Mobility21 Final Research Report

Peter Zhang

July 2023

Funding Agency Mobility21, A USDOT National University Transportation Center.

Project Timeline July 1, 2022 - June 30, 2023.

Principal Investigator Peter Zhang, Carnegie Mellon University.

ORCID 0000-0002-0422-834X.

Student Contributors Yidi Miao, Hao Hao, Alberto Japon Saez.

Award Number #400 Controlled Deployment of Analytical Solutions for Essential Transportation Services in Low-Income Neighborhoods.

Contents

1 Project Description

In the landscape of Pittsburgh, Pennsylvania's transportation sector, Heritage Community Transportation (HCT) has been a critical service provider in low-income neighborhoods on the eastern side of the city. HCT's services play a crucial role in these communities, linking individuals with employment, healthcare, and other essential services. However, like many public transit organizations, HCT has been significantly impacted by the COVID-19 pandemic, facing a steep decline in ridership and uncertainty in funding.

The equitable provision of transportation is a critical aspect of a city's infrastructure. It underpins access to basic services such as healthcare, education, and employment opportunities. In east Pittsburgh's low-income neighborhoods, HCT has been a necessary element in ensuring this access. The transit services it provides are more than just a means of getting from point A to B; they are a vital part of the social and economic framework of the community.

Yet, the organization has been dealt a severe blow due to the effects of the COVID-19 pandemic. HCT's ridership had dropped by a 50%, a clear indicator of the pandemic's impact on public transportation usage. Compounded by uncertainties in public funding, HCT now finds itself in a precarious position.

In response to these challenges, HCT engaged in a project with our team, composed of researchers, faculty, and students. The project's aim was to recommend a service change to one of HCT's routes, attempting to reinvigorate its operations and restore its capacity to serve the community. This collaboration also presented a valuable opportunity for team members to contribute to and learn from a real-world, complex transportation problem.

While this project represents an essential step towards addressing HCT's challenges, the broader issue of ensuring equitable transportation in Allegheny is far from being completely resolved. It's a complex, city-wide issue requiring a comprehensive, collaborative approach involving city officials, transportation professionals, and the community at large.

In summary, the project serves as a practical initiative to help HCT navigate its current challenges and continue to provide essential services to east Pittsburgh communities. It is an endeavor to learn, innovate, and drive positive change in the realm of equitable transportation. Meanwhile, the county faces the ongoing task of prioritizing and improving transportation equity across all its neighborhoods. The challenges are considerable, but so is the necessity of the task at hand. For HCT, the project collaborators, and the bigger region, the work to ensure equitable transportation continues.

2 Preliminaries: On-Demand and Fixed-Route Service Analytics

2.1 On-Demand Transportation Service

On-demand transportation services, characterized by dynamic, real-time dispatched trips, have gained popularity due to their flexible and personalized nature. This system's operating principle is generally demand-responsive; hence, their routes and schedules are not fixed but determined by user needs.

The key advantage of on-demand transportation lies in its optimization of wait and travel times. Since users can request rides at their convenience, the service reduces wait times and, by delivering users to their destinations directly, minimizes in-vehicle time. This principle can be represented as follows:

$$W_{\text{ondemand}} = W_{\text{waiting}} + W_{\text{in-vehicle}} \tag{1}$$

Where:

- W_{ondemand} is the total wait time in the on-demand service,
- W_{waiting} is the wait time for the vehicle to arrive, and
- $W_{\text{in-vehicle}}$ is the time spent in the vehicle.

However, the on-demand model faces challenges. The cost per trip is typically higher due to the lower load factor (i.e., the ratio of utilized vehicle capacity to available vehicle capacity), which can be expressed as:

$$LF = \frac{P_{\text{onboard}}}{P_{\text{capacity}}} \tag{2}$$

Where:

- LF is the load factor,
- P_{onboard} is the number of passengers onboard, and
- P_{capacity} is the total vehicle capacity.

2.2 Fixed-Route Service

Fixed-route services form the backbone of traditional public transportation, operating on a predetermined path and schedule. The regularity of these services simplifies planning for both operators and users. However, fixed-route services are more prone to inefficiencies, with higher wait times and in-vehicle times. The total wait time for fixed-route services can be expressed as:

$$W_{\text{fixed}} = W_{\text{waiting}} + W_{\text{transfer}} + W_{\text{in-vehicle}} \tag{3}$$

Here, W_{transfer} is the wait time at transfer points between different modes of transit.

2.3 Implications for Equitable Transportation

Equitable transportation aims to ensure all citizens have access to essential services and opportunities. This goal introduces a complex optimization problem: balancing coverage, cost, and convenience, often measured in terms of wait time and load factor. This challenge is further complicated by spatial and temporal variability in demand, differing user needs, and budget constraints.

While on-demand services may provide a solution for underserved areas or populations with specific needs (e.g., non-traditional working hours), the higher cost per trip can limit their usage by low-income individuals. Fixed-route services offer a cost-effective solution for mass transit, but their fixed schedules and routes might leave gaps in coverage and offer less flexibility. Hence, integrating these services can potentially enhance equitable transportation:

$$C_{\text{total}} = C_{\text{fixed}} + C_{\text{ondemand}} \tag{4}$$
$$W_{\text{total}} = W_{\text{fixed}} + W_{\text{ondemand}} \tag{5}$$

Where:

- C_{total} is the total cost,
- C_{fixed} is the cost for the fixed-route services,
- C_{ondemand} is the cost for the on-demand services,
- W_{total} is the total wait time, and similarly for W_{fixed} and W_{ondemand}.

Transportation planners could therefore leverage mathematical models and algorithms to design and manage integrated transportation networks, minimizing C_{total} and W_{total} while maximizing coverage and accessibility.

2.4 A Numerical Example

Consider that we still have four passengers, each at distinct origin points $O_1(0,0)$, $O_2(2,2)$, $O_3(5,1)$, and $O_4(7,3)$, and they all wish to travel to a common destination $D(10,2)$.

In the on-demand service scenario, assuming that our vehicle has enough capacity to pick up all passengers, we would want to find the shortest possible route that visits each passenger once before heading to the destination. This is a version of the Traveling Salesperson Problem (TSP).

For simplicity, we use the Euclidean distance to calculate the distances between points:

$$d(O_i, O_j) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

We first calculate the distances between all the origin points and the destination point:

$$d(O_1, O_2) = \sqrt{(2-0)^2 + (2-0)^2} = 2\sqrt{2}$$
$$d(O_1, O_3) = \sqrt{(5-0)^2 + (1-0)^2} = \sqrt{26}$$
$$d(O_1, O_4) = \sqrt{(7-0)^2 + (3-0)^2} = \sqrt{58}$$

And similarly, we calculate the remaining distances.

Next, we can solve this instance of the TSP and determine the optimal pick-up sequence. Assuming this sequence is $O_1 \to O_2 \to O_3 \to O_4 \to D$, the total travel distance D_{ondemand} would be:

$$D_{\text{ondemand}} = d(O_1, O_2) + d(O_2, O_3) + d(O_3, O_4) + d(O_4, D)$$

Assuming the vehicle travels at a speed of $v = 1$ unit per minute, the total travel time W_{ondemand} for the on-demand service would be:

$$W_{\text{ondemand}} = D_{\text{ondemand}}$$

Furthermore, if we assume the operational cost of the vehicle is proportional to the total distance traveled, the cost C_{ondemand} would be:

$$C_{\text{ondemand}} = k \cdot D_{\text{ondemand}}$$

Where k is the cost per unit distance.

This demonstrates how an on-demand service can provide a faster, more efficient service by optimizing the travel route *if* demand is known ahead of time and there is sufficient capacity. In addition, the potential higher cost due to lower load factors is still an issue, emphasizing the need to integrate different transportation models to achieve an optimal solution.

Now we can also incorporate the "social cost" such as passenger travel times into the overall decision problem. Recall that we assumed the optimal pick-up sequence (found through solving the Traveling Salesman Problem) to be $O_1 \to O_2 \to O_3 \to O_4 \to D$.

We already calculated the total travel distance D_{ondemand}:

$$D_{\text{ondemand}} = d(O_1, O_2) + d(O_2, O_3) + d(O_3, O_4) + d(O_4, D)$$

In order to compute the individual travel times of each passenger, we need to determine the distance each passenger travels along this route. Here, the travel time of each passenger is proportional to the cumulative distance from their origin to the destination along the route.

