Adjusting Transit Networks for Undeserved Communities using Evolutionary Algorithms

Jacob Emmerson
jte27@pitt.edu

Chris Hinson
chh183@pitt.edu

1 Introduction

With our project, we sought to detect transit equity inequalities within the Pittsburgh region's transit network, and generate potential network modifications to minimize this inequality, and provide maximal, equitable accessibility to transit for all citizens. In keeping with the spirit of AI4SG, we explored possible applications of various AI methodologies for this variant of the Transit Route Network Design Problem (TRNDP), and finally settled upon a genetic algorithm approach. This gives us a natural, computationally inexpensive optimization technique for the large system and search space we were faced with. We could model the cities transit routes as genes of an organism, and through mutation and equity focused fitness evaluation, evolve the network as a whole towards a more globally equitable state.

We first set out our problem formulation and proposed methodology. We discuss the materials and prior work we pulled upon in our problem consideration, and then lay out our model representation and mathematical problem formulation. We then discuss the challenges we encountered in cleaning our data and implementing our algorithm, and finally we discuss our results, their interpretation, and the limitations and possible future expansions of our work.

2 Materials

In building our genetic algorithm, we first needed to build a ground truth representation of the transit network, representing the system as it exists today so that we may optimize and build upon it. We first found census tract data (which is the geospatial granularity we chose to use as a first approximation of neighborhoods) available directly from the census.gov website, which gave us both a geographical partitioning of the area, as well as numerous sociopolitical stats for these partitions. All of our transit network data was pulled from Pittsburgh's Regional Transit Authority's website, and we primarily took segment data, which encompasses all available routes, at multiple times of day, multiple segments of the week, and allows us to build up the network system from its smallest components.

Our approach necessitated a large amount of ground truth data about the world and system as a whole. We can split these datasets into two main distinctions; socio-demographic data, and transit network data.

2.1 Socio-demographic

With our eye towards social equity, it was crucial to have a solid data baseline on socio-demopgrphic factors throughout the geographical region

2.1.1 Census Tracts

We needed a way to subdivide our total geographic area into measurable plots upon which we could draw demographic information. While considerations like neighborhoods lend themselves better to intuitive applications, they are ill-defined, and would require immense effort to parse meaningful data for. Instead, we chose to use census tracts. These provide subdivisions approximately the size of a neighborhood, with standardized and easily available data from the census bureau, and typically seem to mirror what a common citizen would consider to be rough neighborhood boundaries. We pulled geographic data for 402 census tracts covering the Allegheny county region, encompassing (nearly) the entire city of Pittsburgh transit network.

2.1.2 Transit user demographics by census tract

Via the census.gov website, we were able to find a wide range of demographic data for all the census tracts within the allegheny county region. We narrowed this down to 3 groupings, those who said they carpooled, those who said they commuted via public transit, and those who commuted alone. From these groupings, we further subdivided by census tract, and pulled the racial data, median income, and owner-occupied housing percentages. This data, in combination with the PRT equity index, provided the backbone for our equity analysis and demand prediction data.

2.1.3 City of Pittsburgh zoning

One dataset we considered an interesting possibility was city of pittsburgh zoning zones. With information about areas which were zoned for commercial use, single family housing, high-density housing, industry etc., we considered attempting to embed extra data about desirable destinations. Unfortunately, in pulling and parsing this data, we realized that due to the patchwork of local authorities on the matter, there was no centralized, standard dataset which covered the entire transit network boundary. The best we could do was zoning information for the city of pittsburgh, sans the surrounding allegheny county info. While this dataset proved unusable, we still considered it a worthwhile inclusion, and possible future work could seek to expand and standardize it, so that it may be included in demand calculations.

