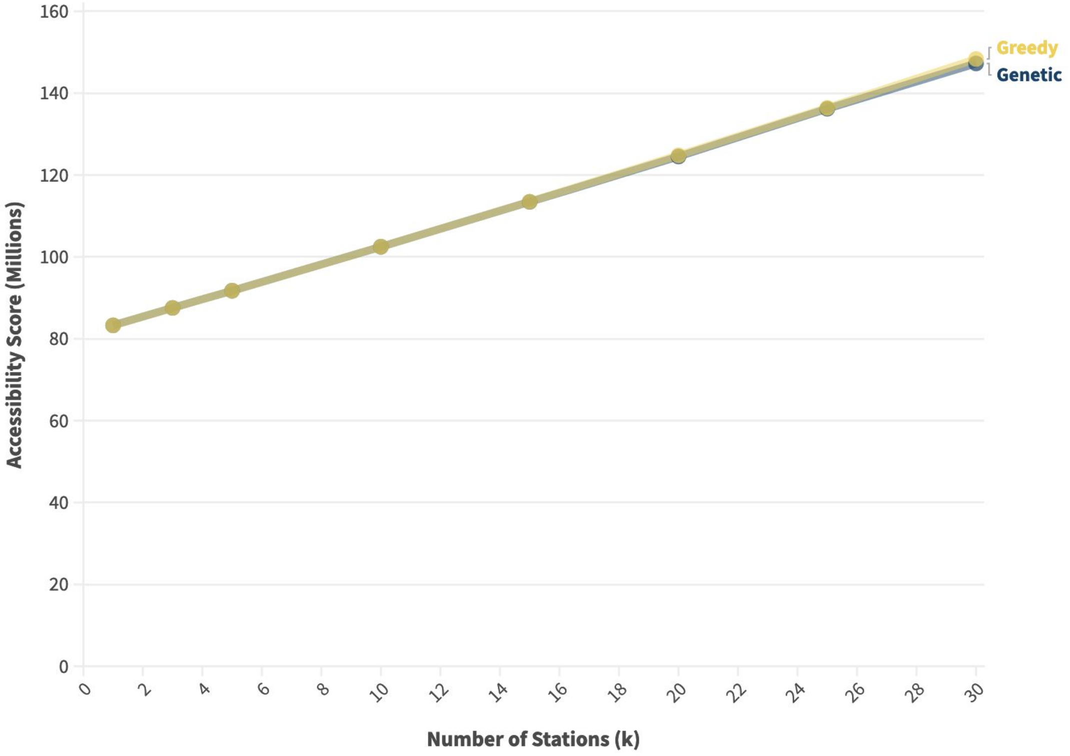


Crossover: Follow the colors.

Mutation: Follow the grey.



Greedy and genetic algorithm produce very similar results



The gain in *accessibility* — the number of grocery stores, jobs, etc. — that riders experience by moving between their origin and destination station is virtually the same between the two algorithms.

