Assessing Minnesota's Public Transportation

Lucas Major

Abstract

This project builds a model to measure the Met Council 7-county regions access to public transit. The model calculates a score for each parcel in the region based on its relationship to public transit stops. Tax parcel datasets and a transit stop dataset are necessary to accomplish this. To calculate scores, two independent methods will be implemented and compared: A weighted linear combination model, and a regression model. These methods will take into account the distance from a parcel to transit stops, the number of transit stops within a distance of a parcel, the type of the transit stop, and the frequency of the route. To build the regression model, current Zillow transit scores [1] of properties in Minneapolis are used as the training data. These Zillow values are also used to assess accuracy of the weighted linear combination. Both methods are measured on accuracy, and it is found that a polynomial model gives the highest accuracy, as well as having the lowest computational time. However, it is believed the accuracy of the model decreases in the regions outside of the immediate Twin Cities. The weighted linear combination has a lower measured accuracy overall, and higher computational time, but it is better equipped to handle all areas of the dataset.

Problem Statement

Public transportation is important in the foundation to creating more affordable, healthy, and livable cities. A transit score is a measure of how accessible public transit is for a given place. Currently, websites such as Zillow and Redfin assign these transit scores to properties in Minneapolis and Saint Paul. However, they do not have scores available for properties in surrounding cities, even though many of these cities have transit access. As a result, the goal of this project is to create a model that can assess the transit accessibility of all parcels within the Met Council 7 county region.The scores will be assigned based on the following attributes: the property's distance from transit stops, the number of transit stops within a distance, and the frequency of the route, and the type of transit (bus/light rail). To complete this, shapefiles of each parcel in the region will be needed. Additionally, the location of every transit stop is needed, along with the attributes type and frequency.

Creating a weighted linear combination is one method of modeling the transit scores of the region. This type of method works well when more than one attribute must be taken into consideration in a model [2]. Then, each attribute is assigned a weight based on its importance. In this project, the model weights and other parameters are assigned through repeated runs of the model. For each run, the parameters are varied, and then the results are compared to the reference data which in this case is the Zillow transit scores. This process is repeated until the highest accuracy is achieved.

Secondly, a regression analysis [3] is another method that can calculate transit scores. Regression is the process of estimating the relationships between a dependent variable and a set of independent variables. In this case, the dependent variable is the Zillow transit score, and the

independent variables are the attributes listed above. These models require training data, and in this project I collected the Zillow transit score of 200 random parcels in Minneapolis.

Table 1. Data Requirement

#	Requirement	Defined As	(Spatial) Data	Attribute Data	Dataset
1	Parcel Data	All Parcels in 7 county region	Polygon	Location	Mn Geospatial Commons
2	Transit stops	All transit stops in 7 county region	Point	Location, Type, Frequency	Transit Stops
3	High Frequency Route	All routes that run every 15 minutes or less	Line	Location	Mn Geospatial Commons
4	Reference Data	Zillow transit scores	none	Transit Score	Zillow

Input Data

The first dataset is a shapefile containing a point for every transit stop in Minnesota. Relevant features of the dataset are the location and active status. Later on, features indicating the type and frequency of each stop will be added. The frequency will be added using the High Frequency Network Dataset, a shapefile that has a line for every route that has service every 15 minutes or less. The third dataset is a polygon shapefile of every parcel in the counties listed in the table below. The parcels for the remaining two counties, Ramsey and Dakota, had to be gathered from a different file, so they are listed separately. Finally, I manually collected Zillow transit scores on 200 random parcels in Minneapolis to use for my accuracy assessment and model training.

Table 2. Input Data

#	Title	Purpose in Analysis	Link to Source
1	Transit Stops	Shapefile containing all Minnesota public transit stops as points	Mn Geospatial Commons
2	High frequency network	Shapefile containing the high frequency transit network	Mn Geospatial Commons
3	County Parcels - Hennepin, Scott, Carver, Anoka, Washington	Parcels to calculate transit score for	Mn Geospatial Commons

4	County Parcels - Ramsey	Parcels to calculate transit score for	Mn Geospatial Commons
5	County Parcels - Dakota	Parcels to calculate transit score for	Mn Geospatial Commons
6	Zillow transit scores	Used for accuracy assessment in weighted combination, training data in regression models	Zillow