2.1.4 PRT Equity Index

The second half of our core equity analysis came from Pittsburgh Regional Transit's own equity analysis. This study considered the mobility afforded by the Pittsburgh public transit to 8 at-risk demographics, and importantly, in their published datasets, grouped the information by census tract. The demographics considerd were People with Disabilities, People in Poverty, Minority Race and Ethnicity Persons, Households without Vehicles, Older Adults, Persons under Age 18, Persons with Limited English Proficiency, and Female Householders.

2.1.5 Google Maps Places API

In the search for even more desirability data to train our demand network upon, we enlisted the Google Maps Places API. This api takes a coordinate set (which for our purposes was the coordinates of a single bus top) and returns a set of easily parseable Places within a set distance from this point. This data, importantly, includes information about the types of these places, such as restaurant, business, industry, hospital, park, etc. This gave us some extra information to calculate the demand of a segment connecting two stops, and was helpful in training our demand prediction net.

2.2 Transit network

The second half of our necessary datasets was information on the transit network itself. This data proved to be, unexpectedly, a far larger hurdle in terms of processing and cleaning compared to its demographics counterpart. Confusing, undocumented, incomplete, inconsistent, and at times outright conflicting data was present in almost every dataset we processed. This was an unnecessarily painful process, and we strongly feel that PRT has a responsibility to its constituents to keep better maintained, publicly available datasets.

2.2.1 Stops

The most important facet of a transit network, for us, was the boarding stops. These serve as the interface between humans and the transit network, and as such, fundamentally underpin how equitable service can be. As such, we focus most of our process on this granularity, analyzing and modifying transit routes in terms of the stops which comprise them. PRT maintains a database of every boarding stop for every route for every mode of transit within their system. This dataset formed the backbone of our route construction, and importantly, since every stop entry included its coordinates, we could effectively and efficiently classify which census tract every stop was contained within, and intrinsically link it to demographic information about citizens who are close enough to use it.

2.2.2 Segments

PRT also maintains a database of segments. They classify a segment as a connection between two bus stops, and maintain a list of every route which traverses this segment. We initially planned to utilize this information to reverse engineer the path of travel for every route. We could simply find all the segments which listed a given route as traversing them, and then consider these segments together to be a route. The key issue here was the lack of imposed ordering. In writing an algorithm to sort these nodes by finding their terminal segments, we quickly found a glaring problem. Nearly 40% of routes did not have well defined terminals.

The problem here was two-fold. First, some routes are either entirely a loop, or contain within them a small loop or divergence. These routes simply have no well defined single start and end, and as such, we cannot sort them into path of travel based on their member segments alone. Secondly, and more glaringly, a huge amount of lines simply had very incomplete data, leading to multiple "holes" in their path of travel. We eventually decided this data was unworkable, and had to come up with another solution.

2.2.3 Routes

PRT maintains a list of all routes within its network, however, the entries for each route are rather sparse of useful information. They contain a line name, some bookkeeping information, and then geometry data defining the path of travel. We initially discarded this dataset as it does not embed information about where along this path of travel the route stops. Faced with the shortcomings of the segments dataset however, we devised an ample approximation. We reference back to our bus stops set, and find all the stops which are denoted as being members of this route. We then iterate over all the line segments which comprise the geographic geometry of this route, and assign the stop to be on the line segment closest to it. Since these line segments are ordered in direction of travel, this approximation also yields an ordering of the route's stops at the same time. This proved to be the final push we needed for an accurate representation of the network as a whole.

2.2.4 Feasible Stop Network

For our mutations to decide on a new stop, we needed a network of feasibly reachable stops. This "feasibility" aspect is discussed more in our future work section, as it is a subtly complex problem, which could have wide reaching impacts on the overall solution quality. Our first approach to building this network was to take a real-world approach utilizing the google maps pathfinding api. However, we quickly realised this was not feasible for us due to the size of the network and googles pricing. This was a disappointment, as high fidelity information about walk-able time between two points is extremely hard to calculate. We ended up settling for a much coarser approximation of including any stop which is within a half kilometer radius.