Project Motivation and Context

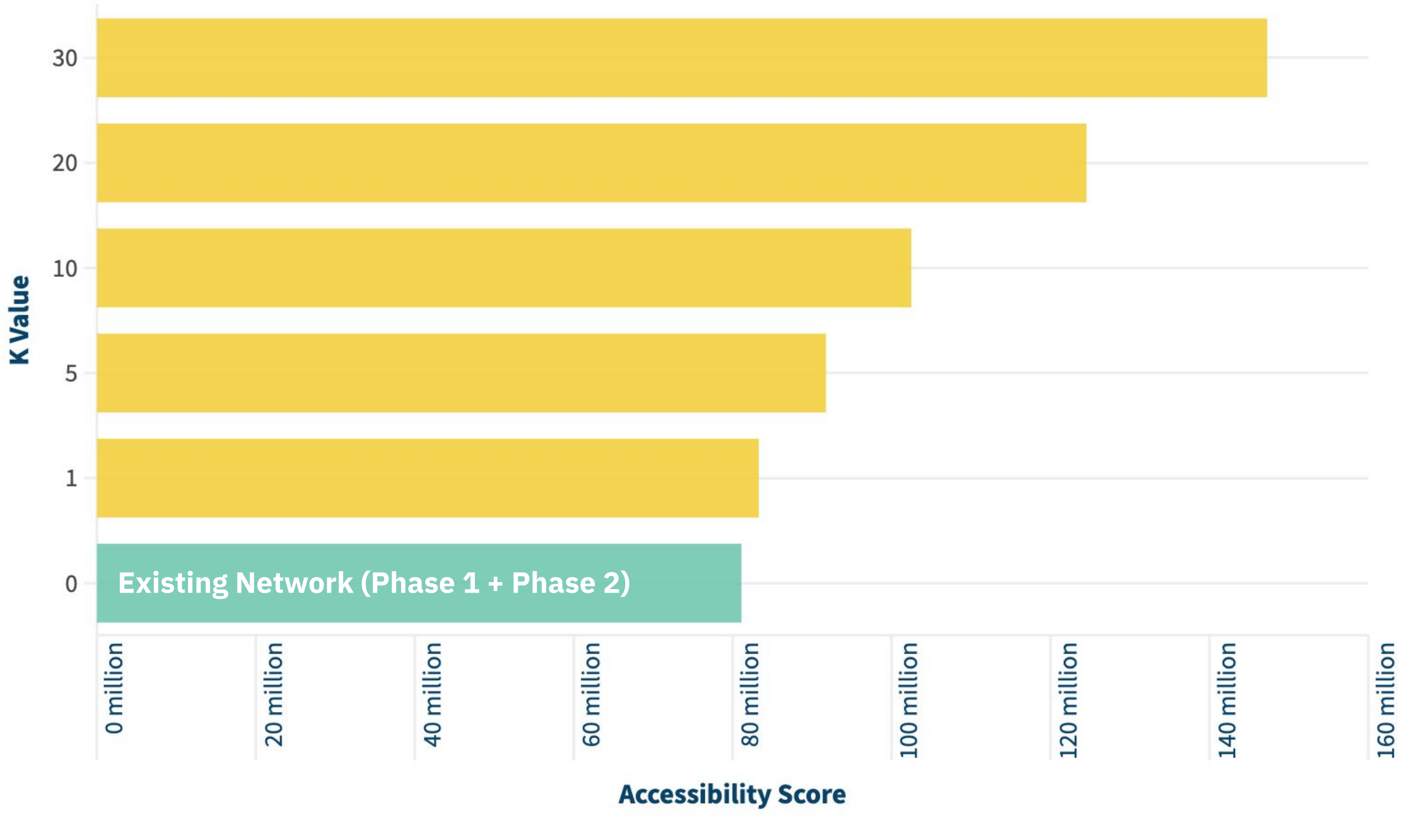
Data overview

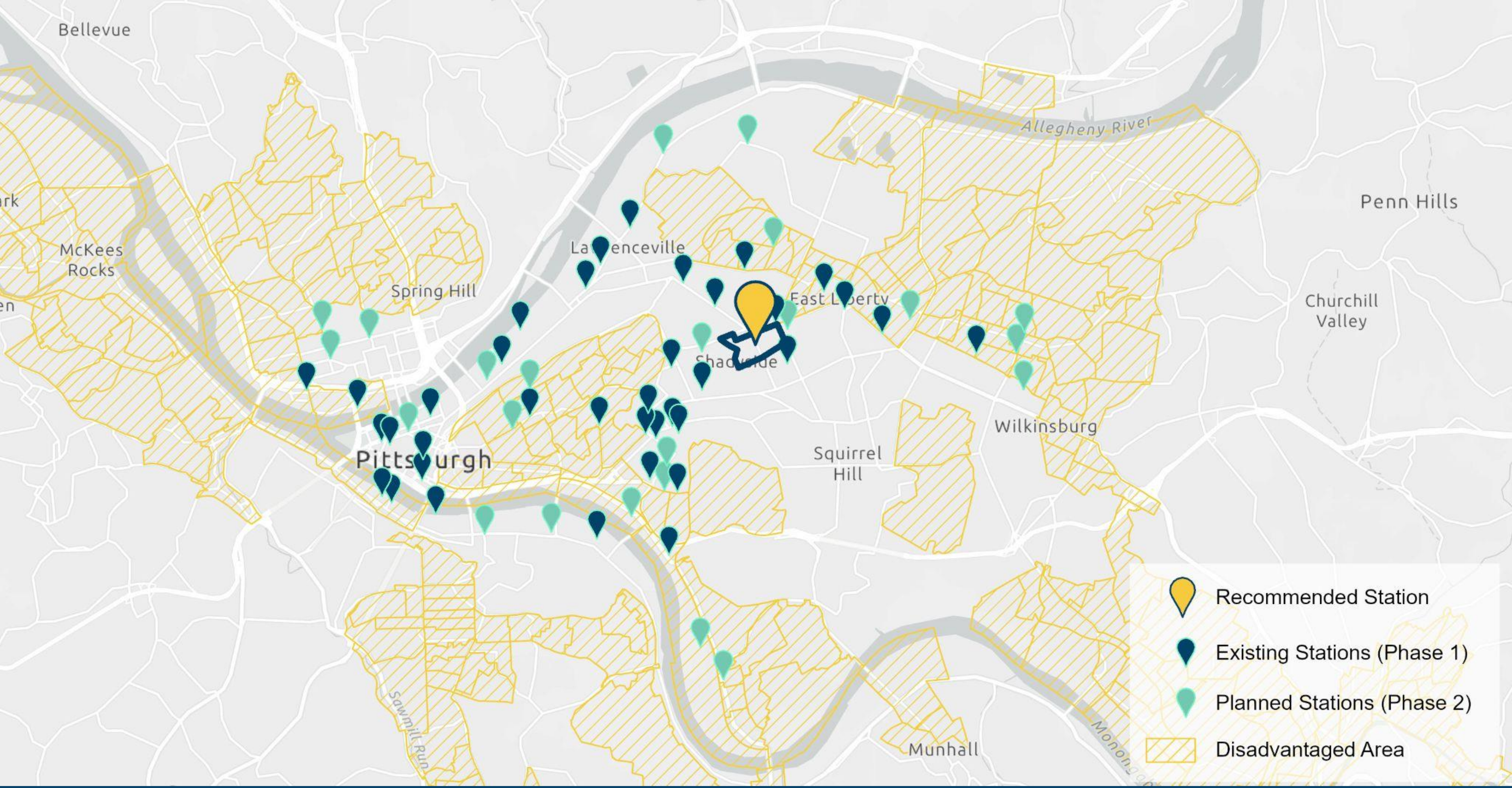
