# A Statistical Analysis of the Effects of Hennepin County Community Works Projects on Housing Values

Simon Bilsky-Rollins
Junxiong Liu
Charlotte Mann

Laura Chihara, supervisor

Carleton College
June 6, 2016

# Executive Summary

- The Hennepin County government has enacted several Community Works projects around the county within the past 20 years. This study, a modification and extension of the efforts of the group from winter term 2016, aims to identify the impact these projects have had on the market value of parcels near them.
- We focused on the Midtown Greenway project area, specifically a random sample of 1,000 residential parcels located within two miles of the project.
- We used a hierarchical linear regression model that takes into account year-to-year change in estimated market value (EMV) and differences between neighborhoods, measured by census tracts. Our response variable is the log of EMV adjusted for inflation. Our explanatory variables are time (each year from 2002 to 2015), each parcel's individual attributes, and neighborhood attributes.
- We found that within one mile of the project area, an increase in distance away from the project area corresponds with an increase in EMV.
- Between one and two miles away from the project area, an increase in distance away from the project area corresponds with a decrease in EMV.
- While this model is certainly an improvement over the model from winter term, we still encountered statistical problems that we were unable to resolve.
- We only had time to analyze the Midtown Greenway project; however, there are several other Community Works projects on which a similar study could be conducted with few modifications.

# 1 Introduction

Since the early 2000s, the Hennepin County Community Works has enacted numerous projects geared at livening surrounding communities. These projects have sought to make transportation more efficient and/or pedestrian-friendly (Lowry Avenue, Midtown Greenway, SW Light Rail, etc.) and create shared public spaces for community members to enjoy (Fort Snelling, etc). Of these projects, seven are in progress, two have been completed, while three have not yet started or are of unclear status. With the amount of time and capital invested in these projects, measuring their impact is of high concern. Our study aims to investigate how these project areas were associated with positive growth or stabilization of the estimated market value (EMV) for properties in close proximity to the project areas.

Our analysis in Winter '16 used a multiple linear regression and focused on one completed project area, the Midtown Greenway Project. We investigated the property market value changes in the immediate vicinity of the project area using descriptive parcel data and modeled how these changes were correlated with a parcel's linear distance to the project area. We found that within 1.5 miles of Midtown Greenway, as the distance to the Greenway increases, the change of EMV (between 2006 and 2015) increases. This result was opposite of what we expected, and the statistical model we used was subject to many limitations, including correlation between the property values for parcels near each other and only modeling the difference in EMVs between two specific years (for details please see the Discussion part of our W'16 report).

To address these limitations, we will use a model with a hierarchical structure based on a more comprehensive dataset. The parcel's linear distance to the Midtown project area remains the primary focus in our analysis. In addition, we will examine the importance of other factors such as the difference between years and characteristics of the neighborhood in determining the impact of the project. We will conclude by highlighting some limitations of our approach and suggesting possible future analyses.

# 2 Data

Based on the spatial data that delineate the centered lines of each project, we used GIS, Excel and R to produce and format our data. Currently, our data includes Consumer Price Index (CPI) data from 2002 to 2015. For each parcel, our dataset includes original total EMV values and total EMV values adjusted for inflation (based on CPI) for each year from 2002 to 2015, "individual" parcel attributes that include the parcel's exact spatial location (x,y coordinates), use classification, building square footage, acreage, etc, in addition to "neighborhoods", classified by census tracts, attributes that include the name of the tract, diversity index, median household income, and median age.

# 3 Exploratory Data Analysis

We limited the scope of our preliminary analysis to the Midtown project area, as it was one of the two projects completely finished by the time of our analysis. We hoped to gain a more in-depth understanding about a single a project area and establish a solid statistical framework from which future analyses of the remaining project areas could move forward.

The Community Works project located in the Midtown project area is the Midtown Greenway. The Greenway is a 5.5-mile trail built in a former railway corridor, along which bikers and pedestrians can enjoy either a mostly car-free commute to work, or simply spending time outdoors isolated from vehicular traffic. The project was enacted in several stages. Phase 1 of the greenway was completed in 2000 (two years before the first available year of historical parcel data). This first phase was comprised of bicycle and pedestrian pathways from 31st St. and Chowen Ave. to 5th Ave. The second phase was completed in 2004, extending the paths from 5th Ave to Hiawatha Ave. Phase 3 expanded the Greenway past Hiawatha Ave. to the Mississippi River in 2006. Further improvements continue to be implemented, such as the opening of Martin Olav Sabo Bridge, which bypasses a busy intersection, and the opening of the Freewheel Midtown Bike Center, which offers bicycle repair and locker/shower facilities for Greenway commuters. All of these aspects of the Greenway could have contributed to its impact on the surrounding community.