3 Related Works

Research on transit route optimization has coined the term transit route network design problems (TRNDP). These problems generally represent transit routes as network graphs; the parameters of the graph depend on the objective of optimization. Prior work has noted that mathematical programming formulations of this optimization problem cannot be efficiently applied due to the inherent complexity of problem [5], thus heuristic or metaheuristic algorithms are favored for optimization.

Since our objective is to improve accessibility in underserved communities, we propose a solution which expands upon the prior work on equity-focused TRNDP [6]. Our primary contributions focus on an improved heuristic which additionally accounts for income and location biases. Additionally, the heuristic should attempt to find reasonably implementable routes (e.g., proposed solutions should not generate entirely new routes but instead extend and improve the existing routes). This may be implemented as a constraint when evolving current routes or culling bad routes by accounting for similarity to the currently implemented transit routes.

Additionally, prior work on TRNDP utilizes pseudo-transit [6, 9] networks which are easily constructed and accessible or apply prior methods on current transit systems in major cities. However, here we explore the potential of applying prior methods with our contributions on a real-world network subject to current constraints.

4 Method

For our initial work, we formulate the TRNDP utilizing census tracts rather than a road network; however, future work will involve a more fine-grained representation of bus networks. Our current contributions are focused on equity-encoding in our transit network prior to optimization to ensure that proposed solutions are considerate of undeserved and high demand areas.

Let $G = (V, E)$ denote a graph with a set of edges E and vertices V. An edge $e_{ij} \in E$ has a weight equal to the center-point distance between v_i and v_j for $v \in V$. We want to find a set of bus lines which maximizes the demand coverage in the network and accounting for equity factors. To prevent infeasible solutions, we penalize our solution quality by the number of bus stops in a given line; intuitively, a good bus line should maximize coverage by adding only beneifical bus stops.

4.1 Genetic Algorithm for Equitable Neighborhood Service (GAENS)

Prior research on the TRNDP has found brute force solutions to be inefficient; thus, evolutionary algorithms have been proposed for optimization over transit networks [4, 7, 9, 1, 10, 6, 5]. In our approach, we utilize a *genetic algorithm* to optimize over a set of bus lines. Genetic algorithms are a type of meta-heruistic algorithms can find near-optimal solutions over a set of configurations by constraining the search space to a reasonable set of decision variables. Using randomness and selection through a meta-heuristic to determine solution utility, genetic algorithms avoid unnecessary computational complexity given that a sub-optimal solution is okay.

Let P denote a population of individuals $p \in P$ where $|P| = n_{pop}$. Each p has a representation of the transit network called a chromosome consisting of L genes. Concretely, an individual is a set of genes $p = \{g_1, \ldots, g_L\}$ where $g_l = \{v_1 \ldots v_i\}$ for a bus line of i-stops.

The initial population, P_0, consists of the current PRT network abstracted using census tracts. Within n_{gen} generations, we want to find the individual which minimizes the total edge weights over the graph. Let our solution be p^* where

$$p^* = \arg\max_{p \in P} f(p)$$

and f is a function denoting the utility of a given individual.

4.2 Demand Modeling

To encode coverage factors into our graph, we learn a function $w : (v_i, v_j) \to \mathbb{R}^+$ which utilize features local to bus stops to make demand predictions are currently unobserved line segments. To learn w, we train a deep neural network on the 81,820 ground truth ridership metrics within our bus segments dataset.

4.2.1 Fitness

To determine which individuals have a higher utility, we evaluate each $p \in P$ at each generation utilizing a fitness function $f : p \to \mathbb{R}^+$.

$$f(p) = \sum_{l=0}^{L} \sum_{r=0}^{R-1} w(v_r, v_{r+1})$$

Intuitively, our fitness function returns the sum of all demand edges for every gene in an individuals chromosome.

4.2.2 Crossover

At each generation, we use *tournament selection* to determine which individuals will be used for crossover. From a random sample of k individuals, the two with the highest fitness become the parents for a new individual in the population which is created using crossover.