Our approach and results

Solutions, Challenges, and Future Work

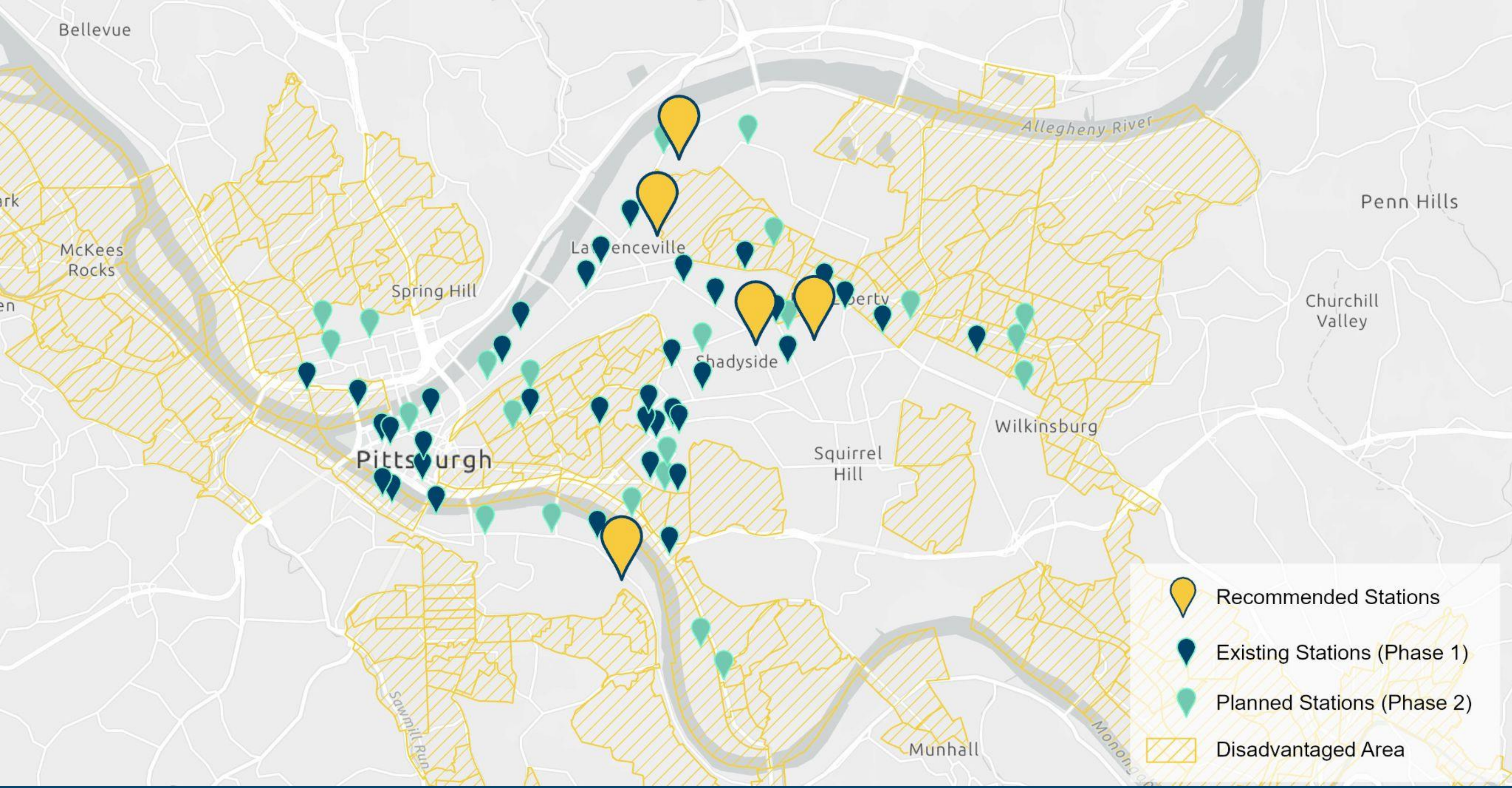
Park for Questions