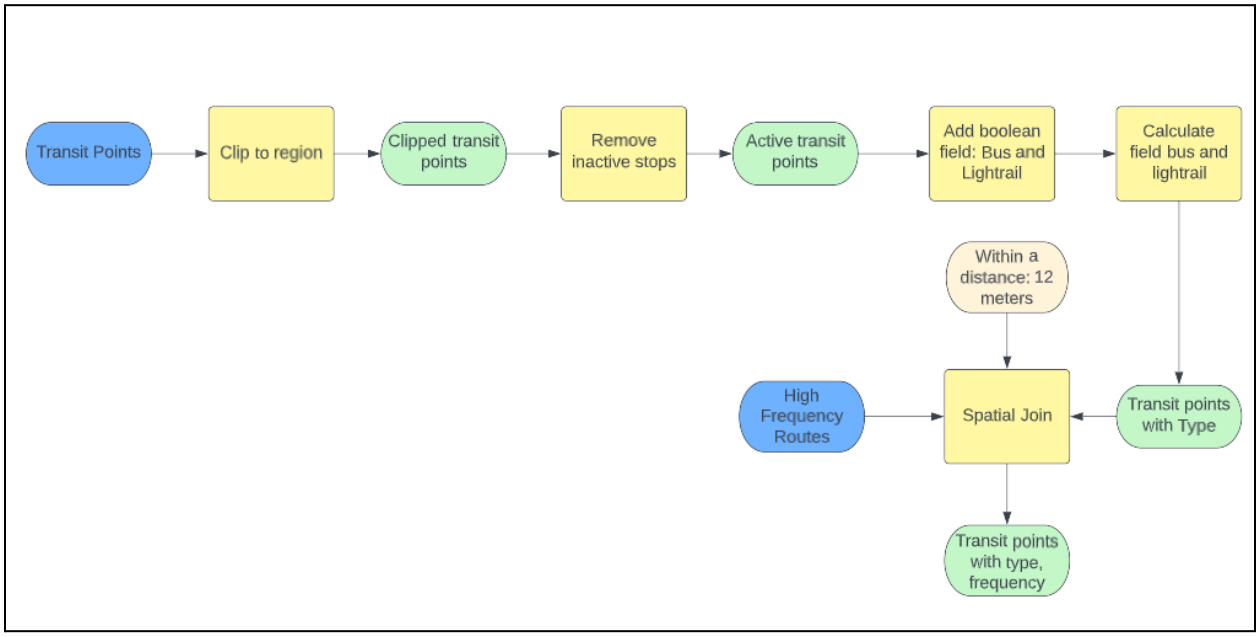
Methods

Data Cleaning/Organizing

The parcel, transit stop, and high frequency network datasets were all gathered via API calls to the Minnesota Geospatial Commons using arcpy. Many of the transit stops had to be removed from the dataset as they were not currently active. For each transit point, an attribute indicating the type was added. This was done by adding a boolean field to the data table for both bus and light rail. Then, all light rail stops were manually selected and given a value of "1" for the light rail field, and a "0" for the bus field. Every other point was then given a "1" for the bus field.

Additionally, a boolean field for high frequency stops was added. The high frequency network dataset is a shapefile containing lines for all routes that run every 15 minutes or less. This dataset was spatially joined with the transit stop dataset with a distance threshold of 12 meters. Then, the high frequency field was labeled as a "1" for stops that were within 12 meters of the high frequency line, and a "0" otherwise. This process is shown in *Figure 1* below.

Figure 1: Data Flow Diagram



Weighted Linear Combination

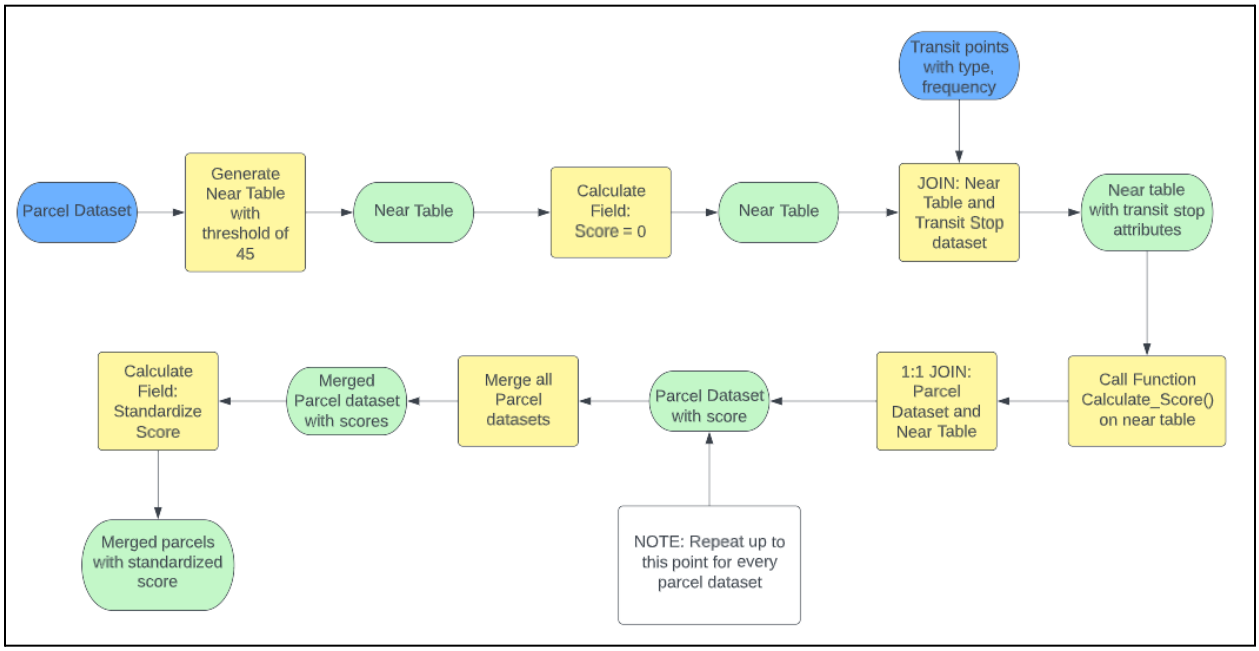
The first method implemented to calculate a score was a weighted linear combination. First, the arcpy function *GenerateNearTable()* was called on the parcel datasets and transit stops, which creates a table containing a row for each parcel, transit stop, and the distance between them. The distance threshold used was 1 kilometer, and the closest 45 transit stops were recorded for each parcel. Then, the transit stops table was joined with the near table. This way, for each parcel, there are up to 45 rows for each transit stop it is within 1 kilometer of. Each of these rows has the attribute distance to, transit stop type, and transit stop frequency. With this information, the scores can be calculated. Below, the weighted linear combination is shown.

$$Score = \sum_{i=1}^{n} (\omega_L(x_{iL}) + \omega_B(x_{iB}) + \omega_F(x_{iF})) * d_i$$

$$\omega_L = 0.5 \quad \omega_B = 0.2 \quad \omega_F = 0.3 \quad n = 45$$

Each x in this equation represents a boolean value for light rail, bus, and frequency, as labeled by the subscripts. These parameters have attached weights determined through model tuning and sensitivity analysis, which will be discussed later. Additionally, for each transit stop, a distance penalty is multiplied. In this case, if the transit stop was between 200 and 400 meters away, the distance penalty is 0.7. If the transit stop is farther away than that, the distance penalty lowers to 0.5. This value was surprisingly gradual, and it was also determined through model tuning. The model then sums each calculated value for the closest *n* parcels to give a final score for each parcel. In general, this model works by giving parcels higher scores for having more transit stops within a 1 kilometer radius, as well as having more valuable transit stops, ie. light rail stops, high frequency stops, and close distance stops. The model calculation was done in python through a written function that calculates scores based on the near tables. The resulting scores were standardized from 0 to 1, in order to compare results with the regression methods. The process is shown below in *Figure 2*.