In genetic algorithms, crossover finds higher quality solutions by combining two of the previously best chromosomes – note that this does not always result in a higher quality solution. For the purposes of network design, we utilize *uniform crossover* which potentially switches every gene rather than a subsection of the chromosome. With the two parents selected, each gene in their chromosome has a probability of being swapped according to a crossover parameter $\xi_c \in [0, 1]$. This results in two offspring, each of which are a combination of their parents.

4.2.3 Mutation

After reproduction, the individuals in the new population may mutate to introduce genetic diversity into future populations. According to a mutation rate $\xi_m \in [0, 1]$, a random sample of $\xi_m(n_{pop}(1-\phi))$ individuals are drawn from the population (excluding the elitists).

Utilizing an adapted algorithm from prior research on route generation approaches [9], we introduce two potential forms of mutation: *midpoint mutation* and *big mutation*. Our midpoint mutation is an extension of the authors' original implementation of a small mutation which selects terminal nodes to shrink or expand. Our motivation for a midpoint mutation is that the current terminal nodes of some bus lines are set due to nearby facilities, thus a change in these nodes may not be truly beneficial. However, expanding terminal nodes can still be permissible in the rare, big mutation. In each individual selected for mutation, the gene with the lowest fitness is mutated according to a probability $\xi_{re} > 0.5$.

- **Midpoint Mutation**: a variation of the **small mutation** with a probability of ξ_{re}. Rather than relying exclusively on terminal nodes, we randomly select a non-terminal node according to a uniform probability distribution. With the new non-terminal node selected, a nearby node is randomly selected and reconnected according to our network graph.

- **Big Mutation**: a less frequent mutation with a probability $1 - \xi_{re}$. Similarly, one of the terminal nodes is selected at random but guaranteed to expand by greedily selecting the new terminal node.

4.2.4 Elitism

To ensure that the solution quality of each generation does not decrease, *elitism* is a popular technique. A small proportion of the most-fit individuals determined by $\phi_e \in [0, 1]$ is passed into the next generation without any modifications (though they can still produce offspring if chosen during tournament selection). Tournament selection is repeated $(n_{pop}(1-\phi))/2$ times to maintain a constant population size.

4.2.5 Culling

Similar to Elitism, we can prevent low utility individuals from reproducing by culling them prior to reproduction. According to $\phi_c \in [0, 1]$, a proportion of the least fit individuals in a population are removed prior to reproduction.

4.2.6 Priority Stops

In order to account for equity biases in our data, we define a set of preferred bus stops as $V^* = \{v_1, \ldots, v_p\}$ for p prioritized bus stops. To determine which bus stops are currently undeserved, we utilize an analysis done by PRT on equity scores. The scores calculated by PRT utilize information such as mobility, age, median income, and ethnicity.

4.3 Demand Representation

To query demands between any two stops, we utilize a demand matrix D where each index represents a given bus stop. Since bus stops cannot have a segment to themselves, $diag(D) = [0, \ldots, 0]$. For any $v \in V^*$, the corresponding column is up-weighted by a constant $\alpha = 1.5$ which was determined through random search.

4.4 Genetic Algorithm

Algorithm 1 GA for Equitable Neighborhood Service (GAENS)

```
n_pop ← number of individuals
n_gen ← maximum number of generations
φ ← proportion of elite individuals
L ← number of genes (bus lines)
k ← number of individuals for selection
ξ_c ← crossover probability
ξ_m ← proportion of mutated individuals
ξ_re ← small mutation probability
ξ_d ← deletion probability in mutations
P ← initial population
P_new ← ∅
p* ← ∅
for i in 1 ... n_gen do
    for p ∈ P do
        Fit ← f(p)
        if p* = ∅ or Fit > f(p*) then p* ← p
        end if
    end for
    P_new ← φ * n_pop fittest p ∈ P
    for j in 1 ... (n_pop(1 − φ))/2 do
        p_a, p_b ← Selection(P, k)
        c_a, c_b ← Crossover(p_a, p_b, ξ_c)
        P_new ← P_new ∪ (c_a, c_b)
    end for
    for j in 1 ... ξ_m * (n_pop(1 − φ)) do
        P_new[j] ← Mutate(p, ξ_re, ξ_d)
    end for
    P ← P_new
end for
    return p*
```