While we fit final models on all years in the data set, we began exploratory analysis comparing property values from the first year in the data set, 2002, and the completion of the project, in 2006. Next, we compared property values in 2006 with property values in 2015, the latest year for which we have parcel data. The analysis described here is on the sample of 1,000 residential parcels within two miles of the Midtown project area that we used in modeling.

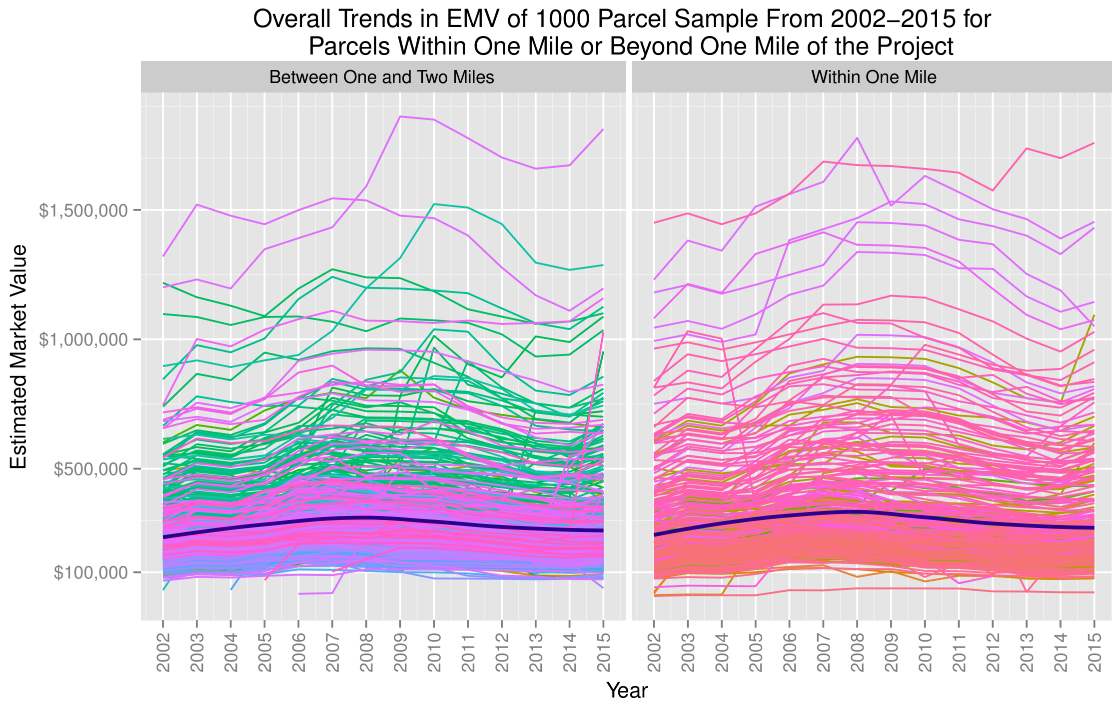


Figure 1. Yearly trends in EMV for the 1,000-parcel sample used in modeling, divided by parcels within one mile and between one and two miles from the Midtown project area.

Between 2002 and 2006, the sample of residential parcels within 2 miles of the Midtown project area saw an average change in EMV of 31.72%, and only and roughly 9% of parcels experienced a net decline. Between 2002 and 2006, the parcels within 1 mile of the Midtown project area saw an average change in EMV of 51.31% and median change was 36.54%, while parcels between 1 and 2 miles from the project area experienced an average change in EMV of 31.98% and median change of 29%. Between 2006 and 2015, 88% of parcels experienced a net decrease in EMV. On average, parcels within two miles of the project area saw an average change in EMV of -14.18%. Parcels within one mile of the project area experienced an average change in EMV of -17.84% while parcels between one and two miles of the project area experienced an average change of -11.95% in EMV. Thus, exploratory analysis suggested that, on average, property values of residential parcels within one mile of the Midtown Greenway project area tended to be slightly less stable than property values of residential parcels between one and two miles of the project area. **Figure 1** illustrates the general trends in EMV for each year from 2002 to 2015 for our sample of 1,000 residential parcels within two miles of the project area. While the general trend in property values is similar between parcels within and beyond one mile of the project area, on average, it seems that EMV within one mile rises slightly more before 2008 and falls more after 2008.