Optimization maximizes accessibility for different network sizes

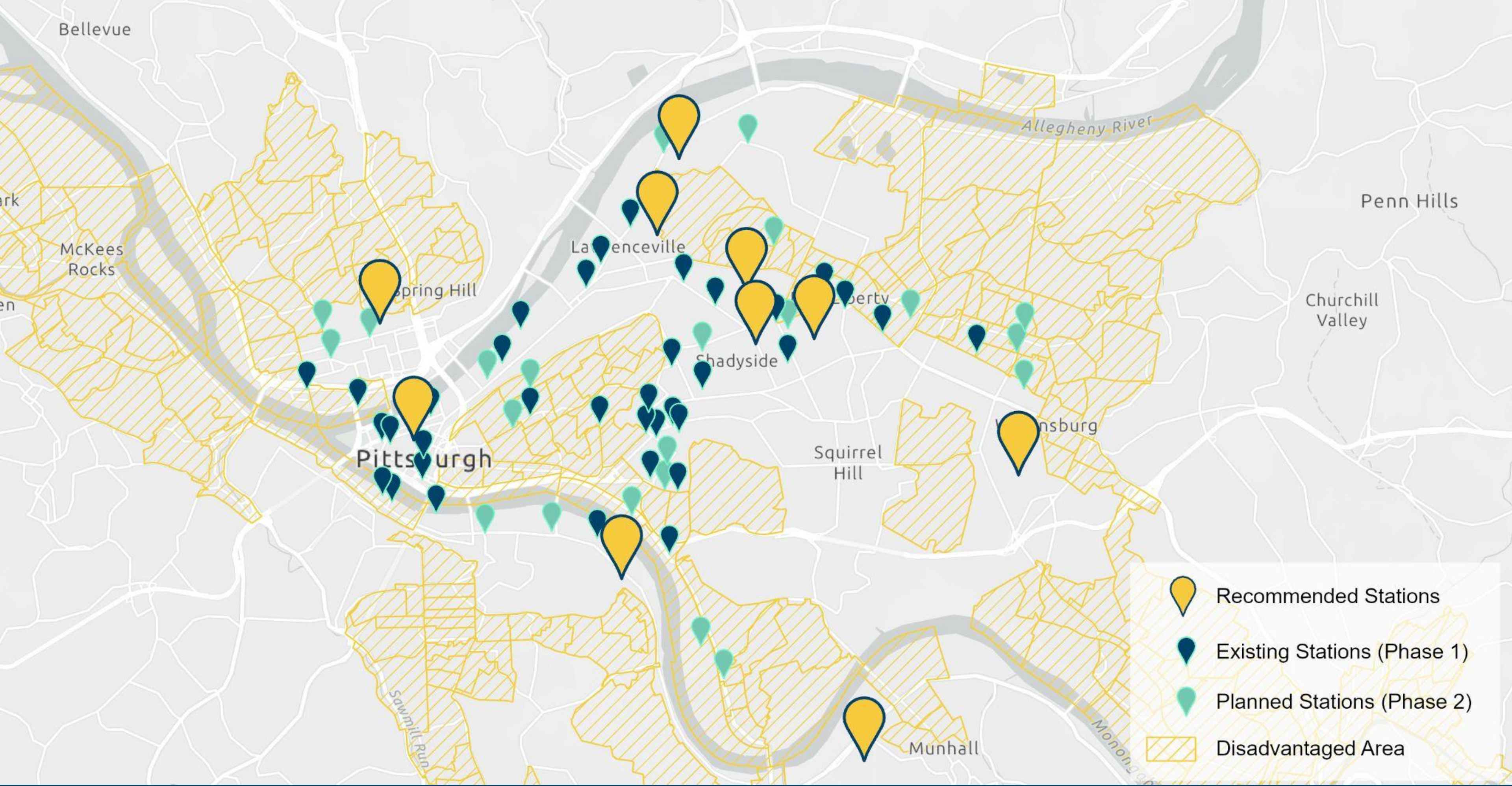