Figure 2: Data Flow Diagram



Regression

The second method used to calculate transit scores was both linear and polynomial regression. Linear regression was first implemented by training a model using 200 randomly selected Zillow scores from Minneapolis. The attributes used were property sale value, distance to the nearest transit stop (distance), the type and frequency of that stop, and the number of stops within a 1 kilometer radius (number). When creating the linear model, the variables light rail, frequency, and number were all statistically significant at the 0.05 significance level. Then, the insignificant variables (sale value, bus, light rail) were removed one at a time so as to not encounter multicollinearity issues.

Table 3: Models

Method	Linear	Polynomial
Initial model	Score ~ Distance + Num + Bus + LightRail + HighFreq + Sale_Value	Score ~ poly(Num, *x*) + poly(Distance, *x*) + Num:Distance + Bus + Lightrail + HighFreq
Final model	Score ~ Distance + Num + LightRail + HighFreq	Score ~ poly(Num, 3) + Distance + Lightrail + HighFreq

Then, a polynomial regression model was constructed. A polynomial model was chosen because of the relationships in the graph in *Figure 3*, which is a plot of predicted vs actual values of the linear model. The polynomial model took in the same parameters, but the variable number was raised to the power of 3, and distance was kept standard. These polynomial values were determined through model tuning. The variables bus, light rail, and frequency were not raised to

a polynomial as they are boolean values. Again, the insignificant variables were removed one at a time from the model. In this case, bus, sale value, and the interaction term between Num and Distance were the insignificant variables. The predicted vs actual values are displayed in *Figure 4*, which shows an improvement over *Figure 3*. The initial and final models for both linear and polynomial regression are displayed in *Table 3*.

Figure 3: Linear prediction

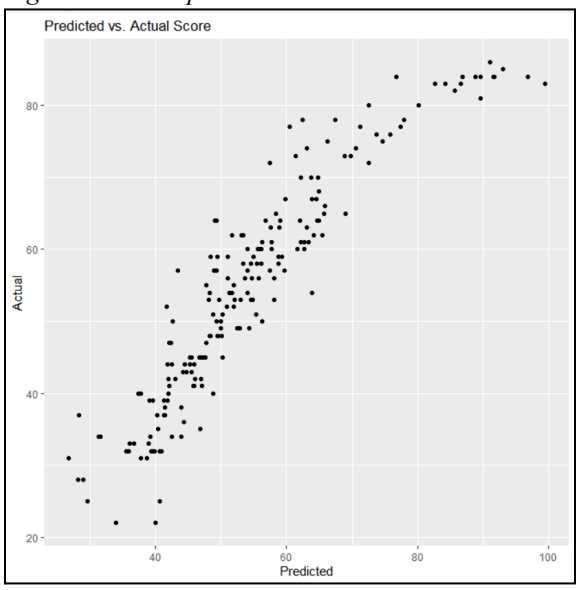
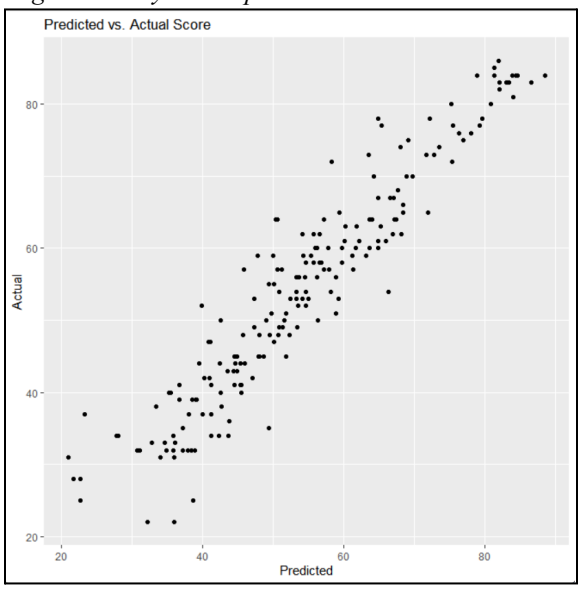
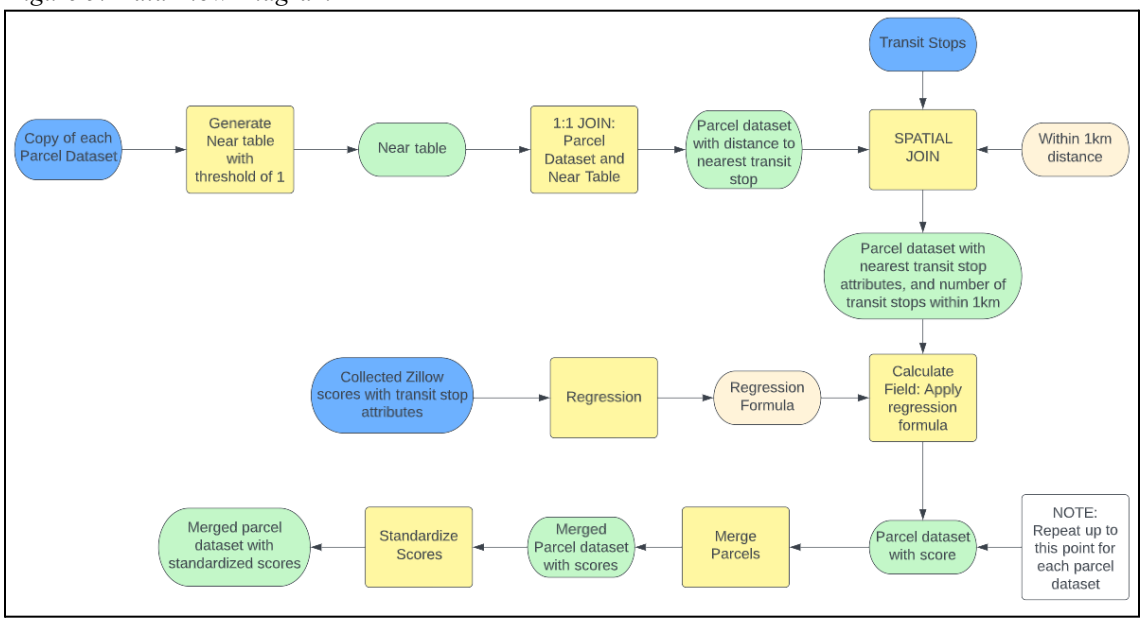