For passenger 1, the travel time is simply the total travel distance D_{ondemand}. For the remaining passengers, we subtract the appropriate distances from D_{ondemand} to get their travel times.

$$\begin{aligned}
W_{\text{ondemand},1} &= D_{\text{ondemand}} \\
W_{\text{ondemand},2} &= D_{\text{ondemand}} - d(O_1, O_2) \\
W_{\text{ondemand},3} &= D_{\text{ondemand}} - d(O_1, O_2) - d(O_2, O_3) \\
W_{\text{ondemand},4} &= D_{\text{ondemand}} - d(O_1, O_2) - d(O_2, O_3) - d(O_3, O_4)
\end{aligned}$$

We can then sum these individual travel times to get the total passenger travel time T_{ondemand} under the on-demand service scenario:

$$T_{\text{ondemand}} = \sum_{i=1}^{4} W_{\text{ondemand},i}$$

This measure T_{ondemand} allows us to assess the efficiency of the on-demand service from the passengers' perspective, capturing not just the total travel distance or time, but the cumulative time passengers spend traveling. This perspective emphasizes the experience of the users, highlighting the importance of individual passenger experience in designing effective and efficient transportation systems.

2.5 TSP Under Uncertain Demand

Now we can discuss how to deal with a situation where demand isn't precisely known. This is common in the real world, where transportation systems need to handle uncertain demand and make decisions based on approximations or predictions.

One way to handle uncertain demand is to use statistical or probabilistic models to predict demand and then design the system to be robust to these uncertainties. In the context of transportation, this could involve using historical data, machine learning algorithms, or other forecasting methods to estimate future demand.

In the specific context of the Traveling Salesman Problem (TSP) and on-demand transportation services, we can consider the Beardwood-Halton-Hammersley (BHH) theorem. This theorem gives us an asymptotic approximation of the minimal length of a traveling salesman tour when the cities are randomly distributed in the plane. In other words, it provides an expected value for the optimal total travel distance when picking up n passengers randomly located in a specific area.

The BHH theorem states that if n points are independently and uniformly distributed in a unit square, then the length L_n of the shortest possible tour through these points satisfies:

$$\lim_{n\to\infty} \frac{L_n}{\sqrt{n}} = \beta$$

Here, β is a constant approximately equal to 0.7120 for a unit square.

Suppose we have a large number of passengers (say 50), and we don't know their exact locations, but we know they are uniformly distributed in a certain area of Pittsburgh. According to the BHH theorem, we can approximate the total travel distance required for an on-demand transportation service to pick up all passengers as follows:

$$D_{\text{ondemand}} \approx \beta \cdot \sqrt{n} = 0.7120 \cdot \sqrt{50} \approx 22.5 \text{ units}$$

This approximation helps us understand the expected travel distance in uncertain demand scenarios, allowing transportation planners to make more informed decisions about resource allocation and route planning.

This form of BHH theorem assumes that the demand points (or passengers, in our case) are uniformly distributed within the given area. When the spatial distribution of the demand is non-uniform, we can no longer use the simple form of the Beardwood-Halton-Hammersley theorem. In such a scenario, we have to generalize the theorem by taking into account the spatial density of the demand.

In the most general form, the BHH theorem states that if we have a sequence of demand points n, independently distributed according to a spatial density function $\rho(x)$, then the expected length L_n of the shortest possible tour through these points is asymptotically given by:

$$E[L_n] \sim \beta \int_A \sqrt{n\rho(x)\,dx}$$

Here, A is the area where the demand points are located, and β is a constant (which is dependent on the specific shape and constraints of the problem at hand). The term $\int_A \rho(x)\,dx$ is an integral over the area A, giving us a measure of the total demand in that area, weighted by the square of the local demand density.

This expression signifies that the total expected travel distance (or the length of the tour) is not merely dependent on the total number of demand points, but also on the spatial distribution of these points.

To give an intuitive explanation: if the demand points are densely packed in certain areas, it makes sense that the expected travel distance increases, as the vehicle needs to make more stops in these high-density areas, thereby increasing the total distance covered.

This generalized version of the BHH theorem can be of great value when dealing with real-world transportation scenarios. Urban environments often exhibit non-uniform demand distribution due to various socio-economic factors. By using the square root of the spatial demand distribution in our calculation, we can make better approximations of the expected travel distances, leading to more efficient route planning and resource allocation.

2.6 Capacitated Version

Now with fixed capacity C, the expected tour length of a TSP problem is

$$E[L_n] \sim \beta \int_A n/C\sqrt{C\rho(x)\,dx} = \beta n/\sqrt{C} \int_A \sqrt{\rho(x)\,dx}$$

Notice that the capacitated version has a tour length that is no longer proportional to the square-root of the demand, but rather linear in demand. In a time-space version of the problem where demand shows up sequentially, we can approximate the demand by the arrival rate of demand (λ) multiplying the time it takes to complete

3 Analysis of Heritage's Demand in Fiscal Year 2023

3.1 Dataset Description

The dataset we are given is a record of shuttle operations. Each record corresponds to a stop made by a shuttle and encompasses a spectrum of information, categorized into various columns:

1. **Shuttle**: This column signifies the route name of the shuttle service. Grouping or filtering the data based on specific routes can be performed using the values in this column.
2. **Date**: Records the date on which a particular journey or stop occurred. It can be used to identify trends and patterns in the data over time, including daily, weekly, or seasonal patterns.
3. **Stop Time**: This is the exact time at which the shuttle arrived at a specific stop. When used in conjunction with the 'Date' column, it provides a complete temporal context for each record (or shuttle stop).
4. **Veh**: Represents the identification number of the vehicle. This column is crucial for tracking individual shuttles and analyzing their performance or usage.
5. **Odometer**: Displays the vehicle's odometer reading at the time of the stop. This data can be used to compute the distance traveled by the vehicle, thereby understanding its operational efficiency.
6. **Address**: Records the physical address where the stop occurred, which can be used for mapping the stops to real-world locations and understanding the spatial distribution of the stops.
7. **Latitude** and **Longitude**: These columns contain the geographic coordinates of each stop. These values are particularly valuable for spatial analysis, enabling the plotting of stops on a map, computation of distances, or understanding the geographic distribution of the service.
8. **Passenger On**: Shows the number of passengers who boarded the shuttle at a particular stop. This column is useful for analyzing demand patterns at different locations or times.
9. **Passenger Off**: Records the number of passengers who alighted from the shuttle at the given stop. Alongside 'Passenger On', it allows for a comprehensive understanding of passenger movements and the shuttle's load factor at each stop.

By analyzing this dataset, valuable insights can be drawn to improve the efficiency and utilization of the shuttle service, predict demand, optimize routes, and address various transportation planning and management issues.

3.2 Data Cleaning

In any data-driven project, data cleaning is an integral and often a preliminary step. A dataset may contain inconsistencies, inaccuracies, or anomalies that can distort analysis results and lead to misleading conclusions. These inconsistencies might arise from several sources such as data entry errors, system glitches, or missing entries. Ensuring that the data is 'clean' is therefore paramount for the success of subsequent stages of a data analysis project, including exploration, visualization, and modeling.

Missing values represent a common issue in many datasets. A dataset with missing values for certain observations might lead to biased or incorrect results. Hence, it's important to address this issue during the data cleaning process. There are several strategies for dealing with missing values, such as discarding the records, imputing the missing values, or using statistical models that can handle missing data.

Another common issue in datasets is the presence of 'scrambled' or 'messy' text. This could be due to encoding issues, data entry errors, or issues with the data collection process. Dealing with scrambled text can be a complex task, requiring text processing techniques to identify and correct these issues.

It's important to remember that there isn't a 'one size fits all' strategy for data cleaning. The approach should be informed by the nature of the dataset, the intended analysis, and the specific research questions being addressed. As such, data cleaning isn't just a process of removing 'bad' data, but rather an iterative process of understanding the data, identifying potential issues, and making decisions on how to handle them.

Finally, data cleaning is not just a preliminary step. It's an ongoing process that might need to be revisited as new data is collected, as the analysis progresses, or as new issues are identified. As such, documenting the data cleaning process is essential to ensure the reproducibility and reliability of the analysis.