5 Results

5.1 Demand Modeling

	MSE	MAE
Training	0.4883	0.5644
Validation	0.4876	0.5649
Test	0.4931	0.5671

Table 1: Demand Error

We can see that with our demand function w, our demand predictions are fairly accurate. After further analysis, the most influential features which influence

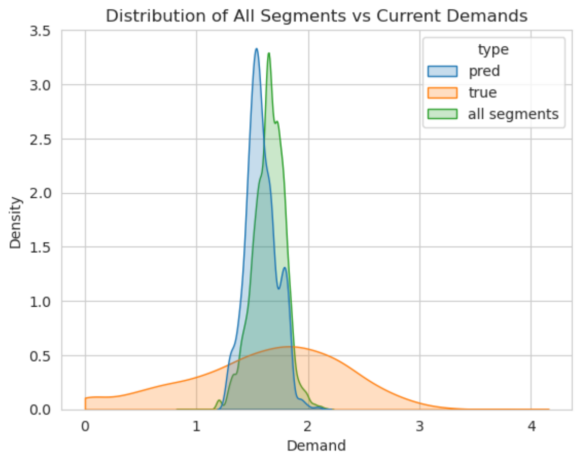


Figure 1: Demand Distributions

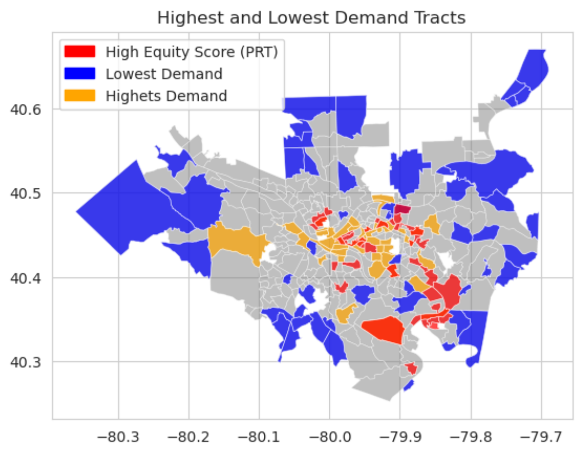


Figure 2: High and Low Demand Areas with Equity Indexes

demand are the information local to bus stops relating to retail services and food accessibility.

5.2 GAENS

Parameter	Value
n_{pop}	100
n_{gen}	100
ξ_m	0.1
ξ_c	0.5
ξ_{re}	0.9
ϕ_e	0.05
ϕ_c	0.05
k	20

Table 2: GAENS Hyperparameters

Given that there are several hyperparameters to tune in this model with a large number of potential combinations, we determined these hyperparameters through random search. The search space was broadly defined by adjusting the initial population size and mutation rates.

Population	Fitness
Initial	12007.26
Proposed_1	12929.55
Proposed_2	12927.36
Proposed_3	12925.86
Proposed_4	12925.86
Proposed_5	12925.86

Table 3: GAENS Fitness

Since we utilize elitism, our solution quality should never be lower than our initial individual. Additionally, since our fitness function is an approximation of the solution quality, it is hard to interpret fitness scores directly. Thus, we additionally evaluate a few routes from our most fit individual out of the top 5 performers.