Figure 4: Polynomial prediction



After building the final model, the rest of the process happens. First, a near table is generated for each parcel to find the single nearest transit stop. Then, the near table is joined with the parcel dataset to have the distance to the closest transit stop for each parcel. Then, a spatial join with the transit stop dataset is completed with a distance threshold of 1 km to have the total number of transit stops within 1 km for each parcel, and the nearest transit stops attributes. Finally, the regression formula can be applied to calculate a score for each parcel. Again, the parcels are merged and the scores are standardized. The data flow diagram for this method is shown in *Figure 5*.

Figure 5: Data Flow Diagram



Accuracy Assessment

An accuracy assessment was performed on the weighted linear combination, linear regression, and polynomial regression methods. For the weighted linear combination, 50 random parcels were sampled in Minneapolis and the corresponding Zillow transit scores were used as ground truth data. These Zillow scores were compared to the scores the model created and then an R^2 and normalized RMSE calculation was performed. This accuracy assessment is important in model tuning and sensitivity analysis, as repeatedly calculating accuracy by changing the model parameters helps to create a more accurate model. Ideally, more than 50 sample points would have been used, but since the computational cost of this model is so large, the number of sample points had to be kept low in order to reasonably re-run the model.

For the regression model, the R^2 and normalized RMSE are calculated based on the model's performance on the collected 200 Zillow scores.

Results

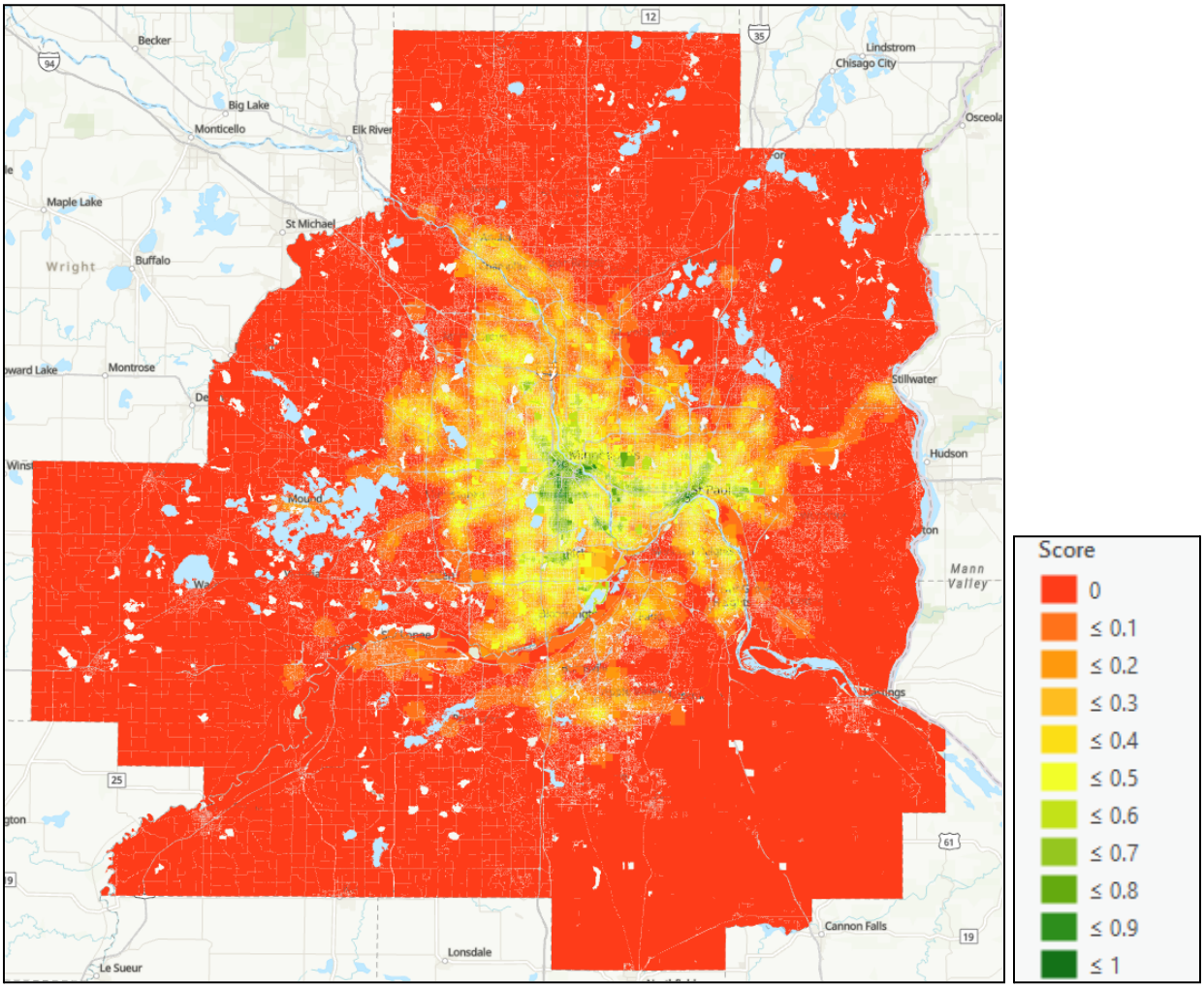
The measured accuracy for the methods is shown below.