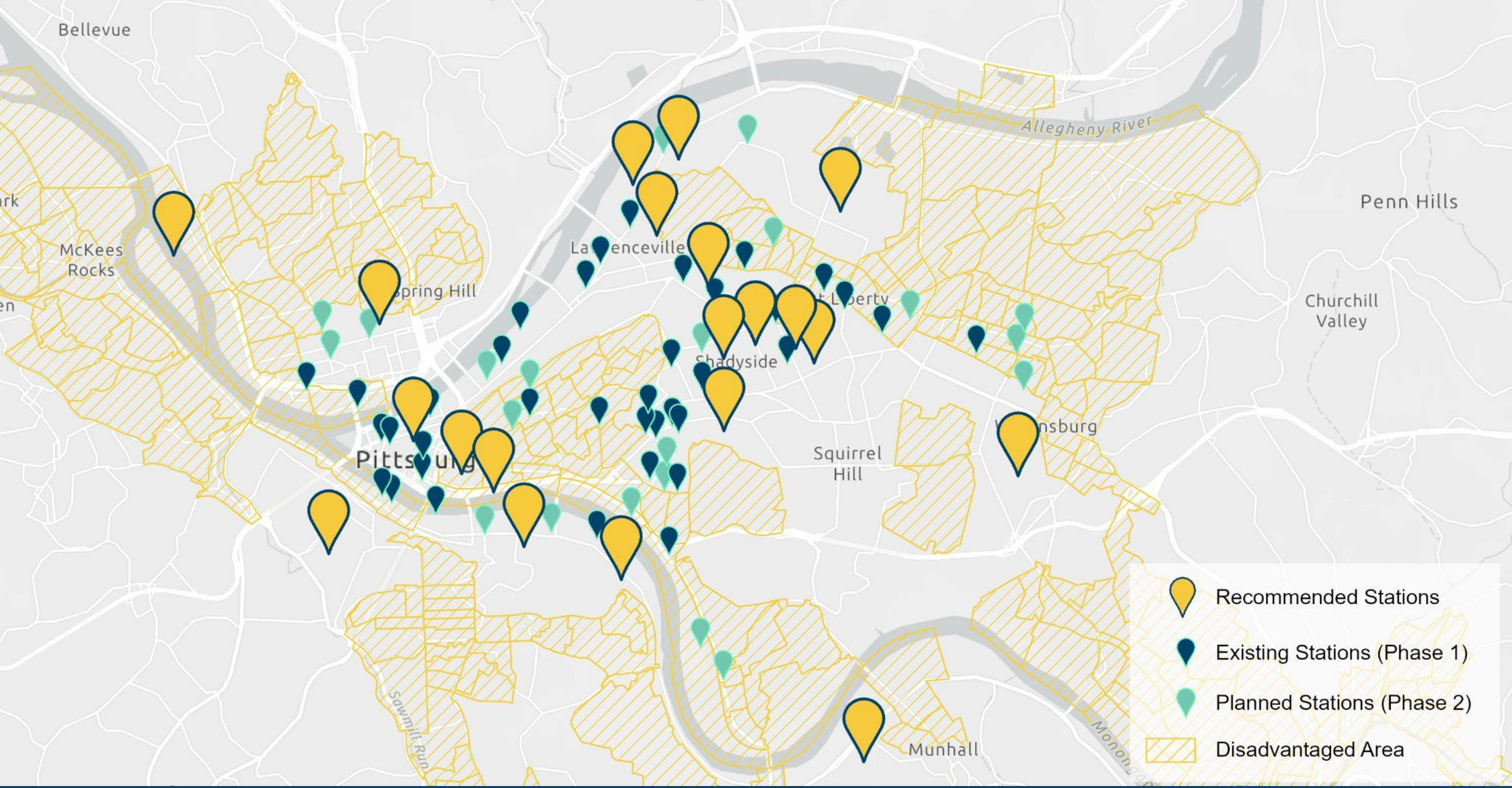
K value: **1** **Accessibility Gain: 2.68%** Method: Greedy Accessibility Score: 83.3 million