3.3 Creating Demand Maps for Each Day and Each Route

This Python script employs the pandas, folium, numpy, and matplotlib libraries to generate and visualize demand maps for different shuttle routes on varying days (e.g., Figure 1). The script processes a dataset comprised of information about shuttle stops, the number of passengers, among other details. Here is a detailed breakdown of the script's functionality:

1. *Setup:* The output directory for the maps is designated, and it is ensured that this directory exists.
2. *Data Preparation:* The 'Date' column in the dataset is converted to a datetime format. The unique dates and shuttle routes present in the dataset are identified, which will serve as the basis for generating the demand maps.
3. *Iteration over Dates and Shuttle Routes:* For each unique date and shuttle route, a subset of the data is constructed, which contains only the records for that particular date and route. Moreover, entries with empty 'Shuttle' names are removed.
4. *Data Filtering and Aggregation:* The dataset is further filtered to include only records where the location (latitude and longitude) falls within the boundaries of Pennsylvania. The latitude and longitude values are rounded to three decimal places to group nearby locations together. The 'Stop Time' is transformed into seconds past midnight and normalized to lie within the interval $[0, 1]$. The dataset is then aggregated by unique location, with the 'Pgr On' values for each location summed up.
5. *Map Creation and Visualization:* With the folium library, a map is created, which is centered around the first location in the dataset. Circle markers are added to the map for each unique location, with the color of the marker representing the 'Stop Time'. Each marker has a popup that contains information about the location, stop time, and the number of passengers. A separate marker, placed at the mean latitude and longitude of all locations, shows the total passenger count for the day.
6. *File Saving:* Each map is saved as an HTML file in the designated output directory, with the filename containing the date and route for easy identification.
7. *Index Page Creation:* Upon the creation of all maps, an index HTML page is created to facilitate easy access to all the maps.

The code employs a series of data cleaning, data transformation, and data visualization techniques to create an insightful visualization of the demand for different shuttle routes on varying dates. This allows for a profound understanding of the patterns in shuttle usage, which can inform decisions about route planning and scheduling.

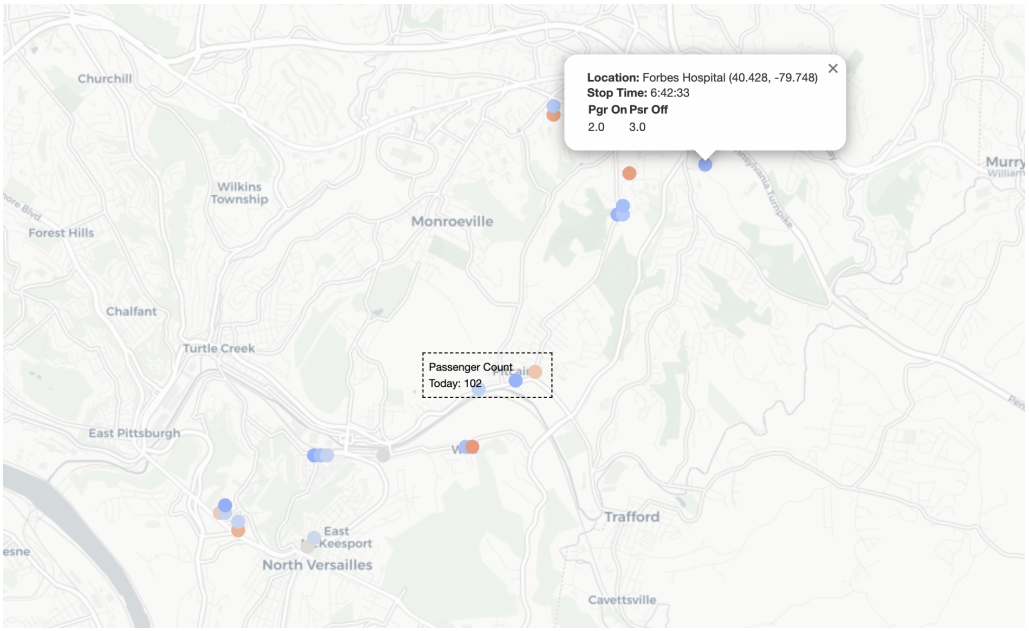

Figure 1: Monroeville Demand for July 1, 2023.

3.4 Calculating Average Daily Mileage

The Python script makes use of pandas to calculate the average daily mileage driven for each shuttle route. The data for this operation is taken from a pandas DataFrame ('df1'), which consists of columns for 'Date', 'Shuttle', 'Veh B' (the vehicle identification), and 'Odom' (odometer readings). Here is a step-by-step breakdown of the script:

1. *Date Conversion:* The 'Date' column in 'df1' is converted into datetime format, if it's not already.
2. *Data Grouping and Aggregation:* The DataFrame 'df1' is grouped by the date, shuttle, and vehicle. For each group, the maximum and minimum odometer readings are extracted. This operation results in a new DataFrame.
3. *Total Miles Calculation:* The total miles driven each day are calculated by subtracting the minimum odometer reading from the maximum odometer reading for each day. This data is stored in the 'Total Miles' column.
4. *Flattening Column Headers:* The column headers, which are multi-index after the aggregation, are flattened to single index for simplicity.
5. *Total Miles Calculation for Each Route:* The DataFrame is grouped by 'Shuttle', and the 'Total Miles' are summed up. This gives the total miles driven for each route.
6. *Count of Unique Dates Calculation:* The DataFrame 'df1' is grouped by 'Shuttle', and the number of unique dates for each route is calculated. This gives the total number of days each shuttle operated.
7. *Average Miles per Day Calculation:* The total miles for each route is divided by the number of unique dates for each route, resulting in the average miles driven per day for each route.

The end result of this script is the average miles driven per day for each route. This insight could be beneficial for planning vehicle maintenance schedules and assessing the efficiency of the routes.

Route	Average Daily Mileage (miles)
East Pittsburgh	207.66
McKeesport	266.49
Monroeville	210.68

Table 1: Average miles driven per day for each route, from July 1 2022 to May 31 2023.

3.5 Passenger-Miles Calculation

The next snippet of Python code calculates the daily passenger-miles for each shuttle route and plots them. Passenger-miles is the sum of the distances ridden by each passenger. The key steps in the code are:

1. *Data Cleaning:* The code starts by creating a copy of the dataframe and removing rows with missing or empty values in the 'Shuttle' column.
2. *Next Stop Distance Calculation:* The code then calculates the distance to the next stop by shifting the 'Dist To Prev Stop' column up by one row within each group of 'Date' and 'Shuttle'.
3. *NaN Replacement:* Any resulting NaN values (which will occur in the last row of each group because there's no "next stop") are replaced with 0.
4. *Passenger-Miles Calculation:* The passenger-miles for each stop is calculated by multiplying the number of passengers ('# Psr') with the distance to the next stop ('Next Stop Dist').
5. *Grouping and Summation:* The dataframe is then grouped by 'Date' and 'Shuttle', and the 'Passenger-Miles' are summed up to calculate the total passenger-miles for each day for each shuttle route.
6. *Plotting:* The code then plots the passenger-miles over time for each shuttle. It also calculates a 3-day running average of passenger-miles and plots that too. The plotting is done using Matplotlib and Seaborn, with separate subplots for the raw data and the running average.

Summary Statistic	
mean	239.64
std	74.63
min	82.50
25%	187.10
50%	230.20
75%	290.50
max	451.00

Table 2: Passenger-miles summary statistics for East Pittsburgh.

3.6 Estimating the Cost Effectiveness of On-Demand Service Mode

Now let's see if on-demand mode can provide better service. In particular, let's just focus on the mileage and focus on March 1, 2023 for McKeesport.

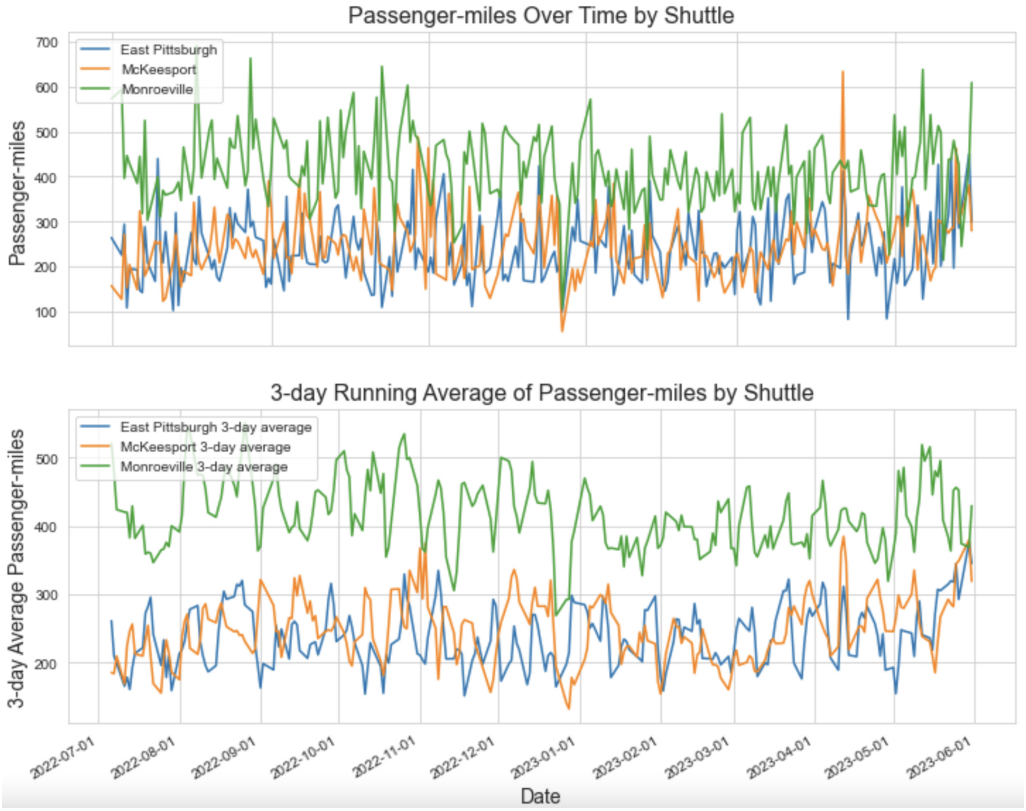

Figure 2: Passenger-miles (total miles travelled by all passengers) over time for each route.