Table 4: Accuracy

Method	Weighted Linear combination	Linear model	Polynomial model
R^2	0.75	0.86	0.90
RMSE	0.21	0.09	0.08

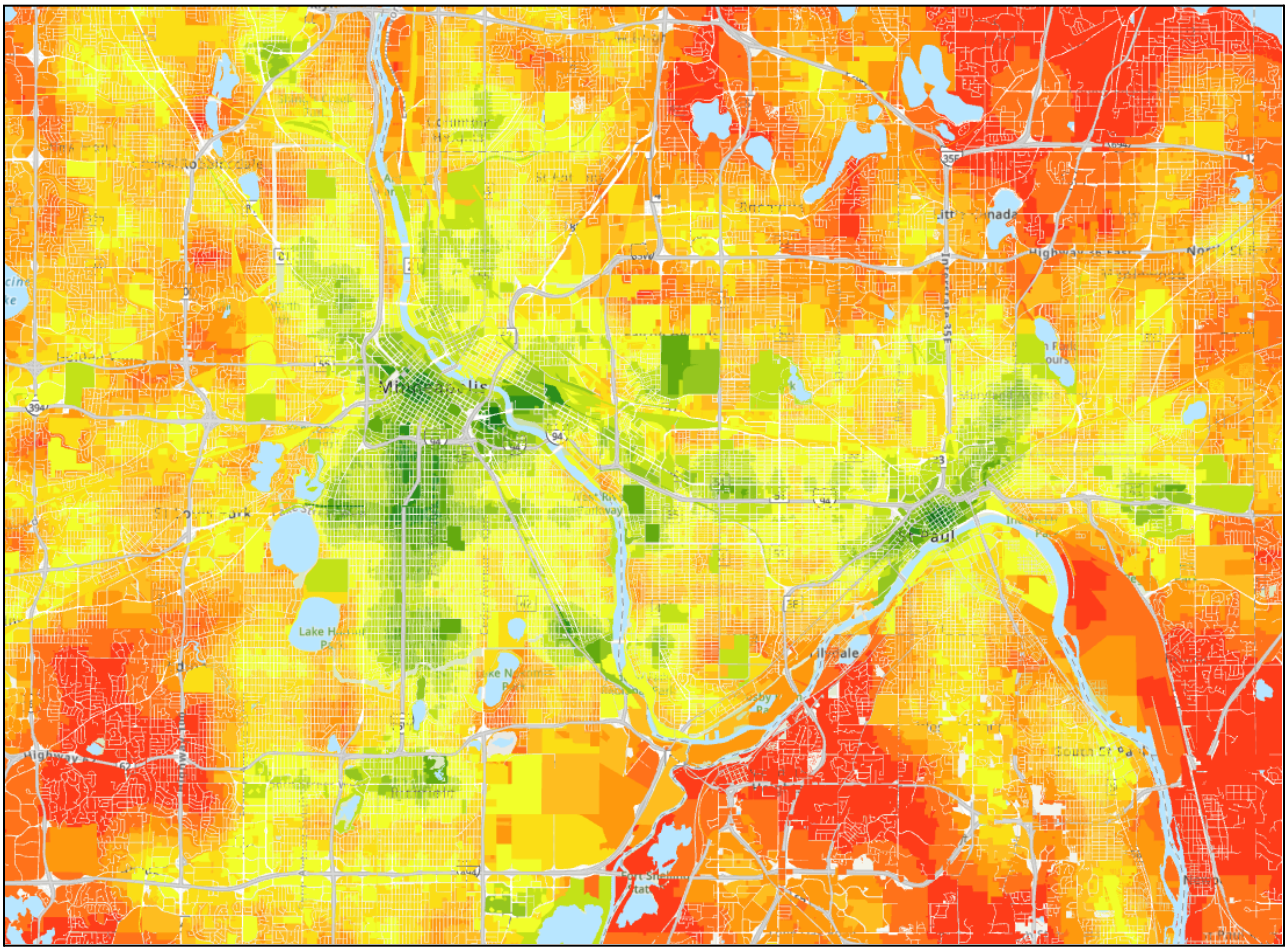
The results for the weighted linear combination are shown below.

Figure 5: Weighted linear combination



The results of the whole 7-county region make logical sense upon initial viewing. Much of the region is red, as much of the region has no transit access. However, pockets of color can be seen where bus routes travel out of the twin cities region. Another notable feature is the city of Bloomington is estimated to have decent transit access almost everywhere, even though properties in it don't have transit scores on Zillow.