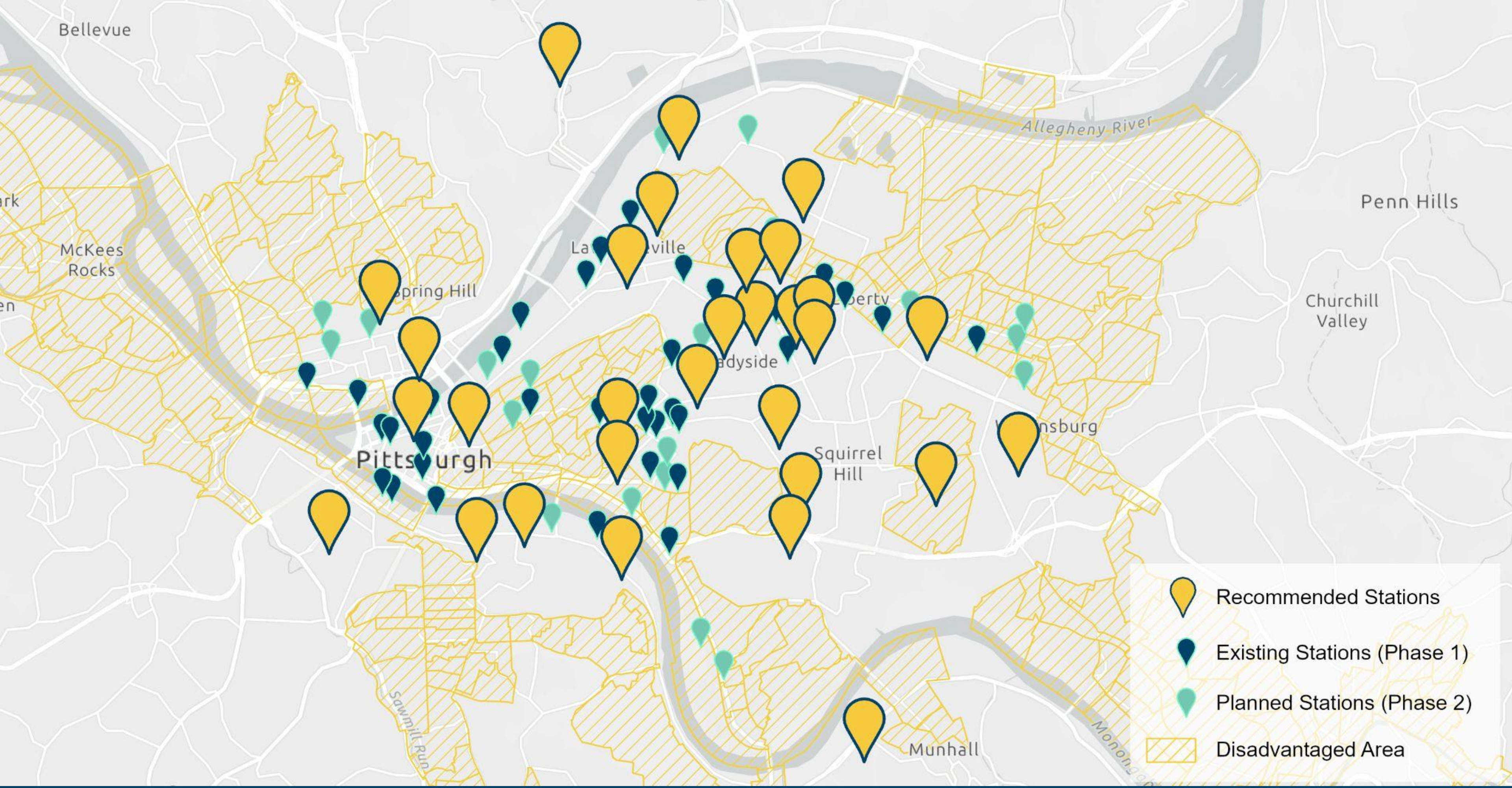
K value: **5** **Accessibility Gain: 13.12%** Method: Genetic Accessibility Score: 91.7 million