First, we need to calculate the distances between each pair of stops, and associate each passenger getting on a bus with a potential drop-off location. We will make an educated guess on the drop-off location for each passenger, by randomly assigning one of the following stops as the drop-off point. After that, we can use a route optimization algorithm or method to find the shortest path that visits each stop that has passengers getting on or off. This is a simplified approach and would not take into account multiple vehicles, or the timing of passenger pickups and drop-offs. It would provide a rudimentary way to understand the complexity of the task and see one possible routing plan.

Here is the description of the Python code that uses a simplified way to randomly assign a drop-off point for each passenger:

1. *Data Preparation:* The script starts by making a copy of df1, and then it removes rows where the 'Shuttle' column is either missing or empty. It then filters the data to only include data from March 1, 2023, and from the 'McKeesport' shuttle route. The index of the DataFrame is then reset to make iteration easier.

2. *Distance Calculation:* An empty list 'distances' is initialized to store the distances traveled by the vehicle for each passenger. The script then iterates over the rows of the DataFrame. For each row, it simulates each passenger getting on the bus by iterating over the range of the number of passengers getting on at that stop ('Pgr On').

   For each passenger getting on the bus, it stores the current stop's latitude and longitude as the pick-up point. It then iterates over the following stops until it finds one where passengers are getting off ('Psr Off' > 0). It assumes this is the drop-off point for the passenger, and stores its latitude and longitude. It also decreases the number of passengers getting off at this stop by 1, to simulate the passenger having gotten off.

   It then calculates the geodesic distance (i.e., the shortest distance on the earth's surface) between the pick-up and drop-off points using the geopy.distance.geodesic() function, and adds this distance to the distances list.

Summary Statistic	
mean	248.84
std	72.65
min	55.60
25%	197.80
50%	239.00
75%	292.30

Table 3: Passenger-miles summary statistics for McKeesport.

Summary Statistic	
mean	418.01
std	83.09
min	106.80
25%	362.30
50%	413.05
75%	470.58
max	692.10

Table 4: Passenger-miles summary statistics for Monroeville.

3. *Total Distance Calculation:* After iterating over all the rows (and thus all passengers), it sums up the distances in the distances list to get the total miles that vehicles would have to travel under these assumptions. This total is then printed out.

This script is making a simplifying assumption that passengers get off at the next stop where any passengers get off. In reality, a passenger's drop-off point could be later, and would depend on their personal destination. However, without further data on each individual passenger's destination, this provides a reasonable approximation.

Assuming the passengers get off at their nearest next stops (get off at next s top if possible, otherwise wait until subsequent stops). Total miles that vehi cles have to travel to pick up and drop off passengers for this day and this route is about 67 miles.

Alternatively, we can also assign the passenger's get-off locations more randomly while still respecting the drop off data to provide another perspective.

1. *Creating Weighted List of Drop-off Indices:* Before iterating over the passengers, the script creates a list of indices representing possible drop-off points. Each index is duplicated in the list to match the number of passengers disembarking at that stop, according to the 'Psr Off' column. This list is then shuffled to randomize the order of the indices.

2. *Assigning Drop-off Points:* As the script iterates over each passenger, it no longer seeks the next stop where any passenger disembarks. Instead, it randomly selects and removes an index from the drop-off indices list and utilizes it to identify the latitude and longitude of the drop-off point.

3. *Distance Calculation:* The geodesic distance between the pick-up and drop-off point is calculated, much like in the previous script, and added to a list of distances.

4. *Total Distance Calculation:* The script finally sums all the distances from the list to yield the

total miles that vehicles would need to traverse under these assumptions. This total distance is then printed.

Total miles that vehicles have to travel to pick up and drop off passengers is 65.10.

The radius of the McKeesport service region is about 1.5 miles. So the conservative estimation is, the travel distance between a drop off location and the next pickup location is on average 1.5 miles. Given that there are about 30 to 40 passengers per day, the to and from travel distance is on the order of 60 miles too. So in total on-demand routing would require about 120 miles for this day.

From the previous mileage calculation, we find that the mileage for McKeesport is 270 miles. More than 2 times the on-demand distance. So roughly speaking, the on-demand modes would cut the vehicle mileage by half.

Analysis from Previous Project Cycle (for Fiscal Year 2022)

4 Ridership and Service During COVID-19

Two datasets were provided by Heritage Community Initiatives and its transportation arm, Heritage Community Transportation (HCT).

- Ridership data collected from March 1, 2019 to December 31, 2021.
- Pass-up data collected from June 25, 2020 to June 1, 2021.

The primary data source used to create visualizations throughout this section was the ridership dataset which consists of GPS tracks of vehicles and ridership information in East Pittsburgh, Monroeville, and McKeesport. Basic information such as date, stop time, route, stop location, number of passengers boarding and alighting, number of passengers on the shuttle, and distance to the previous stop were recorded.

In addition to the ridership data, pass-up data is also a specific metric to HCT's service. A pass-up is recorded by the shuttle operator when there are passengers waiting at a stop, but they are unable to board due to capacity limit. Information such as date, stop time, route, stop location, outbound direction, number of passengers passed up, and maximum capacity were recorded.

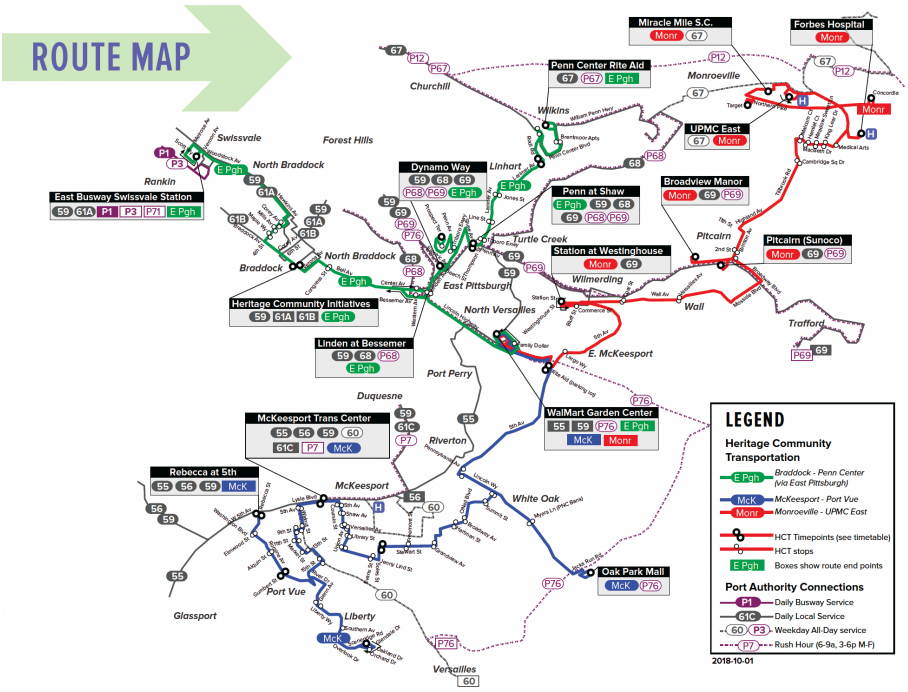

Figure 3: Heritage Community Transportation has three fixed-routes in Allegheny County: McKeesport, Monroeville, and East Pittsburgh. Service map from the Heritage Community Initiatives website (accessed April 2022).

Figure 4 depicts HCT's weekly ridership from 2019 to 2021. The clear drop in March 2020 shows that HCT ridership dropped more than 60% compared with pre-COVID ridership.

In Figure 5, the bubble maps show the spatial distributions of demand, before and after COVID-19 outbreak. The size of the bubble represents the number of riders boarding at a given location. The key observation is that spatial distribution did not shift across locations before and during COVID-19. The magnitude of demand changed across locations almost uniformly.