5.3 Social Equity

We have seen some success in first approximation evaluation of census tracts transit equity by using the factors included in the Community Needs Index maintained by Pittsburgh's Department of Human services, weighed against the raw number stops within a census tract. Expanding upon this metric with factors that take into account available connections to other tracts, as well as some outside factors of our own such as commercial real estate zoning, and percentage of car ownership will be a major part of our second phase of research as we seek to build our fitness function.

5.4 Ground Truth Data

We were able to successfully build an initial system population using available data. We here present two visualizations of the system's initial population.

6 Discussion

From our most fit individuals, we see a slight increase in demand coverage for our proposed routes. Generally this is caused by additional mid-point mutation routes diverging into a road which is not accessible by alternative bus routes. We see this in the 81 and 83 bus lines where proposed bus stops follow a similar route to the current lines but result in a more diverse set of stops in neighborhoods like where there is low-accessibility without the means of a car. It is important to note that since genetic algorithm use a meta-heuristic to learn and find near-optimal solutions based on our decision variable that our proposed solutions may not necessarily be optimal in practice due to confounding factors or unaccounted for constraints.

7 Future Work

7.1 Tuning and Extensibility

One of the largest problems we continuously ran into with this project was the sheer amount of possible data

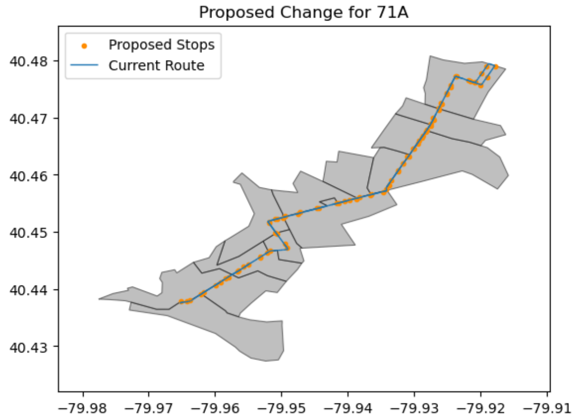


Figure 3: Generated Network for 71A

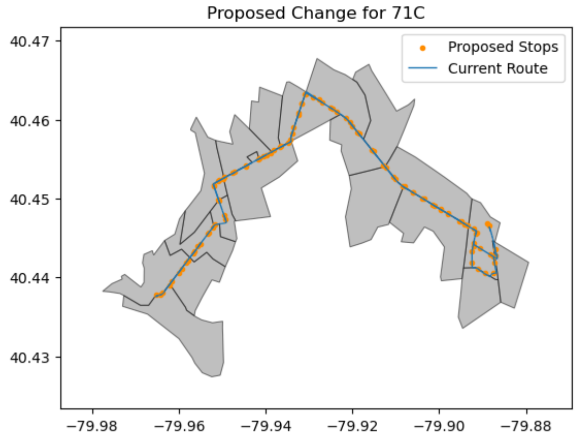


Figure 4: Generated Network for 71C

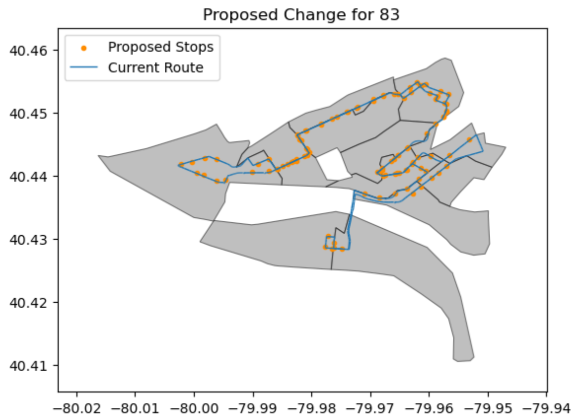


Figure 5: Generated Network for 83

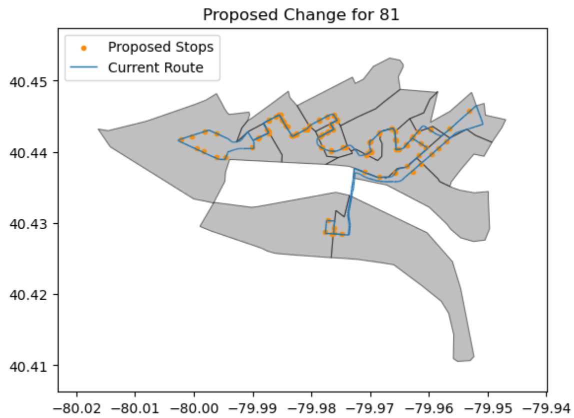


Figure 6: Generated Network for 81

Figure 7: Graph representation of the geographic layout of the system

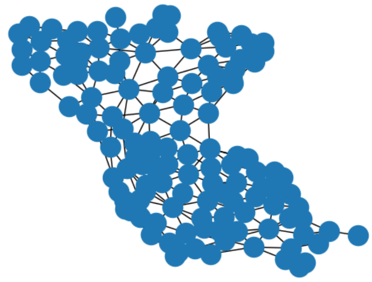
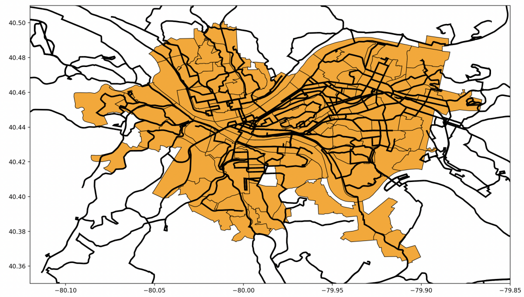


Figure 8: Human "friendly" geographic layout of the system including existing routes