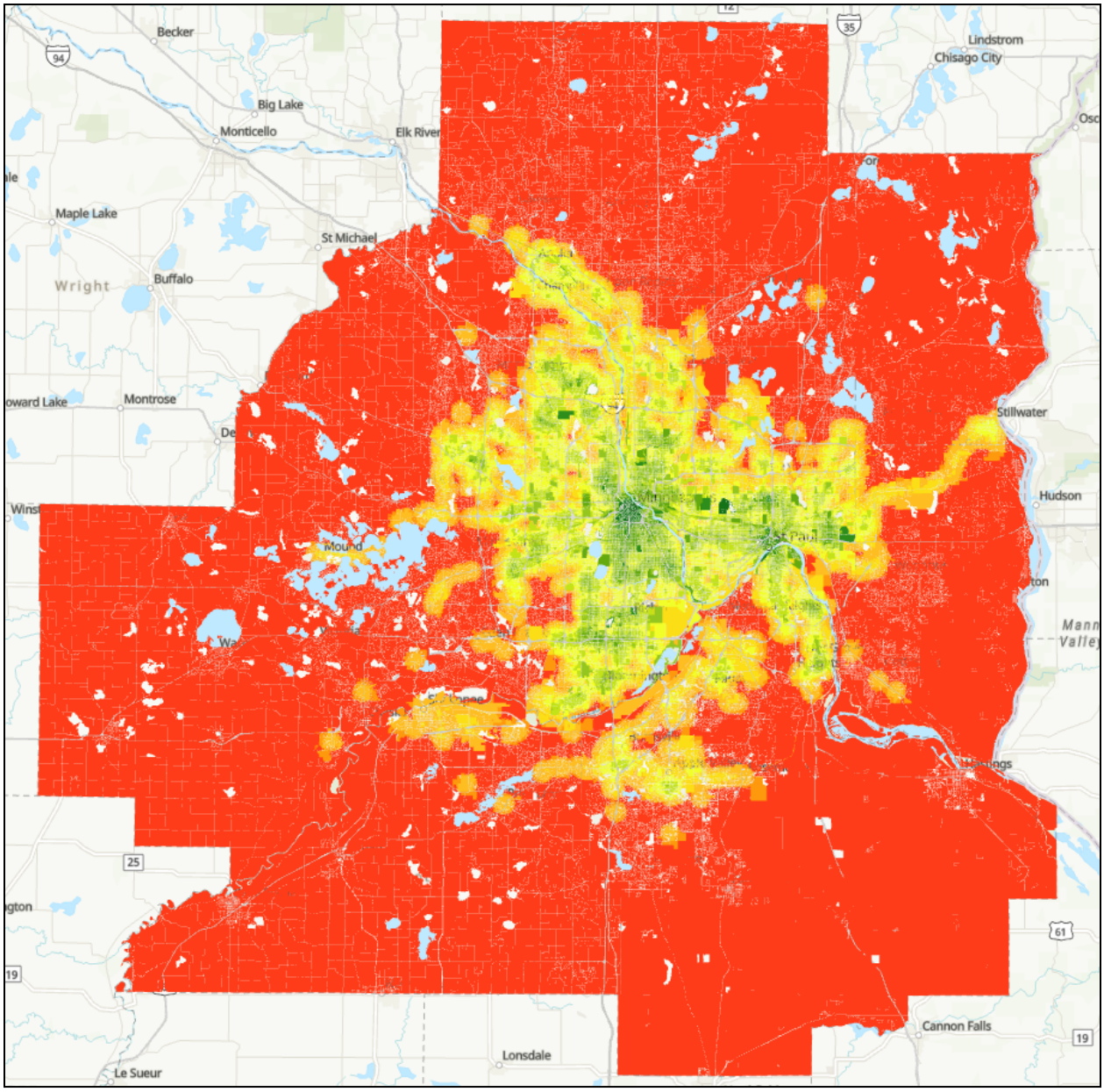
Figure 6: Zoomed in weighted linear combination



Upon a closer look, it can be seen that close to downtown Minneapolis and Saint Paul have good transit access, with much of south Minneapolis having transit accessibility as well. A notable phenomenon is larger parcels generally have higher scores, and this is because the distance from a parcel to a transit stop is measured from the edge of the parcel, so large parcels have a strong chance of having many parcels within their 1 km radius.

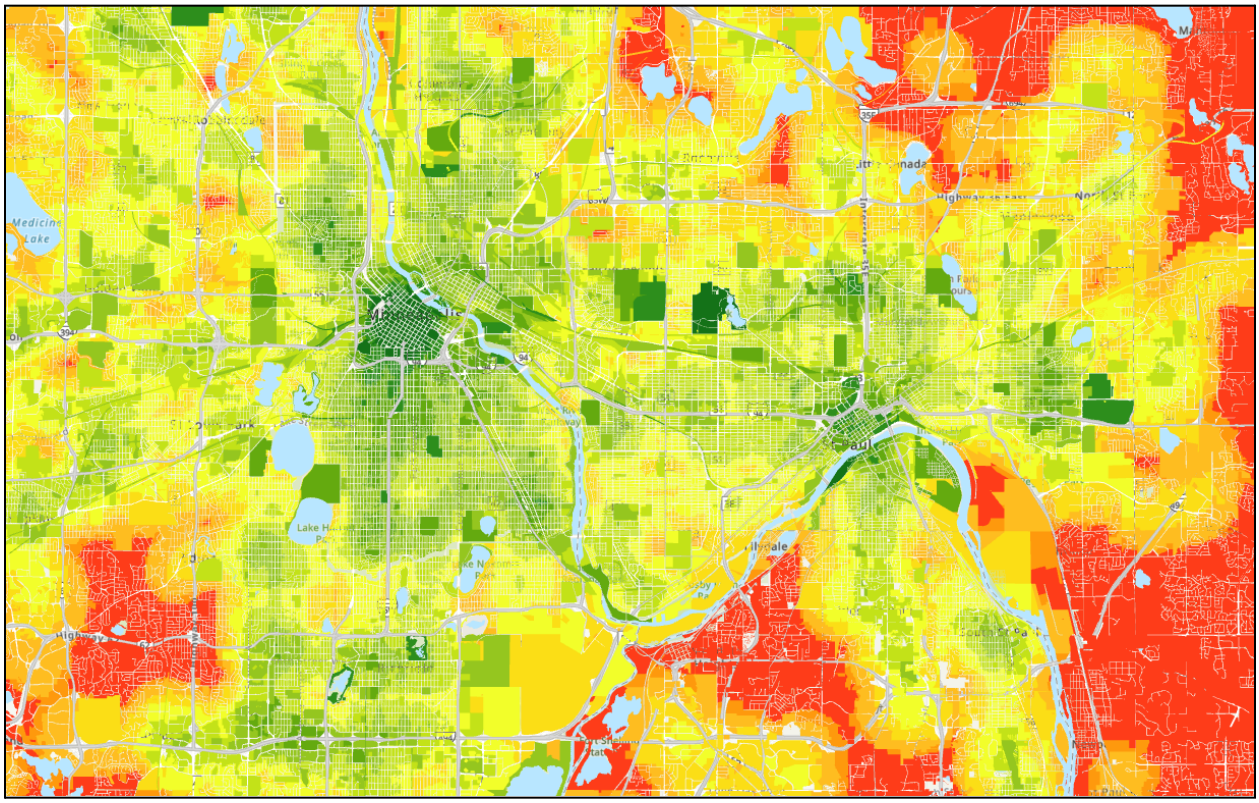
Below, the results of the polynomial regression model are shown. The polynomial result is shown over the linear result as the polynomial model had higher accuracy.