4.1 Supply Side Analytics

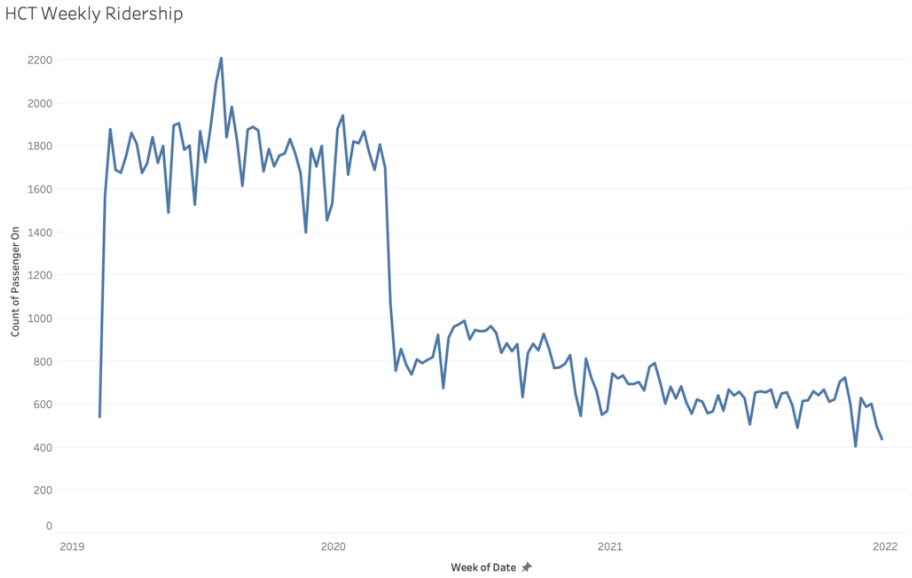

Figure 4: HCT ridership (weekly average) from March 2019 to December 2021.

To understand the sustained low demand level, and its potential causes from the demand side and supply side, we first looked at the supply side of the story: did HCT's service quality and reliability change during COVID? **Overall, HCT provided service in a sustainable and reliable way through COVID.** We illustrate the steps we took to arrive at this conclusion.

Figure 6 demonstrates the change in weekly ridership and major events such as capacity changes or constructions that happened during the time frame. The figure underscores the relationship between the change in ridership and the events. We observe that HCT's shuttle capacity limit in 2020 is unlikely the reason behind the *sustained* low ridership level during COVID-19. The observation can be supported by regression discontinuity analysis around capacity decrease and increase events, as well as weekly passenger count from 2019 to 2021.

In addition, one implication from this analysis is that it is possible that given enough time, ridership is able to bounce back once COVID-19 concerns subside in the future – if we assume residents' work / shopping / appointments patterns resume to pre-COVID style and their transportation choice revert back too. But of course that may not be true, and our study will focus on identifying robust service modes.

Therefore, service changes may have led to short-term ridership fluctuations, but do not explain medium/long-term changes. We further look into more detailed operational statistics to look for the possibility of more granular service change. Our conclusion is that HCT maintained its service level through COVID.

In particular, the length and variability of shuttle's run times are important metrics to evaluate the transit service reliability. Our analyses include the travel time of two scenarios: round trip time and time between any pair of stops. Based on ridership data, we know that shuttles did not have to stop as frequently due to ridership decrease, thus we hypothesized that it could lead to faster travel times and therefore deviate from the shuttle schedule. Analysis shows that this is not the case. HCT has maintained its service reliability through COVID.

Figure 7 and 8 illustrate the distribution of travel times of round trips and a pair of stops, respectively. Surprisingly, the average round trip travel time increases by 2 minute after COVID-19, though the difference may not be statistically significant. On the other hand, travel time between a pair of stops (in this case, Giant Eagle Oat Park and Soles at Stewart on the McKeesport route) remains the same before and after COVID-19.

In addition, we also examined the arrival time of the shuttle at all major stops. The conclusion

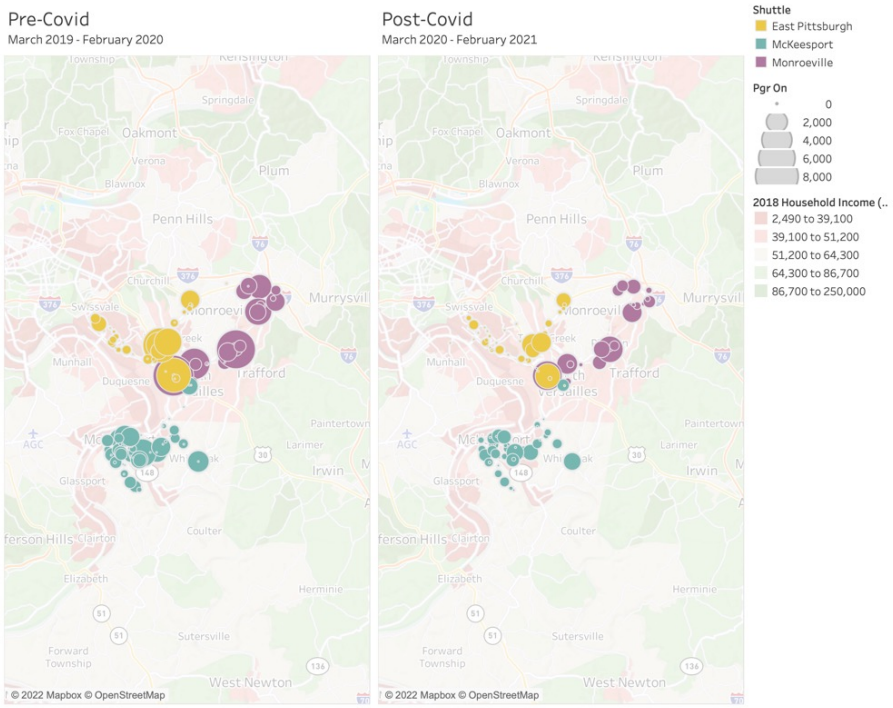

Figure 5: Travel demand distribution across HCT's service area. Spatial demand distribution did not shift, and only decreased in magnitude uniformly.

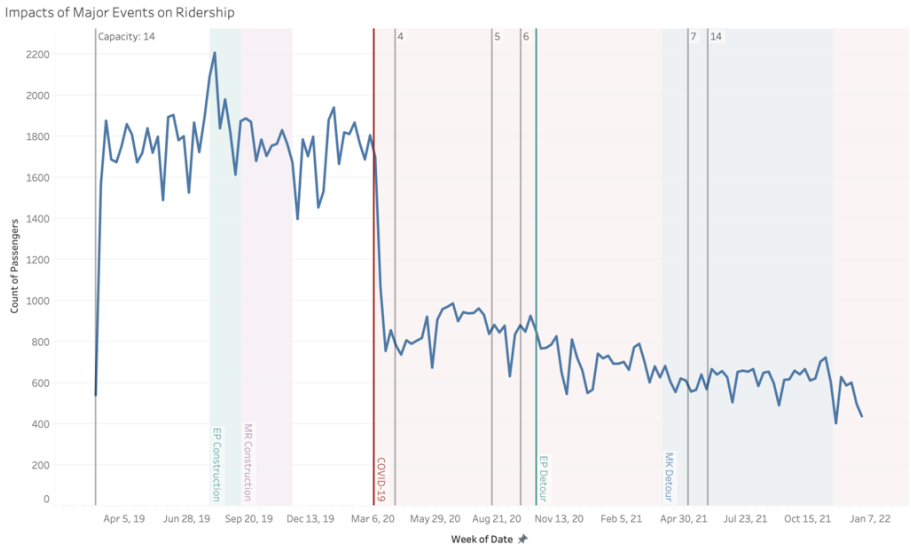

Figure 6: HCT's weekly ridership overlayed with major events. EP refers to the East Pittsburgh route, MK refers to the McKeesport route, MR refers to the Monroeville route. Demand level did not recover after shuttle capacity limit was relaxed, indicating that the travel demand level is still low, and / or the choice of transportation mode has changed.

