Improving Public Transit Systems through Clustering and Polynomial Regression

George Trieu
Queen's University
g.trieu@queensu.ca

Pavle Ilic
Queen's University
20pi@queensu.ca

Cahal Ng
Queen's University
20cn16@queensu.ca

Owen Rocchi
Queen's University
19omr6@queensu.ca

Duncan Cheng
Queen's University
17jc66@queensu.ca

***Abstract*—As urban mass transportation systems worldwide evolve, they must adapt to the ever-changing needs of their residents. This paper explores a machine-learning approach using clustering and polynomial regression. This model fits transportation corridors and hubs to optimize the population served using points of interest and hot spots. Scoring of station and line candidates is performed through a summation of Gaussian distributions between the candidate and each point of interest. A novel approach to generating transportation interchanges through heuristics is also explored.**

I. Introduction

As urban centre populations continue to develop into the 21st century, it is essential to design efficient and cost-effective mass transportation. With differences in geography and routines being local to each region around the world, there is no universal system for planning public transportation. Urban and city planning today leverages highly advanced tools, such as Geographic Information Systems (GIS) to develop transit systems. Often, this process takes years to plan and decades to execute, all while urban compositions and the needs of the public continue to evolve. Building transit for the future involves designing for people today and tomorrow, not from the past. The most significant reason car usage continues to grow in large metropolitan areas is because of the inconvenience associated with using mass transit options – a symptom of poor and outdated public transit planning. This paper aims to explore different machine learning methods to bring mass transit proposals and implementation plans to governing bodies faster. Options such as heavy rail and light rail are the focus of this paper.

A. Motivation

Scientists in Japan experimented with a slime mould simulating the Tokyo railway network using oat flakes and a single-celled mould called Physarum polycephalum [1]. The connections the mould generated mimicked the current Tokyo transit system, a system designed by thousands of engineers.

With the vast amount of geographical data that is publicly available, machine learning could be applied to these large data sets to generate optimized solutions for urban planners. An application that can give realistic solutions to rapid transit lines in any city is the focus of this experiment. As the need for better and more efficient transit grows around the world, it is imperative to make use of machine learning.

Leveraging machine learning tools to design efficient urban mass transit routes can be much more powerful than manual planning. Human-planned public transit tends to stick to pre-existing trends – whether that may be arterial roads, railroad lines, or geography. Machine learning algorithms rely on a higher dimensional feature set that is hard for humans to perceive, resulting in far more diverse and unique ideas.

B. Related Works

There have been numerous works of literature related to transit route planning. An example is "Improving Transit Accessibility with Machine Learning" by Google AI Blog, where the prediction of transit wait times utilizes real-time traffic forecasts and data on bus routes and stops [2]. The model is split into a sequence of timeline units per bus, with its wait duration independently forecasted [2]. Additionally, a guide on transit station site selection to address good practices in the model city of Eau Claire, Wisconsin, USA helps outline the importance of amenities and points of interest around a station [3]. This guide supports a lot of the planning of Eau Claire's transit system through the analysis of other transit systems across the United States [3].

C. Problem Definition

With the lack of automation and the use of artificial intelligence in transit planning, the solution is to build a model to design an optimized subway system for any city in the world. This provides a tool for urban and city planners to make informed decisions when designing and implementing a new subway system. It will also give members of the public an opportunity to compare any current subway system with the most efficient and cost-effective version. The output of the model is a system to improve the subway infrastructure of cities, so a greater percentage of the population can be served through transit. To ensure the designed system is convenient for passengers, interchanges must be implemented to facilitate a smooth and easy transition across subway lines.

Overall, the results of this experiment are measured by an arbitrary metric that represents the number of people on average each station serves. In particular, the City of Toronto is used as the pilot city for this experiment. The system generated from the model is compared to the current Toronto Transit Commission (TTC) system to determine the effectiveness of the model.

II. Methodology

The data sets are created through OpenStreetMap API requests. In particular, Points of Interest (POI) like community centres, libraries, and restaurants are found for a given city and saved to individual CSV files to be used by the model. The attributes are filtered to contain the latitude, longitude, and name of each POI, with all other attributes being dropped.

A scoring function assigns coordinates a measure of usefulness. The scoring function is based on a Gaussian Distribution:

$$f(d) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{d}{\sigma})^2}$$

where d is the Haversine distance between a point and a POI, and σ is the standard deviation. The function $f(d)$ generates high scores for points close to the POI and generates scores approaching zero as the distance d increases. This scoring function is applied to a grid search function to give scores to each grid point throughout a city when compared with the locations of POIs. Through hyperparameter tuning, a standard deviation of $\sigma = 0.9$ is used for this experiment.