Figure 7: Polynomial model



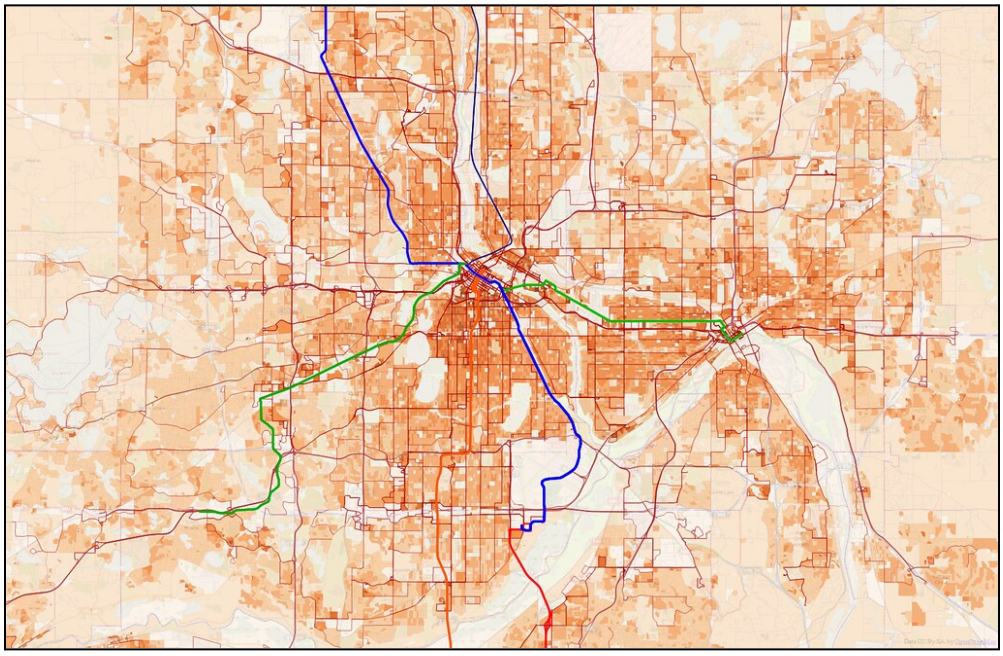
The polynomial model results appear similar in many ways, however there is a noticeable sharp drop-off between areas of no transit vs areas with transit compared to the weighted linear combination, which is more gradual.

Figure 8: Zoomed in polynomial model



When zoomed in, we can still see that the areas with the highest transit access are in Minneapolis and Saint Paul. Generally, this model tends to give higher scores all around compared to the weighted linear combination.

Figure 9: Population density map



Most importantly, with these results, we can identify weak points in transit. For example, when viewing a population density map as shown in *Figure 9* we can see south west Saint Paul has a relatively similar population density to south Minneapolis just across the river, yet the transit accessibility is much more limited. Additionally, there is a large region in Edina with no transit access at all, as shown in *Figure 8*. While the population density is relatively low for this area, it is similar to the region around Plymouth, yet Plymouth has moderate access to transit.

On the other hand, there is a strip of parcels in Bloomington that has a low population density. However, this same region has high transit scores as shown in *Figure 8*. This could possibly be a region where the transit access over represents its actual need for transit, although this particular region is located along 494 where many office buildings are located. As a result, while not many people live in this region, the demand to travel to this region for work is high, so that could be a reason why the transit access is much higher than the surrounding area.

Results Verification

When viewing both sets of results they make logical sense. The areas close to both downtowns have the highest transit accessibility, while much of the region does not have any at all. Further, the color that spreads out of the cities follows bus routes. Additionally, to ensure the score calculations were correct, I chose 10 random parcels and manually calculated their scores for both the weighted combination and regression model.

Sensitivity Analysis

In the case of the weighted linear combination, the model parameters were determined through repeated runs of varying the model parameters and comparing the result to the Zillow transit scores. As mentioned earlier, 50 random parcels were sampled and their Zillow scores were compared to the model's predicted scores. To perform this analysis, every model parameter was adjusted. This includes the weights, the search radius, the distance penalty, and n, the number of transit stops to look at for each parcel. The procedure to run this analysis was to pick one parameter, and then re-run the model varying that parameter over a large domain, then calculate accuracy.