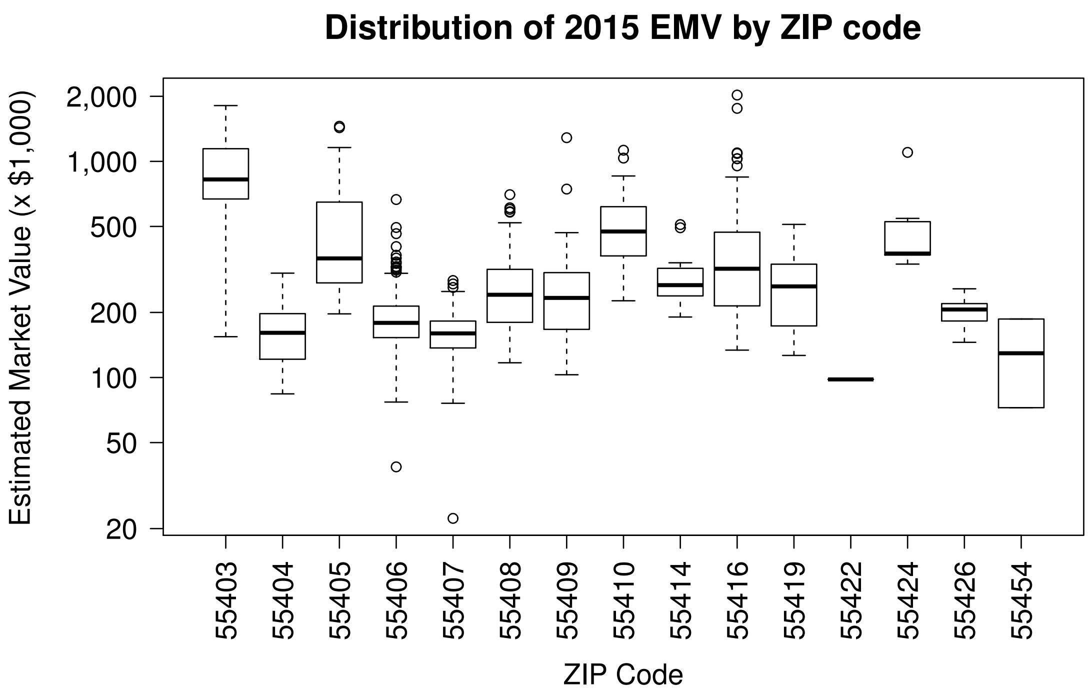


Figure 2. Distribution of 2015 EMV by ZIP Code for the subset of 1,000 residential parcels within 2 miles of the Midtown project. The edges of the boxes denote the 25th and 75th percentiles while the dark line in the middle of a box marks the median EMV.

**Figure 2** shows the distribution of 2015 EMV by ZIP code on this sample of residential parcels, which illustrates the variability between average EMV according to the ZIP code, which could be used as a rough indicator of the neighborhood that a parcel belongs to. Because there seems to be variability between EMV, which can be explained by ZIP code, we expected that a hierarchical modeling approach would be necessary to properly model the data.

Tables 1 through 7 in Section 8.1 display summary information for what we considered the best variables describing the parcels. These variables are all used in our regression model.

## 4 Methods

The distance information used in this term's analysis is updated from Winter '16 as we used GIS to calculate the distance from the parcel using the centerline of the project area, rather than a polygon that represented a 0.5 mile buffer around the center line. Thus, this new distance contains more detailed information than previously as the distance from parcel within 0.5 miles of the project to the project was previously calculated as 0. We used ESRI 2015 Tapestry Segmentation data at the census tract level to gather the variables we were interested in at the tract level and used GIS to merge this data with the original parcel data.