considerations. Equity and inclusion is a **deeply** nuanced sociopolitical field, beyond our ability to fully consider ourselves experts in. With this in mind, we intentionally created our algorithm in a manner so as to be easily extensible if one wanted to consider new data aspects. Because our genetic algorithm is so fundamentally underpinned by the idea of calculating a member fitness, it becomes extremely easy to find new data, so long as it is mappable into the geographic system, and integrate it into fitness calculations. Similarly, our demand prediction model can be retrained upon additional factors just as easily. We welcome and encourage such work, and are open to communication surrounding factors we may have excluded or failed to consider.

In addition to the inclusion of new raw data, we feel that there is always room to fine tune both of our core AI data structures, the demand predictor and our genetic algorithm, by testing differing hyperparameters. We did a minimal amount of this to find an acceptable solution for our research, but feel that it could be taken farther.

7.2 Deployment

As with any AI4SG project, we wish we could test the impact of our research in the real world. While we can run endless simulations and predictions for what may create a better transit network, without making the changes in the real world to measure ridership and feedback, there is a certain credibility lacking in our results.

We feel that a small scale deployment test would be a perfect way to test our results. We could find a single, minimal change that we determine will have the maximal impact upon the network, and then implement this change during off peak hours so as to not inconvenience as many people if we find our results to be incorrect. One possible implementation of this might be to implement our changes only on the weekend PM routes.

7.3 Feasible Stop Network

One limitation of our algorithm came from our limited ability to determine what gene mutations may be feasible and valid. This problem ended up being far more complex than it seemed to us at first glance, and we ultimately ended up using only a general distance approximation. The complexity comes largely from the fact that navigation in a city is complex and hard. One must consider the road network, dedicated busways, bus lanes that effect speed of travel, conditional closures, costs incurred by extensions and diversions, costs incurred by extra driver pay, the bureaucracy and NIMBYism behind creating new bus stops, and a huge number of other factors not listed here when deciding "can a bus reasonably drive from this stop to this location". Ultimately, this problem could yield its own research project, and an objectively correct answer is likely impossible to find, when accounting for all factors.