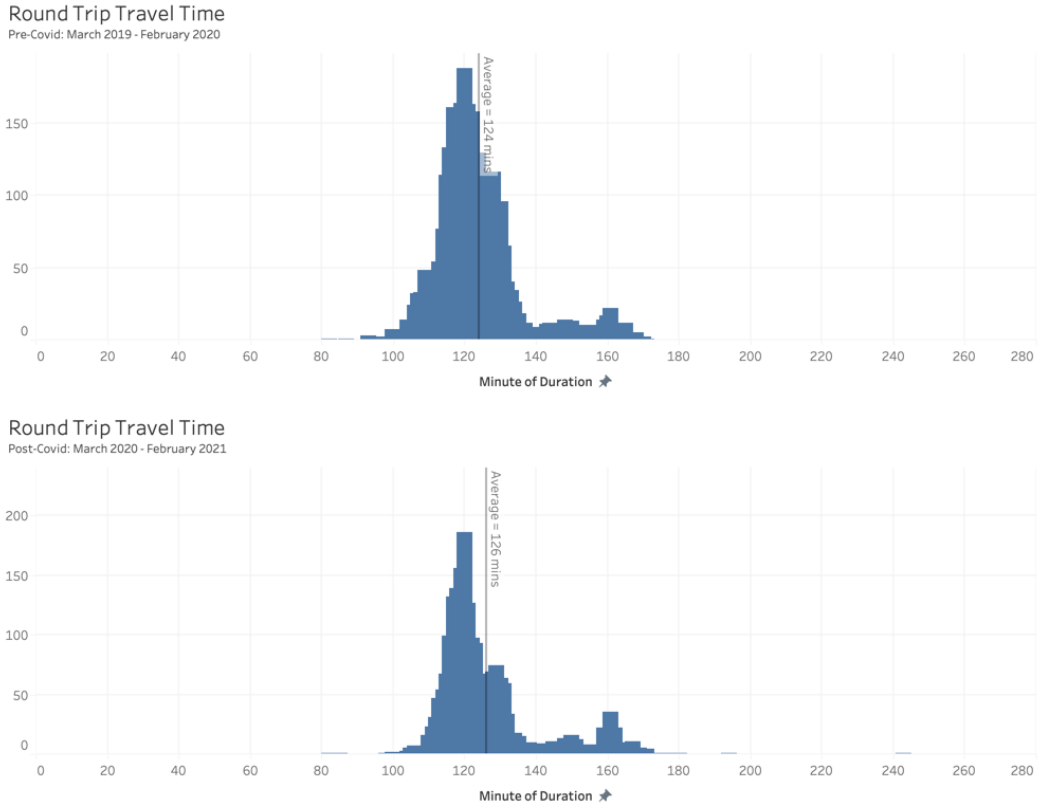

Figure 7: Average round trip travel time remained roughly the same after COVID-19.

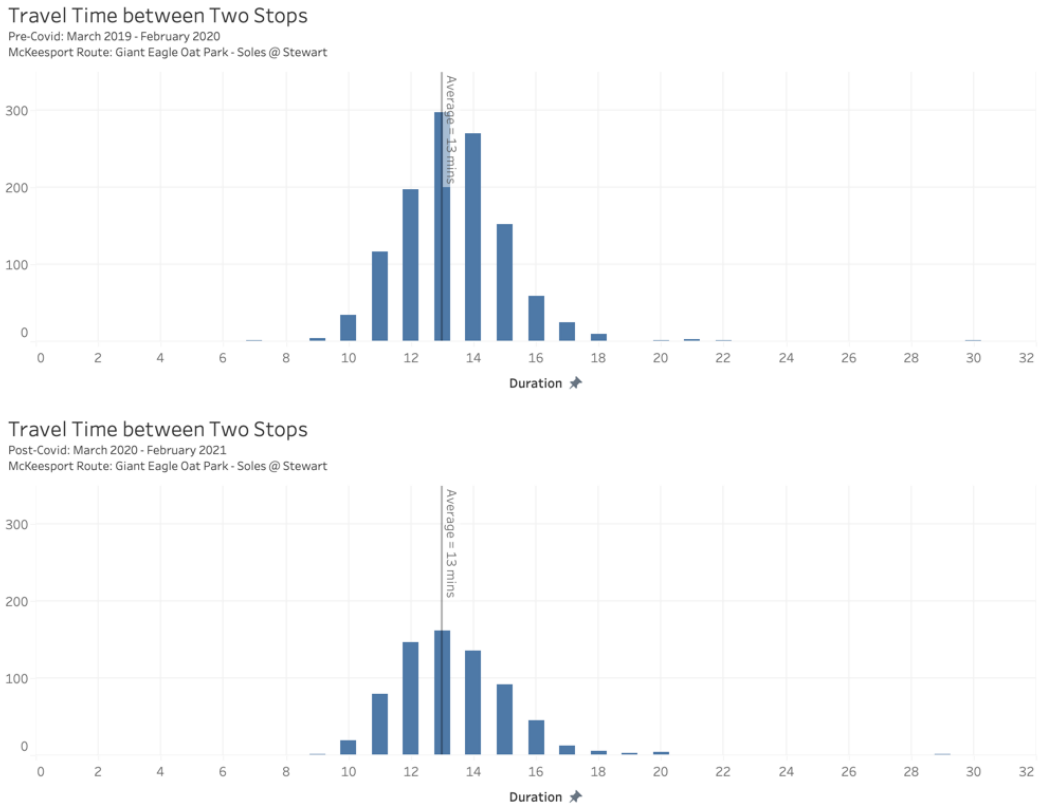

Figure 8: Average travel time between two stops remains the same before and after COVID-19. This plot shows one example: between Giant Eagle Oat Park and Soles at Stewart.

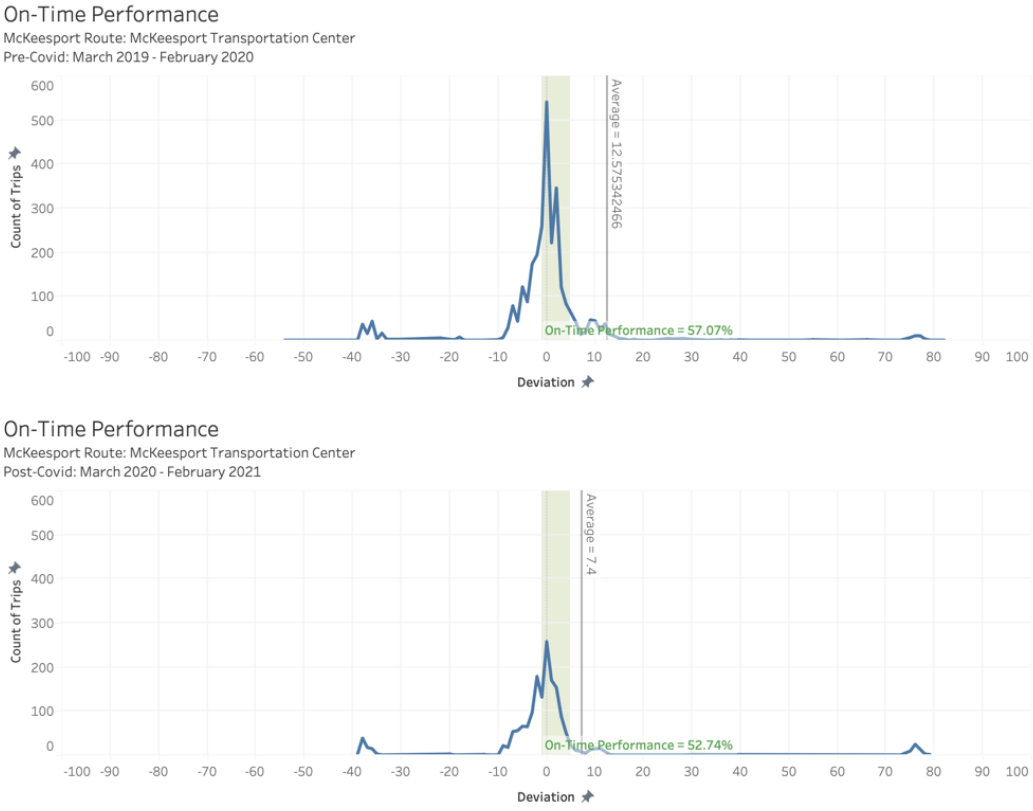

Figure 9: Arrival time distribution at McKeesport Transportation Center. The pattern remained unchanged through COVID.

is the same: arrival times did not change through COVID. Figure 9 illustrates this with the McKeesport Transportation Center stop.

4.2 Demand Side Analytics

Given that HCT's service remained steady and reliable through COVID, yet the demand level is at a sustained low level, we look at the travel demand. There are at least two layers of factors: changes in people's overall demand level, and changes in people's transportation mode choice. **Our conclusion on the demand side is, travel demand has slowly recovered but is significantly below pre-COVID level (by early 2022), and some residents have shifted their transportation choice away from affordable mass transit, and opted to use more expensive but also more flexible and safer individual transit options (jitneys, TNC rides).**

We examined demand level in HCT's service region quantitatively and qualitatively. Quantitatively, we compared the ridership level of HCT with the ambient demand level (measured by SafeGraph's mobility data) and Port Authority's ridership level on the routes that connect with HCT's region. Figure 8 indicates that there is a very slow recovery of travel demand in the region. Therefore, compared with the gradual decline in HCT ridership, we hypothesize that residents have shifted their transportation service choice slightly away from HCT for now.