To determine the high-scoring locations throughout an urban centre, a grid search is performed. This involves dividing the city into an N x N grid. Each point in the grid is scored based on the distance from itself and the locations of each POI using the following equation:

$$s_{i,j} = \sum_{k=0}^{N} f(d)$$

where $s_{i,j}$ is the score at (i, j), N is the number of POIs, and $f(d)$ is the scoring function. The plot of generated grid points for the City of Toronto is shown in Fig. 1.

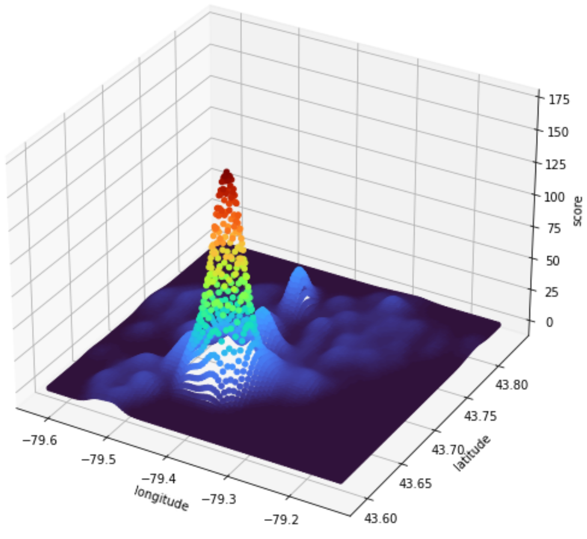

Fig. 1. Scored grid points from the City of Toronto

In any mass urban transportation system, lines in a system should be dispersed throughout the urban area, with an additional level of concentration in the densest part (often the downtown core). Dispersion of transit lines is achieved through clustering. A modified k-means algorithm is used, that considers the scores of each point x_i as a weight w_i in $\mathbb{R}^2$ space [4]:

$$\text{centroid} = \sum_{i=0}^{N*N} \min_{\mu_j \in D}(w_i * ||x_i - \mu_j||^2)$$

where μ_j is the mean of the samples in one of the clusters in the set D. The weighted k-means algorithm converges on a solution where cluster centres are distributed densely in high-score areas, and more sparsely distributed in low-score areas. This is synonymous with having more stations and lines in downtown areas, and fewer stations and lines in the suburbs. The clusters represent the neighbourhoods each line will go through. This implies the number of lines in the system is equal to the number of clusters. The clusters are numbered at random from 0 to C - 1, where C is the number of clusters. Sorting these clusters by average score is important for generating interchanges. After sorting, cluster 0 has points that make up the highest average score, cluster 1 has the next highest, etc. The clusters for $C = 6$ in the City of Toronto are shown in Fig. 2.

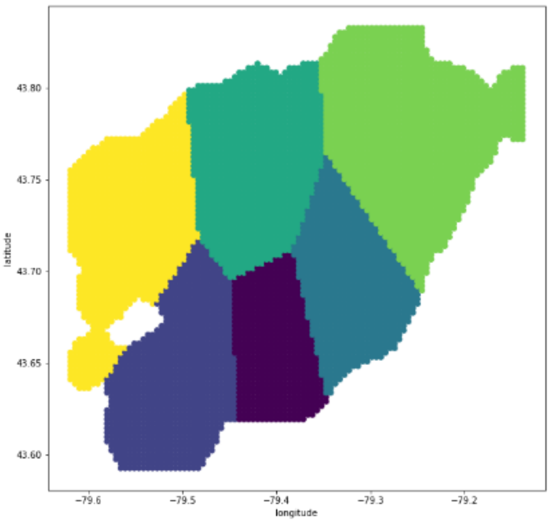

Fig. 2. Clustered grid points from the City of Toronto

After clustering, fitting transit lines is the next step in developing a transit system. A plausible approach is to perform weighted polynomial linear regression independently on each cluster. Weighted polynomial regression is chosen due to the use of weighted grid points. An equation for the matrix calculation of weighted polynomial regression is shown below where m is the degree of the polynomial and n is the number of known data points. It is determined by creating a residual function, summing the squares of the residual, forming a

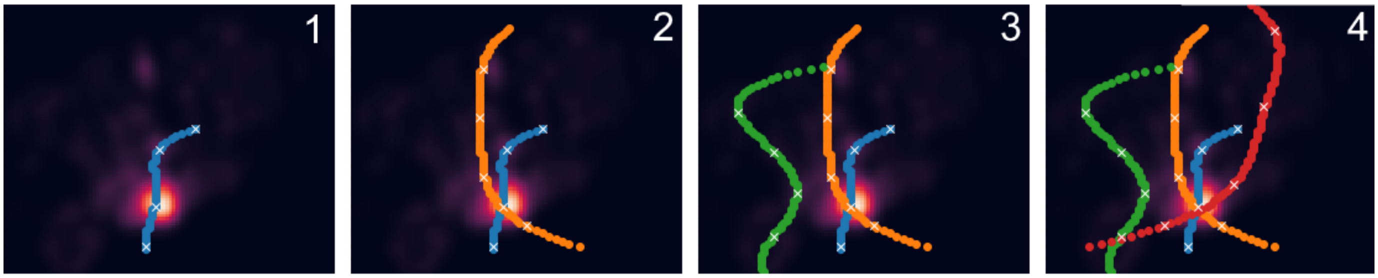

Fig. 3. Line Progression with C = 4 in the City of Toronto

parabola, and determining the coefficients β_i of the parabola [5]:

$$\begin{bmatrix} \sum_{i=0}^{n} w_i & \cdots & \sum_{i=0}^{n} w_i x_i^m \\ \sum_{i=0}^{n} w_i x_i & \cdots & \sum_{i=0}^{n} w_i x_i^{m+1} \\ \vdots & \ddots & \vdots \\ \sum_{i=0}^{n} w_i x_i^m & \cdots & \sum_{i=0}^{n} w_i x_i^{2m} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} w_i y_i \\ \sum_{i=0}^{n} w_i x_i y_i \\ \vdots \\ \sum_{i=0}^{n} w_i x_i^m * y_i \end{bmatrix}$$

Polynomial regression is used instead of linear regression as it allows for better coverage of the grid points compared to a straight line. A polynomial of degree 7 is used for this experiment to maximize coverage without overfitting. Additionally, the relationship between latitude/longitude and score is not linear in each cluster. The points of the regression lines are snapped onto the existing grid to simplify work with the data in later steps.

Interchanges are an important feature in designing mass public transit systems. By performing polynomial regression on each cluster separately, some unintended consequences are introduced. The resulting system of lines does not overlap as each line is designed to optimize for its constituent grid points (i.e. the set of constituents for each cluster is disjoint). This implies there are no interchanges in the aforementioned system. An iterative approach to generating lines is considered instead. Each line is generated successively from previous lines and utilizes information from previously generated lines. Line generation starts from cluster 0 - the cluster with the highest average score. Potential interchanges or interchange candidates are identified on each line generation through

$$I_l = \underset{i \in L_l}{\operatorname{argmax}}(s_i)$$

where I_l is the interchange candidate for line l, L_l is the set of points in line l, and s_i is the score for the i^{th} point on the line. The score of the grid point associated with I_l is multiplied by a factor α such that the new score s is more favourable to the regression of the next line. The interchange candidates also decay after each iteration. This encourages the generation of future lines to consider other interchanges and not just the first interchange. After each iteration, the score of each interchange candidate is reduced by a decay factor ρ, such that the new score for the i^{th} interchange candidate is $s_i(1-\rho)$. This results in the score of the i^{th} interchange candidate after k iterations to be

$$s_i^{(k)} = \alpha s_i (1-\rho)^k$$

The modified grid point $V_i^{(k)}$ associated with each score at each iteration $s_i^{(k)}$ is added to the set of points for all future lines, i.e.

$$\forall k \in \{0, ..., C-1\}, L_{l+k} = L_{l+k} \cup \{V_0^{(k)}, ..., V_i^{(k)}\}$$

An example of this iterative line generation process for the City of Toronto is shown in Fig. 3. The interchange candidates are represented in the figures by a white x.

To generate stations, the N_s highest-scoring locations on a line are found, and the distance between them is calculated. A station is chosen if the distance between two locations is greater than the distance parameter. The N_s value is a parameter that can be tuned, resulting in a line having at most N_s stations. This method does not guarantee that there will be exactly N_s stations per line as it may not be a large enough distance apart from other high-scoring locations.

III. Results

As stated in the problem definition an arbitrary metric that represents the number of people served on average per station is used to compare the current transit system with the one generated by the model.

The metric is based on iterating on all stations of each line and averaging all the scores of the grid points that lie R radius from the station.

$$s_{\text{station}} = \frac{\sum_{s \in S} s}{||S||}, S = \{s_{i,j} \mid i, j \in \mathbb{Z}, \sqrt{i^2 + j^2} \leq R\}$$

The score of the system is simply the mean score of all the station scores.

Fig. 4 shows the system generated by the model for the City of Toronto, consisting of four lines.