This analysis employs a hierarchical linear model to assess any association between the construction of the Midtown project and property values in the surrounding area. This model accounts for correlation between property values for residential parcels in the same census tract, which we use in place of a neighborhood, and thus allows us to control for characteristics of the tract in the model. Because the Midtown Greenway was constructed in multiple phases in different years, we modeled the property value for each year from 2002-2015 for each parcel. For the analysis, we adjusted all years of property values to 2015 dollars using the Consumer Price Index, to account for inflation.

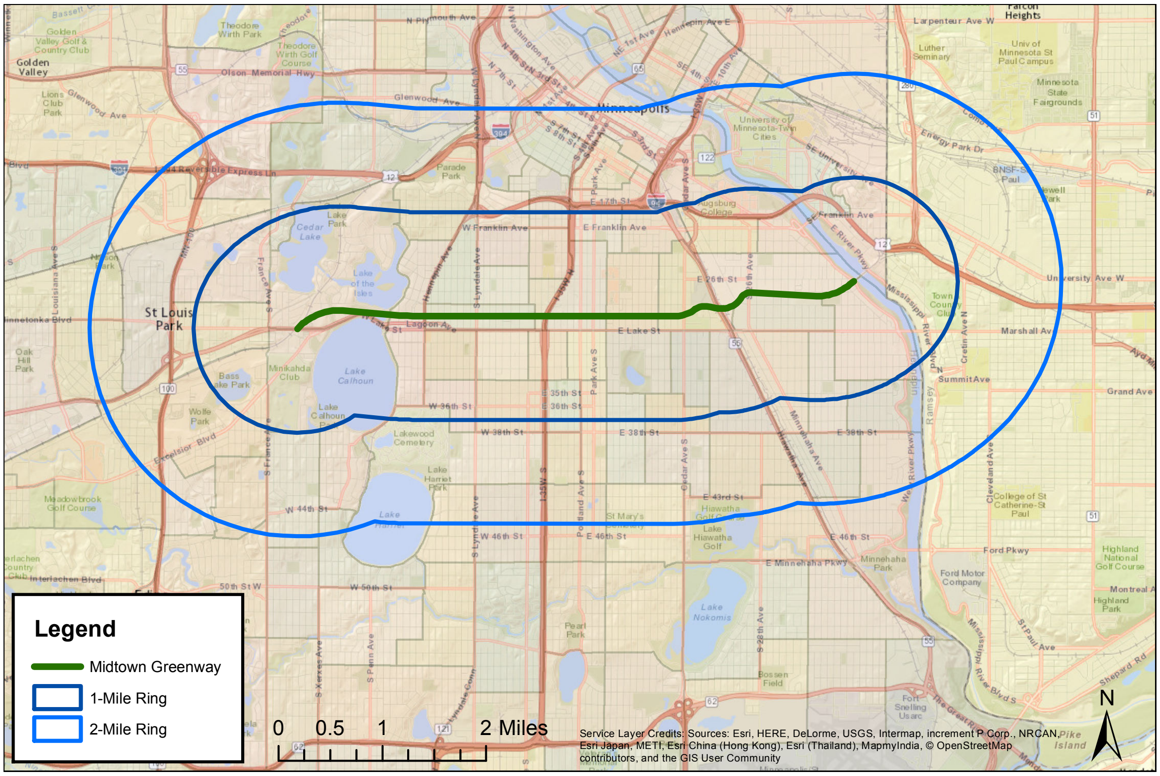


Figure 3. Map showing the Midtown Greenway centerline and 1- and 2-mile buffer rings around the greenway.

We first chose to subset only residential parcels of which there are 271,103 observations. From these residential parcels, we excluded any parcel with EMV less than $20,000 for any year and any parcels that are greater than two miles from the project area. This final subset includes 32,744 parcels. Because the model now fits on 14 observations, one for each year for each parcel, and because the modeling technique is computationally heavy, we chose to take a random sample of 1,000 parcels from this subset, so our final data set for modeling includes 14,000 observations. We ensured that this sample is representative of the subset of residential parcels within 2 miles of the project area by comparing summary statistics of our primary explanatory variables from the sample with the subset as a whole, and confirmed that the distributions of the variables are similar.