We looked into survey data from HCT's service region, and observed that indeed some residents have shifted away from HCT, and have chosen to use jitney, Uber/Lyft, and other transportation modes more often. But the trend is not significant.

Thus two questions remain: Given that people are using ridehailing options more often, can

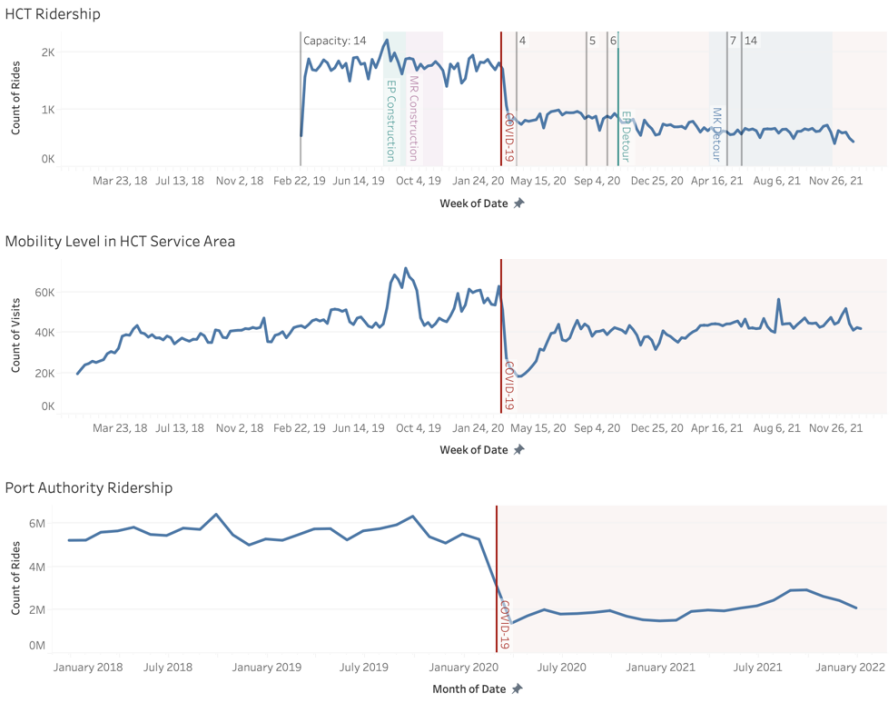

Figure 10: The comparison of HCT ridership, ambient mobility level, and Port Authority ridership. While there is a slow recovery of travel demand, HCT's ridership remained on a slow downward trend. This analysis and additional survey data indicate a shift in riders' transportation choice away from cheap mass transit ($0.25 per rider for HCT) to more expensive and flexible personal transit options such as jitneys and TNC rides.

HCT's service still provide value to its community? If so, should HCT change its service modes to adapt to the demand? We answer the first question with an in-depth top-down analysis of the economic benefits of first-mile and last-mile transportation service. We also propose a research plan for more demand-aware transportation service modes for community transportation provider like HCT.

5 Economic Value of First-Mile and Last-Mile Transportation

In the first subsection, we visualize the income levels and vehicle ownership in the neighborhoods that HCT serves. We show that **HCT's service region include many vulnerable neighborhoods that have low household income and low private vehicle ownership**.

Next, we use data analytics tools (GoogleMap API) to quantify the maximum potential benefit of providing efficient first-mile and last-mile service.

5.1 Income and Mobility Visualization

Figure 11 shows the income level of some neighborhoods around Pittsburgh, focusing on HCT's service region. The dark green dots are HCT shuttle stops. Majority of the neighborhoods served by HCT have an annual household income between $13,400 and $50,400.

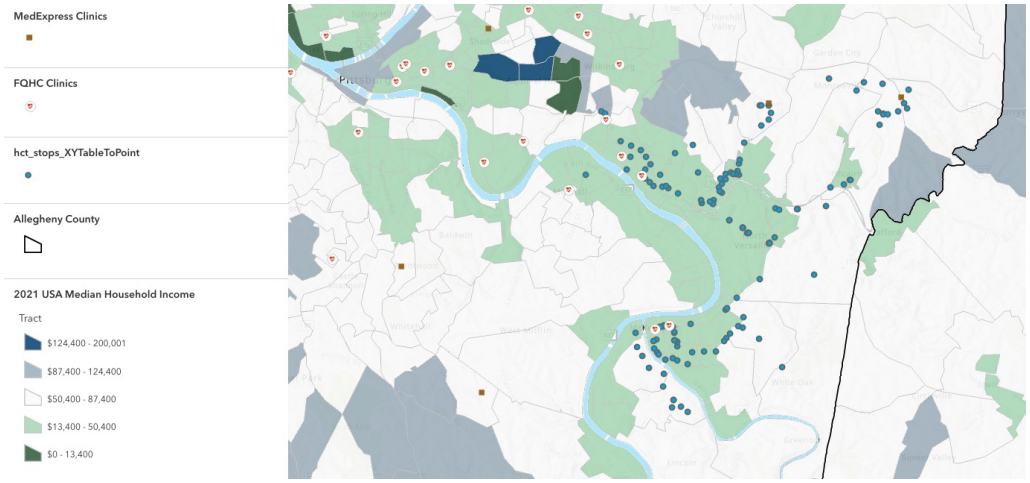

Figure 11: HCT's routes serve residents of low-income neighborhoods. Income and map data from US Census 2020. Map layer created by Esri, accessed April 2022. Overall map created by authors of this report on ArcGIS website.

Figure 12 shows that HCT's service region covers many neighborhoods with high percentage of households that do not have access to vehicles.

5.2 Maximum Potential Benefit for Reducing Commuting Time

In the previous section, we show that residents have low levels of income and access to private vehicles. In this subsection, we further show that they can save time from using first-mile and last-mile transportation service to and from public transit (Port Authority).

In particular, we quantify the maximum potential benefit that a first-mile and last-mile service can provide for these residents in terms of commuting between home and work. We focus on job-related commuting because of data availability. By focusing on this, we do not imply that HCT should focus on job commuting only. To quantify the total benefit that HCT is providing in terms of other activities (e.g., access to health care, grocery shopping, leisure, and education), additional datasets are required and they are not available at the moment.

In the remainder of this section, **we quantify the maximum benefit that a first-mile and last-mile transportation service provider can provide**.

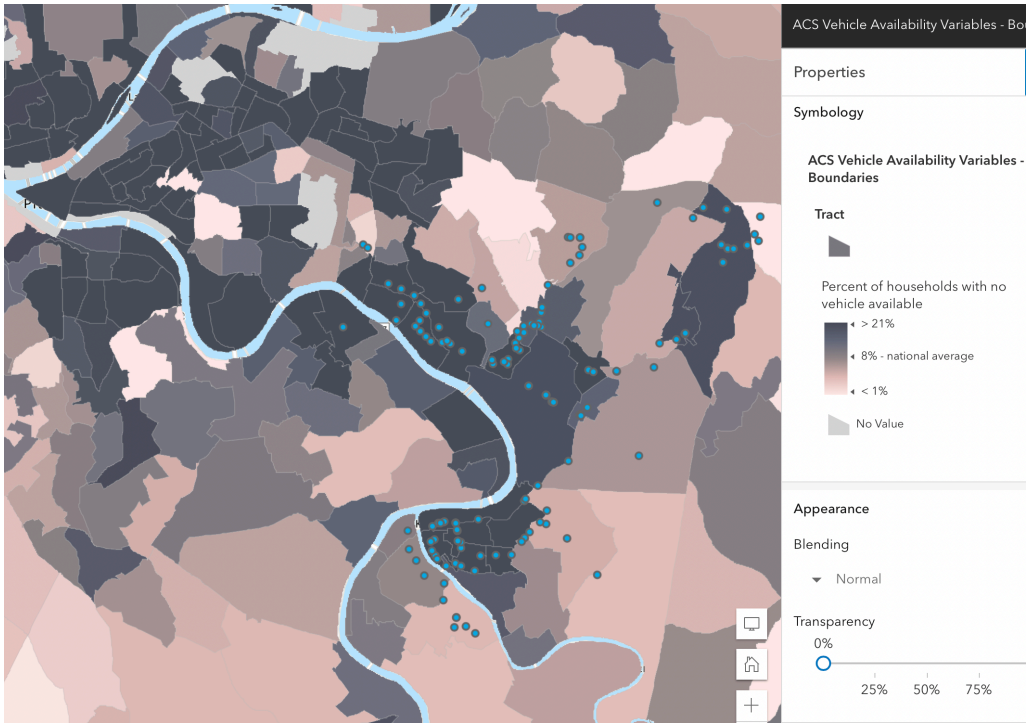

Figure 12: HCT's routes serve residents with fewer vehicles. Vehicle ownership data from American Community Survey, and map layer created by Ersi. Date of API call to the data: March 17, 2022. Overall map created by authors of this report on ArcGIS website.