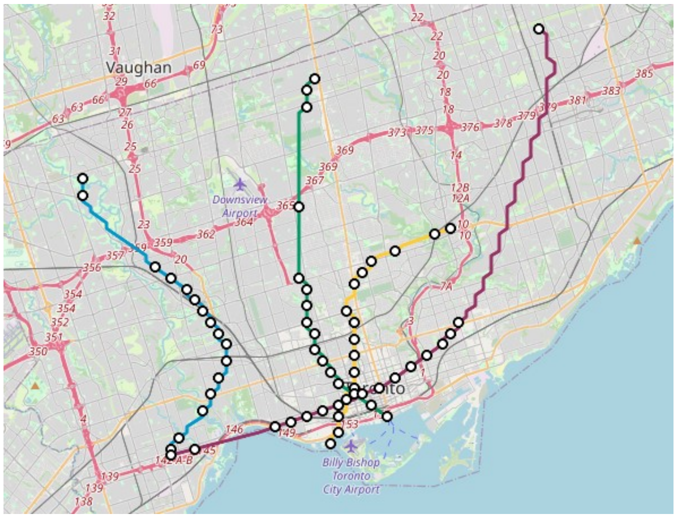

Fig. 4. Generated System for the City of Toronto, $C = 4$

To ensure an accurate comparison between the model and the current TTC system the results were calculated with $C = 4$ as the TTC consists of 4 subway lines. The model uses $R = 1$, representing the distance customers would walk to use a transit station. The current TTC system with 65 stations scored 44.15 using the system scoring metric. Table I shows the system scores generated by the model for Toronto, where N_s represents the number of stations per line. The score differs based on how many stations are chosen per line. A simplification is made where each line is chosen to have the same number of stations. The exception is cases where the line can not support N_s stations due to its length, resulting in fewer stations than N_s. For $N_s \leq 15$, the transit system generated from this experiment outperforms the existing TTC system.

TABLE I
Results for generated transit systems for the City of Toronto, $C = 4$

N_s	Score	N_s	Score	N_s	Score
8	69.75	12	53.40	16	42.83
9	64.70	13	50.30	17	40.91
10	60.46	14	47.64	18	39.13
11	56.68	15	45.07	19	37.48

A transit system for Berlin with $C = 6$ is generated as shown in Fig. 5 to demonstrate the viability of a mass transit system in a different city on a different continent.

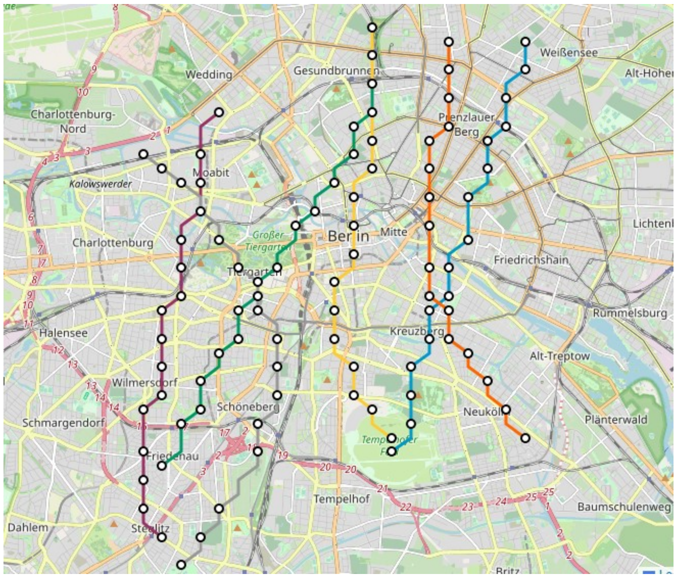

Fig. 5. Generated System for the City of Berlin, $C = 6$

IV. Conclusion

An optimized subway system model is created for any city by combining a K-means clustering and polynomial regression algorithm to be able to determine transit corridor placement throughout a city. Polynomial regression is conducted on each cluster, resulting in distinct transit lines. These algorithms are used on a set of scored grid points generated using a grid search function. A heuristic is used to connect the transit lines together. Stations are placed in the order of highest score to lowest score while maintaining a minimum distance from one another. The results show that the model for this experiment outperforms the current TTC system for $N_s \leq 15$.

For future improvements, a larger data set can be used to ensure that all high-traffic areas are considered. This also ensures that smaller cities, such as Kingston, have more accurate lines. An improvement to improve the defined bounds of a city can also be considered. This ensures that lines and stations strictly appear over areas where transit can be built. The bounds should also take into account locations of water.

References

[1] L. S. News Science, “Slime Mold Grows Network Just Like Tokyo Rail System,” Wired, Jan. 22, 2010. https://www.wired.com/2010/01/slime-mold-grows-network-just-like-tokyo-rail-system/

[2] A. Fabrikant, “Predicting Bus Delays with Machine Learning,” Google Research, Jun. 27, 2019. https://ai.googleblog.com/2019/06/predicting-bus-delays-with-machine.html

[3] City of Eau Claire, “Transit Centre Site Selection Study,” City of Eau Claire, Eau Claire, Wisconsin, United States, May 2016. Available: https://wisconsindot.gov/Documents/doing-bus/local-gov/astnce-pgms/transit/ec-site.pdf

[4] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830, 2011, Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[5] A. Que, “Mathematics of Polynomial Regression,” polynomialregression.drque.net, 2021. http://polynomialregression.drque.net/math.html