K value: **10** **Accessibility Gain: 26.36%** Method: Genetic Accessibility Score: 102.5 million

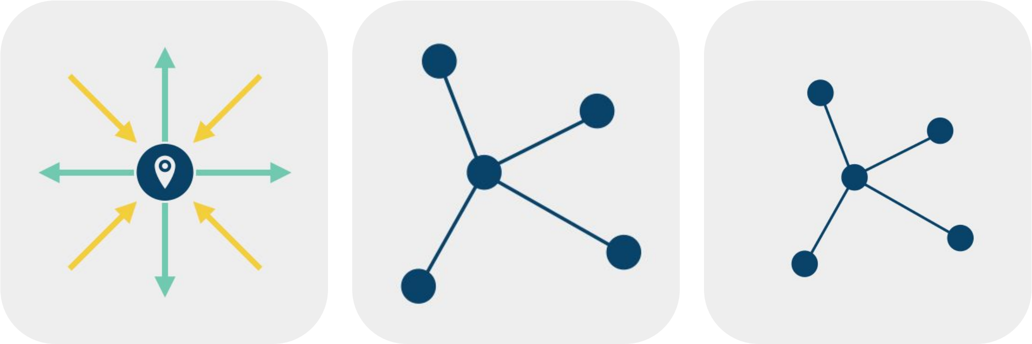


K value: **20** **Accessibility Gain: 53.53%** Method: Genetic Accessibility Score: 124.5 million



K value: **30** **Accessibility Gain: 81.57%** Method: Genetic Accessibility Score: 147.3 million

Add this data-driven model to the POGOH toolkit!



Inputs

- Past POGOH trips
- Demographic Data
- Block Group Characteristics
- Distance or time between stations
- Opportunities near stations
- List of candidate stations

Outputs

- List of optimal station locations at the block group level
- Accessibility score for the recommended network
- Predicted trips across the recommended network

Limitations and Challenges

Amount of data impacts predictive power

- Less than a year’s worth of data
- Few stations in current network

Limitations of demographic assumptions

- We assume that the rider has the “average” demographic features of the Block Group from which their trip originated

Additional Recommendations

Allegheny County data merge
Enables more granular analysis

Focus on MJM riders
Outreach, learning session with MJM riders who ride frequently

Data as an entrypoint
Use data-driven recommendations as a starting point for conversations with community members

Future Work

More Data / More Precise Data

- More data over time and from Phase 2 of the POGOH network expansion
- Allegheny County Data Merge
- Identify University of Pittsburgh riders
- More geographic granularity

Adjust Methodology

- Improve or customize methodology for identifying candidate stations
- Revenue-based optimization

Thank you!