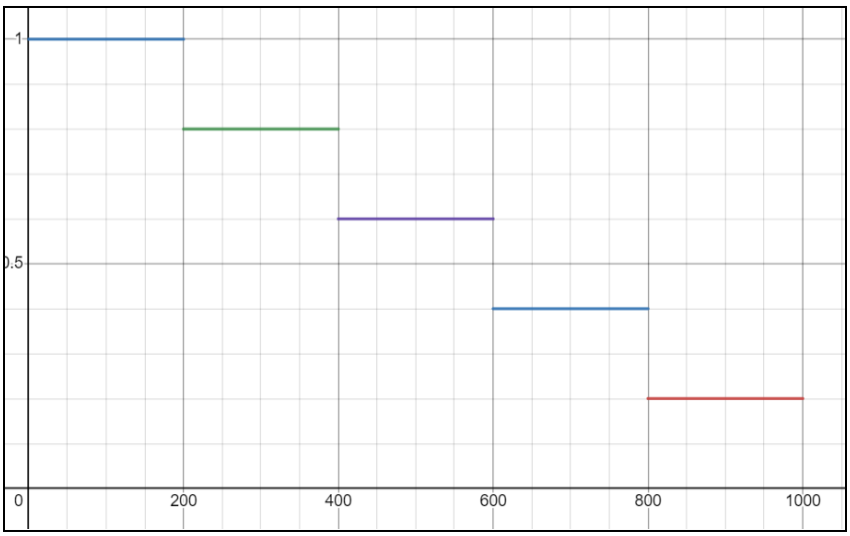
The optimal weighting scheme makes sense, as light rails and high frequency stops were found to have more value than regular bus stops. The model accuracy would decrease when the bus weight was higher than frequency and light rail weights, which again makes logical sense.

The search radius also greatly affected the results. From this analysis, both increasing and decreasing the radius from 1km significantly lowered the model accuracy.

Most of these parameters involved just changing a number, however modifying the distance penalty was unique. The distance penalty acts like a decay function, where as the distance increases, the value of the distance penalty also decreases. An example of what this could look like is shown below in *Figure 10*, where the value of d is the y-axis, and distance is the x-axis in meters. In order to tune this parameter, both the cutoff ranges and the associated d values had to be adjusted. Initially, the distance decay function looked similar to the one shown in *Figure 10*,

however it was found through this analysis that a very gradual decrease gave the best accuracy. The model was also tested without the distance penalty, but the accuracy was much lower, with an R^2 of around .60.

Figure 10: Example distance decay function



Additionally, it was found that adjusting the *n* value can have a very large effect on the model accuracy. This makes sense, as the most significant variable in the regression model was the number of transit stops within the search radius. Increasing *n* increases the accuracy, however it also greatly increases the computational time.

For regression, the insignificant variables were removed, and when creating the polynomial model, all combinations of polynomials 1-5 were tested between Number and Distance. Then, an ANOVA table was used to test if the R^2 value of the polynomial model was statistically higher than the linear model, which it was.

Limitations

As mentioned in the results, the weighted combination model has a more gradual drop-off, while the regression model has a very sharp one. This phenomenon essentially boils down to the regression model lacking training data from these areas. The model was only trained on parcels from Minneapolis, as that is the only area Zillow has scores. Since Minneapolis is generally covered very well with good transit access, most of the training data had parcels with a high number of transit stops within 1 km, and as a result high scores. Therefore, the model performs very well at estimating scores for regions with high transit. However, since it was not trained on many areas with low transit, it tends to overestimate scores for these areas, leading to the sharp drop off in these outskirt regions.

Discussion and Conclusion

Both models were able to create a good visualization on the current status of Minnesota's public transit. For example, the entire city of Bloomington's public transit access was modeled, even though they do not have transit scores on Zillow. Results like this are important in gauging how well a population is being served, as well as a measure of where future developments are needed.

When strictly viewing accuracy, the polynomial regression model performs the best. Additionally, the computational time is lower, as the total runtime for the weighted combination model was around 3.5 hours to calculate a score for every parcel. However, one drawback to the regression model is the amount of training data. As mentioned above, it is believed the out of sample prediction for parcels on the outer edges of the twin cities region, mainly places in Scott, Carver, and Washington counties may not be as accurate, because the regression model was not trained on areas with this low of transit access. This is why despite the lower measured accuracy, the weighted linear combination model may be a better model for estimating a transit score. This is mainly due to how the score is calculated, as it sums individual scores between a given parcel and all of its surrounding transit stops with their unique attributes. As a result, it has good potential for higher accuracy as it accounts for more variables compared to the regression model.

For future analysis, in an effort to reduce computational cost, using a dataset that contains less polygons would be a good idea. In the current parcel dataset, there are over 1 million parcels with varying size. If instead a dataset of less polygons covered the region, the computational limitations could greatly decrease. Additionally, if all polygons were the same size, there wouldn't be issues with large parcels having disproportionately high scores. Another way to improve the project would be to create a measure of accessibility to each transit stop. For example, this project measures accessibility strictly based on a parcel's distance to a transit stop. Instead, accessibility could be measured by how easy it is to walk to the transit stop. Utilizing data like land cover and road networks could help create this type of model.

References

[1] Rosen, R. (2022, August 11). *Your home's transit score now available on zillow*. Zillow Group. Retrieved from https://www.zillowgroup.com/news/your-homes-transit-score-now-available-on-zillow-3/

[2] Malczewski, J. (2000). On the use of weighted linear combination method in GIS: Common and best practice approaches. *Transactions in GIS*, *4*(1), 5–22. https://doi.org/10.1111/1467-9671.00035

[3] Ali, P., & Younas, A. (2021, October 1). *Understanding and interpreting regression analysis*. Evidence-Based Nursing. Retrieved from https://ebn.bmj.com/content/24/4/116

Zillow. (n.d.). from https://www.zillow.com/homes/Minneapolis,-MN_rb/

Minnesota Geospatial Commons. (n.d.). from https://gisdata.mn.gov/