**Figure 4** illustrates what we mean by "hierarchical" in that we consider each year as nested in a parcel, which in turn is nested (associated) with a certain census tract. Section 8.2 includes a more detailed description of the modeling approach.

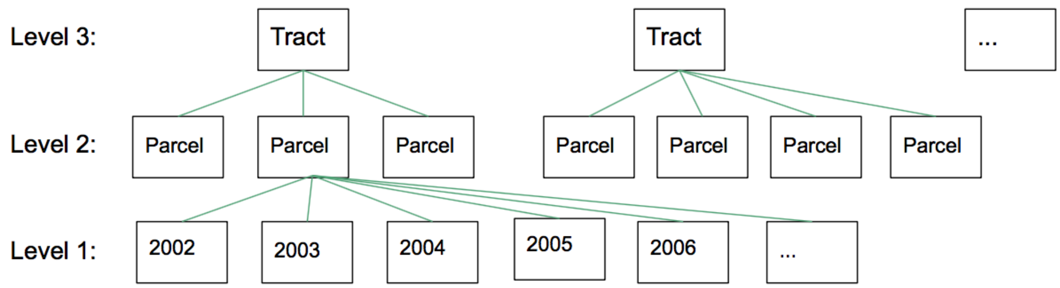


Figure 4. Illustration of the composition of data, with years nested in parcels, nested in census tract.

Our model fits the log of the EMV for each parcel for each year on characteristics of the tract and the parcel. Because this method of modeling allows us to predict the EMV for each year for a parcel and accounts for differences between parcels and tracts, we hope that this could account for any confounding factors that could be contributing to property values other than the specific explanatory variables that we use, in order to isolate the effect of distance from the Midtown Greenway project. Our final model additionally controls for variables such as the square feet of the parcel, the median household income in the tract, the median age in the census tract, and the diversity index that measures the level of diversity in the tract.

# 5 Results

We were able to obtain a final model that explains the year-by-year change in EMV of each residential parcel within two miles of the Midtown Greenway project area. At the parcel level, the model includes the distance away from the project area, an indicator for whether the parcel is within one mile of the greenway, an indicator for whether the parcel is located to the north or south of the greenway, and the log of the size of the parcel as explanatory variables. At the census tract level, the model includes the median household income of the tract, the square root of the diversity index, and the square root of the median age as explanatory variables.

The within one mile indicator attempts to control for the potential drop off in EMVs that would occur at the boundary where accessing the greenway becomes difficult enough that it no longer has a significant effect on property values. The direction indicator attempts to control for differences in neighborhood and market value between a parcel on the north side of the greenway and one on the south side that are both the same distance away from the greenway. Similarly, we used the three census tract statistics as proxies for the desirability of living in a given neighborhood to control for fluctuations in EMV caused by differences between neighborhoods. Finally, we used the size of the parcel (in acres) to control for differences in house size. When we ran our model, we found that both year and distance are statistically significant. With each passing year, the expected EMV of any given parcel increases by about 1% controlling for other variables. The distance parameter has a negative value, meaning that controlling for other variables, the expected EMV of a given parcel decreases the further away it is from the project area. More specifically, the expected EMV decreases by 1.2% for each tenth of a mile.

However, considering the interaction between the exact distance away from the project area and whether the parcel is within one mile of the project reveals that there is actually a more complex relationship between distance and EMV. Within one mile, an increase in distance of one tenth of a mile actually is associated with a 1.3% increase in expected EMV with all other factors held constant. Between one and two miles away from the greenway, an increase in distance of one tenth of a mile still is associated with a 1.2% decrease in expected EMV.

# 6 Discussion

After running our model, we determined that a parcel's distance away from the Midtown Greenway is a significant factor in its EMV. It is important to note that as this is an observational study, our results inform us only of a correlational relationship between change in EMV and distance, not a causal relationship.

One interesting finding from our model is that on average, a parcel's EMV is expected to increase by about 1% each year even after controlling for inflation and other factors in the model. However, this statistic does not tell the whole story. When we performed exploratory data analysis on the year-over-year change in EMV, we found that there was a steep drop in market value at the onset of the recession in 2008, and that overall, most parcels experienced a net decline in EMV between 2002 and 2015. The linear nature of our model made it hard for us to account for this sort of temporary fluctuation in prices, so the year estimate should be interpreted with a grain of salt.