SPECIAL THANKS TO

Sara Khalil

David White

Bike Share Pittsburgh

Dr. Xiaodong Qian

Dr. Lee Branstetter

Cameron Drayton

Questions & Discussion

References

Bike PGH. (2022). *Bicycling and Walking in Pittsburgh 2022: A snapshot of trends and statistics.* BikePGH.org. https://bikepgh.org/wp-content/uploads/2022/02/Cycling-Trends-Snapshot-in-Pittsburgh-2022.pdf

City of Pittsburgh (2018). Pittsburgh Equity Indicators. PittsburghPA.gov. https://pittsburghpa.gov/equityindicators/documents/PGH_Equity_Indicators_2018.pdf

Fan, Zhufeng, and Corey Harper (2022). "Congestion and environmental impacts of short car trip replacement with micromobility modes." Transportation Research Part D: Transport and Environment, Vol., 103. https://doi.org/10.1016/j.trd.2022.103173

Qian, Xiaodong, Miguel Jaller, and Debbie Niemeier. “Enhancing Equitable Service Level: Which Can Address Better, Dockless or Dock-Based Bikeshare Systems?” *Journal of Transport Geography* 86 (June 1, 2020): 102784. https://doi.org/10.1016/j.jtrangeo.2020.102784.

Qian, Xiaodong, and Miguel Jaller. “Bikeshare Destination Choices and Accessibility among Disadvantaged Communities.” Transportation Research Part D: Transport and Environment 91 (February 1, 2021): 102686. https://doi.org/10.1016/j.trd.2020.102686.

Qian, Xiaodong, and Deb Niemeier. “High Impact Prioritization of Bikeshare Program Investment to Improve Disadvantaged Communities’ Access to Jobs and Essential Services.” Journal of Transport Geography 76 (April 1, 2019): 52–70. https://doi.org/10.1016/j.jtrangeo.2019.02.008.

Pittsburgh Region (n.d.) *Public Transit in Pittsburgh*. PittsburghRegion.org. https://pittsburghregion.org/living-in-pittsburgh/transportation/

POGOH. (n.d.) *Media Kit.* POGOH.com. https://pogoh.com/media-kit/

APPENDIX

Phase I: Results

Origin Coefficients

	Coefficient	P-Value
Intercept	7.107	0.000
Total_In_labor_force	0.001	0.548
Employment_Rate	-0.008	0.385
bike_path_density	2.623	0.804
park_density	-0.015	0.932
num_stations	0.773	0.153
pct_young_pop	0.373	0.776
num_bus_stops	-0.015	0.675
Disadvantaged	-0.347	0.614

Destination Coefficients

	Coefficient	P-Value
Intercept	7.239	0.000
Total_In_labor_force	0.000	0.760
Employment_Rate	-0.011	0.246
bike_path_density	0.677	0.949
park_density	-0.010	0.955
num_stations	0.889	0.100
pct_young_pop	0.763	0.562
num_bus_stops	-0.015	0.685
Disadvantaged	-0.617	0.370

Phase I: Results

	Root Mean Squared Error
Scaled Chicago Model (Origin)	1,972
Scaled Chicago Model (Destination)	1,960
Pittsburgh Model (Origin)	992
Pittsburgh Model (Destination)	1,413

Explanation

Phase III: Tuning the Optimization

There are several choices to be made when running the optimization:

- How many recommended stations are desired? (defaults to k = 10)
- Should the solution be produced using the genetic algorithm? Or a greedy algorithm?
- For the genetic algorithm:
    - How many generations should be produced? (default = 100)
    - How many parents should be included in a generation? (default = 4)
    - How many offspring should be produced in a generation? (default = 100)

    Higher values increase likelihood of achieving optimal solution, but also increase run time

Phase III: Our Optimization Parameters

Identify recommendations for multiple different k values:

- When k = 1 we use the greedy algorithm to identify the definite optimal station using an exhaustive search
- When k is greater than 1, we use the genetic algorithm to identify the optimal network.
    - # generations = 100
    - # parents = 4
    - # offspring = 100