Assumptions. We make the following assumptions.

- We focus on job related commuting only for this analysis, due to data limitation. (We do not imply HCT would or should only provide job related transportation service).
- We assume that all the first-mile and last-mile services are on-demand, to calculate the maximum benefit of such service.
- We assume that the individuals in the dataset used are representative of HCT's service region. Further demographic and economic distribution information could be found in the bias analysis.

Tools. We used Google Map API to request the possible transit methods, time, costs, and distance of the workers' daily commute. For the consistency of different transit methods, all the workers are set to arrive at their destinations at the same time.

Method. We calculate the economic value of time saving for commuters by the amount of time saved and their hourly salary. Time saved for transit commuters is calculated by the difference between the first-mile and last-mile walking time (requested from Google Map API via walking mode) and the expected travel time if they take on-demand shuttle service (requested from Google Map API via driving mode). In addition, the expected hourly salary is calculated by dividing the average values of the monthly salary recorded in the wages dataset by the average working hours in the United States.

Data. The data we used for this analysis comes from LEHD Origin-Destination Employment Statistics (`https://lehd.ces.census.gov/data/`), or LODES data, a synthetic dataset that describes geographic patterns of jobs by their employment locations and residential locations, as well

as the connections between these two locations. Specifically, the data includes three data sources, mainly as follows:

1. The Unemployment Insurance Wage Data, which is reported by employers and maintained by each state to administer its unemployment insurance system, providing information on employees and jobs (relationship between employee and firm).
2. The Quarterly Census of Employment in Wages, which publishes a quarterly count of employment and wages reported by employers.
3. Office of Personnel Management (OPM) - sourced data, which covers more government-related employment information.

By the given definition, a job is counted if a worker is employed with positive earnings during the reference quarter as well as in the quarter prior to the reference quarter. In addition, if a worker is employed at more than one job during the referenced period and the core datasets cover those jobs, then all of those jobs will be captured in the dataset. Besides, these datasets currently exclude several groups of workers: uniformed military, self-employed workers, and informally employed workers.

We look at the data collected in 2019 in Pennsylvania for all job types, from residential locations to employment locations. The total number of records is 5,128,507, and the corresponding features includes the number of jobs in different age groups, income levels, and industries by residential and work locations.

Potential Bias by Using These Datasets. We document the potential bias introduced from using the aforementioned datasets for our analysis. The number of records in the 2019 Pennsylvania origin-destination data files is 5,128,507. Each record represents the number of jobs between a work location and a residential location. By adding up the number of jobs from each record, we could compute the total number of jobs in Pennsylvania to be 5,513,582. If a worker is employed at more than one job during the referenced period and those jobs are covered by the core datasets, then all of those jobs will be captured in the dataset, potentially creating duplicate entries for the same person. Based on the data from the U.S. Bureau of Labor Statistics, the size of the labor force at the end of 2019 is 6,571,438. The data we use in this report represent a large portion of the workforce, but still miss a non-trivial segment that we do not know how to recover.

Economic Value. For the service region that HCT runs through, we estimate that **the maximum benefit that a first-mile and last-mile shuttle service could provide is a saving of 16,002.08 hours for 33,905 workers every day** (Figure 15).

In particular, Figure 13 shows:

- First-mile service could save 32.45 minutes per person per day, for 13,948 workers living in but not working in Monroeville, McKeesport, or East Pittsburgh neighborhoods.
- Last-mile service could save 15.09 minutes per person per day, for 16,790 workers working in but not living in Monroeville, McKeesport, or East Pittsburgh neighborhoods.
- First-mile and last-mile service could save 80.23 minutes for 3,167 workers that both live in and work in Monroeville, McKeesport, or East Pittsburgh neighborhoods.

Based on the pattern, first-mile and last-mile services save the most average commute time for local residents who live and work in the same county, and the most total commute time for those who live in these counties and work outside.

To further quantify the economic value, we borrow data from the Bureau of Labor Statistics (https://www.bls.gov/emp/tables/output-by-major-industry-sector.htm, accessed May 2022).

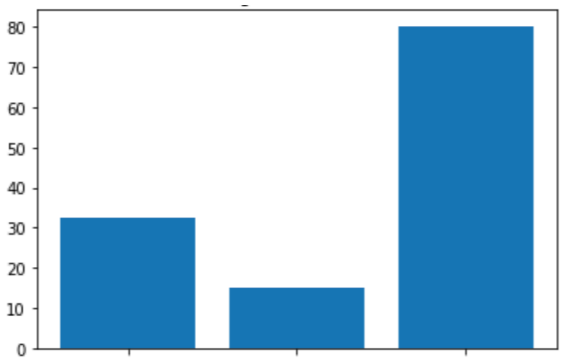

Figure 13: Minutes saved per person per day (y-axis), for those who live in HCT service region but work outside (left column); work in HCT service region but live outside (middle column); and live and work in HCT service region (right column).

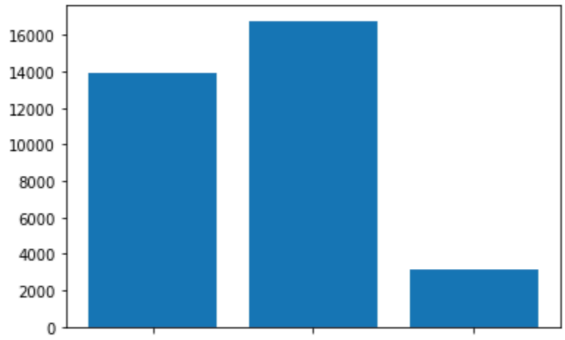

Figure 14: Number of workers (y-axis) separated into three groups: those that live in HCT service region but work outside (left column); work in HCT service region but live outside (middle column); and live and work in HCT service region (right column).

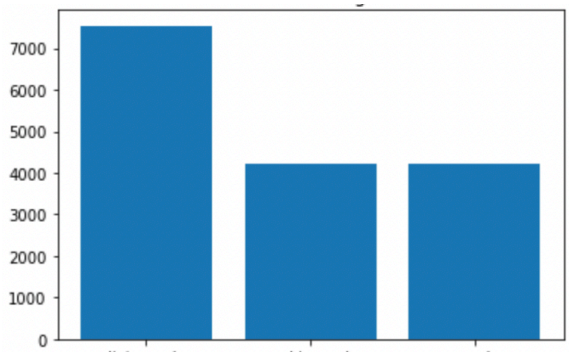

Figure 15: Hours saved per day for all workers (y-axis), for those who live in HCT service region but work outside (left column); work in HCT service region but live outside (middle column); and live and work in HCT service region (right column).

One can show that the economic value created in different sectors to be: Goods Producing industry sectors: $232.70; trade, Transportation, and Utilities industry sectors: $210.20; all Other Services industry sectors: $157.16.

In conclusion, a saving of 16 thousand hours in the HCT service region is the maximum benefit that a first-mile and last-mile shuttle service can support. This translates to $ 3.78 million dollars per day in economic output, which is equivalent to $55.73 per person in HCT's service region.

6 Conclusion and Future Work

This current research aims to understand HCT's status and travel demand after the pandemic happened, to provide actionable recommendations. We analyzed different performance metrics that may have impacted the ridership levels. Overall, we find that HCT's service remained stable amid demand decrease. Since the potential economic value of HCT's service is very high, our overall recommendation is to maintain HCT's current service and wait for demand recovery. In addition, HCT may also want to consider a few service modifications to recover its ridership level faster. In a follow-up analysis (July 2022 to June 2023), we will study these potential service modes in more detail with the support of operations research and machine learning methods.

7 Project Output

The following publications and working papers are supported under this grant.

1. Blanco, V., Japon, A., Puerto, J., **Zhang, P**. A Mathematical Programming Approach to Optimal Classification Forests.
2. Wei, N. and **Zhang, P**. Adjustability in Robust Linear Optimization (submitted).
3. Elci, O., Hooker, J., and **Zhang, P**. Structural Characteristics and Equitable and Efficient Distributions (submitted).

The following academic conference presentations are supported by this grant.

1. "Adjustability in Robust Linear Optimization", INFORMS Annual Meeting, October 2022, Indianapolis.

The following (password protected) website is built to provide interactive data analysis and visualization for the deployment and equity partner:

1. https://www.andrew.cmu.edu/user/yunz2/heritage/