Perhaps the biggest problem with the model we built last term was a lack of independence between the property parcels in our dataset. Since we were looking only at parcels near the project area, we encountered a great deal of spatial correlation. For example, recent sale prices of other homes in the same neighborhood are often used to set the market value of a home. Unfortunately, we were not able to account for this lack of independence last term because the linear regression model we used assumes independence. This term, we attempted to mitigate this problem by separating parcels into different census tracts and using a hierarchical model. This model is not hindered by lack of independence. We used demographic data for each tract to describe differences between tracts, and we are confident that our modelling approach resulted in a model that explains a much greater proportion of the variation in EMV than the model that we produced last term explained.

A foundational part of our model that we were not able to research and justify completely was our choice of where to delimit the EMV range of properties included in our model. In creating our model we chose to exclude properties with EMV value lower than $20,000 without thorough analysis of whether properties below this cutoff actually represented outliers. However, we believe that using this cutoff was a reasonable step to take for two reasons. For one, we suspect that properties with an EMV of $0 represent properties with unknown EMVs rather than properties that are worth nothing. Second, we think that most properties with EMVs between $0 and $20,000 are most likely properties that are abandoned, don't have buildings on them, or are under construction, categories of properties that would not provide very meaningful information to our model. Of course, more exploratory data analysis could be done on logged EMV values to determine whether properties with values lower than $20,000 actually represent outliers, and what range of EMV values is more appropriate. We additionally decided this term not to set an upper limit on the EMV values included.

One problem we faced when compiling our data about the direction of parcels from the greenway was that the greenway is not a completely straight line. To compute the direction data, we used GIS to calculate the heading of the line of shortest distance from each parcel to the greenway, with headings between 90 degrees and 270 degrees representing parcels located north of the greenway, and everything else representing parcels located south of the greenway. However, the heading of this line is harder to interpret when the greenway is not perfectly straight east-west, as is the case on both ends of the greenway (particularly the eastern end) where it curves to avoid obstacles and meet other trails. We were not able to thoroughly investigate this issue, so it is likely that there are some parcels for which we have incorrect direction data. However, we are not extremely concerned about this issue because the proportion of parcels it might affect is relatively small.

Furthermore, the eastern part of both our 1-mile and 2-mile buffer rings exceed the boundaries of Hennepin County and extend into Ramsey County across the river. Since we only had data about parcels in Hennepin County, our model is based on partially incomplete data. As with the issue with the direction data, however, we believe that this affects such a small proportion of the properties within two miles of the greenway that it is not cause for concern. Moreover, since the boundary between the two counties is a river traversed by a limited number of bridges, the access to the greenway is most likely diluted in Ramsey County. Therefore, any impact that the project may have had on properties in Ramsey County is probably quite small.

We also have some concerns about missing or ambiguous data. While the original dataset had information about which properties had garages and basements, we found that such a high proportion of the properties didn't have any data in these columns that that information was not statistically significant. Because the modeling that we used this term did not allow for testing on explanatory variables if there is missing information, we chose to exclude any variables that contained missing information, which means that our model could exclude information. However, since the interest of the project is to evaluate whether or not the community works projects have influenced surrounding property values, any explanatory variables we include other than distance are meant to control for any other possible influences on property value. We feel confident that whether properties have garages or basements is information we can exclude considering the other explanatory variables we are able to include.

# 7 Future Work

So far, we have only created a model that explains the change in estimated market value of residential parcels around the Midtown community works project. A future analysis of non-residential parcels near the Midtown community could potentially contribute to our completed work. Also, the rest of the projects have yet to be modeled and analyzed. New models should be made starting with the other completed project areas, Victory Memorial Drive and Fort Snelling. As for the community works projects that are in progress, it might be more difficult to analyze their effects because the projects haven't been completed and the information about the state of completion is generally difficult to find. Finally, projects that have not started should not be analyzed.

Some community work projects other than Midtown also suffer a similar "straight line problem" as we experience with the Midtown project. The centerline of Victory Memorial Drive, for example, has an L shape. Because of this irregular shape, it becomes hard to decide the direction from the project area for some parcels near the Victory project area. Without a clear basis to decide the directions of all the parcels, the direction parameter might not longer be valid. A potential way to address this issue is to disregard the direction parameter for the projects with irregular centerline shapes, which will of course, lose the information about the direction. Improved ways to deal with this issue should be considered.