7.4 Generalization and Automation

Our final area of consideration for future research came from the fact that we were focused on what is ultimately a rather small geographic region. While our dataset is too specific to Pittsburgh itself to be widely applied to other cities, our methodology is not. The largest hurdle to this goal is the ground data required. We pulled a huge amount of data for Pittsburgh, and it all had to be manually processed and cleaned. This is a time prohibitive procedure. We believe that turning this data into a more standardized expected set of data, and then giving this outline to city officials to produce, could allow our methodoloy to be automatedly and generally applied to any city in the world.

References

[1] B. Bulut et al. "OPTIMIZING BUS LINES USING GENETIC ALGORITHM FOR PUBLIC TRANSPORTATION". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVI-4/W5-2021 (Dec. 2021), pp. 131–136. DOI: 10.5194/isprs-archives-XLVI-4-W5-2021-131-2021.

[2] Tawanna R. Dillahunt and Tiffany C. Veinot. "Getting There: Barriers and Facilitators to Transportation Access in Underserved Communities". In: *ACM Trans. Comput.-Hum. Interact.* 25.5 (Oct. 2018). ISSN: 1073-0516. DOI: 10.1145/3233985. URL: https://doi.org/10.1145/3233985.

[3] Lang Fan and Christine Mumford. "A metaheuristic approach to the urban transit routing problem". In: *J. Heuristics* 16 (June 2010), pp. 353–372. DOI: 10.1007/s10732-008-9089-8.

[4] Alex Karner and Richard A. Marcantonio. "Achieving Transportation Equity: Meaningful Public Involvement to Meet the Needs of Underserved Communities". In: *Public Works Management & Policy* 23.2 (2018), pp. 105–126. DOI: 10.1177/1087724X17738792. eprint: https://doi.org/10.1177/1087724X17738792. URL: https://doi.org/10.1177/1087724X17738792.

[5] Konstantinos Kepaptsoglou and Matthew Karlaftis. "Transit Route Network Design Problem: Review". In: *Journal of Transportation Engineering* 135.8 (2009), pp. 491–505. DOI: 10.1061/(ASCE)0733-947X(2009)135:8(491). eprint: https://ascelibrary.org/doi/pdf/10.1061/%28ASCE%290733-947X%282009%29135%3A8%28491%29. URL: https://ascelibrary.org/doi/abs/10.1061/%28ASCE%290733-947X%282009%29135%3A8%28491%29.

[6] Myeonghyeon Kim, Seung-Young Kho, and Dong-Kyu Kim. "A Transit Route Network Design Problem Considering Equity". In: *Sustainability* (2019). URL: https://api.semanticscholar.org/CorpusID:198797886.

[7] Fettah Kılıç and Mustafa Gök. "A public transit network route generation algorithm". In: *IFAC Proceedings Volumes (IFAC-PapersOnline)* 1 (Jan. 2013), pp. 162–166.

[8] Sanju Maharjan, Fatemeh Janatabadi, and Alireza Ermagun. "Spatial Inequity of Transit and Automobile Access Gap across America for Underserved Population". In: *Transportation Research Record* 2678.1 (2024), pp. 674–690. DOI: 10.1177/03611981231171914. eprint: https://doi.org/10.1177/03611981231171914. URL: https://doi.org/10.1177/03611981231171914.

[9] Muhammad Ali Nayeem, Md. Khaledur Rahman, and M. Sohel Rahman. "Transit network design by genetic algorithm with elitism". In: *Transportation Research Part C: Emerging Technologies* 46 (2014), pp. 30–45. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2014.05.002. URL: https://www.sciencedirect.com/science/article/pii/S0968090X14001156.

[10] Moschoula Pternea, Konstantinos Kepaptsoglou, and Matthew G. Karlaftis. "Sustainable urban transit network design". In: *Transportation Research Part A: Policy and Practice* 77 (2015), pp. 276–291. ISSN: 0965-8564. DOI: https://doi.org/10.1016/j.tra.2015.04.024. URL:

https://www.sciencedirect.com/science/article/pii/S096585641500110X.