Our model could also be improved if we find a convincing cut-off line for the total EMV values. As we have mentioned, our current final model uses $20,000 as a cutoff line, and we automatically treat residential parcels with adjusted EMV total below $20,000 as outliers. Again, we believe that setting a cutoff line is appropriate, but we believe it would be valuable to have more concrete justifications for a cut-off value in the future. One way to deal with this issue is to look through the observations with relatively extreme adjusted EMV total values in our sample, one by one. In this way, we could physically identify which observations could be eliminated based on human judgement. This approach, even though time-consuming, is likely to best illustrate a convincing reasoning behind the process of setting up a cutoff line. As for the ambiguous data and missing data problem addressed above, it is a bit hard to deal with this problem because it is created by the nature of our dataset rather than the modeling process. As we know, the garage variable in our dataset has a lot of missing values. Based on our knowledge, we were not able to find a better way than simply eliminating the variable. In other words, we got rid of this variable mainly because it has too many missing values. Thus, for the future work, other ways should be considered to deal with this issue of missing information.

Finally, even though we already adjusted EMV for inflation based on CPI, our model still does not seem to well explain some non-linear fluctuations during the recession period, nor does it account for the housing bubble in 2008. We think a better account of economic conditions is needed in the future models. We believe that rectifying the problems outlined above should go a long way towards developing a more robust model, and we look forward to doing so.

# 8 Appendix

## 8.1 Summaries of Important Variables

**Table 1**
ZIP Codes

| | |
|---|---|
| 55403 | 9 |
| 55404 | 19 |
| 55405 | 38 |
| 55406 | 266 |
| 55407 | 207 |
| 55408 | 65 |
| 55409 | 87 |
| 55410 | 64 |
| 55414 | 13 |
| 55416 | 192 |
| 55419 | 15 |
| 55422 | 1 |
| 55424 | 7 |
| 55426 | 15 |
| 55454 | 2 |
| Total | 1000 |

**Table 2**
Within One Mile

| | |
|---|---|
| Yes | 382 |
| No | 618 |

**Table 3**
Parcel Acreage

| | |
|---|---|
| Min | 0.03 |
| 1st Quartile | 0.12 |
| Median | 0.13 |
| Mean | 0.15 |
| 3rd Quartile | 0.16 |
| Max | 0.92 |

**Table 4**
Square Footage

| | |
|---|---|
| Min | 582 |
| 1st Quartile | 1111 |
| Median | 1290 |
| Mean | 1463 |
| 3rd Quartile | 1601 |
| Max | 7258 |
| No Data | 831 |

**Table 5**
2002 EMV (x $1,000)

| | |
|---|---|
| Min | 7.8 |
| 1st Quartile | 154 |
| Median | 191 |
| Mean | 241 |
| 3rd Quartile | 272 |
| Max | 1450 |

**Table 6**
2006 EMV (x $1,000)

| | |
|---|---|
| Min | 16.8 |
| 1st Quartile | 210 |
| Median | 252 |
| Mean | 315 |
| 3rd Quartile | 350 |
| Max | 2027 |

**Table 7**
2015 EMV (x $1,000)

| | |
|---|---|
| Min | 22.3 |
| 1st Quartile | 161 |
| Median | 204 |
| Mean | 276 |
| 3rd Quartile | 317 |
| Max | 2027 |

## 8.2 Detailed Description of Hierarchical Modeling Method

The modeling considers whether separate lines need to be fit, with separate intercepts and/or slopes (called the random effects) for each year, for each parcel in a tract. Thus, the model accounts for the effects of differences between tracts and parcels on the variability of the property values for each year. The model fits the EMV for one year for a parcel in a census tract, at this level the only covariate is time. At the parcel level, covariates include parcel's individual attributes including the parcel's acreage, distance and direction to the Midtown Greenway project along with an indicator variable showing whether the parcel is within a mile to the Midtown Greenway. At the tract level, covariates include the Census Tract's Median Household Income, Median Age and Diversity Index.This method of modeling allows the use of ANOVA likelihood tests to evaluate the significance of the variables in the model, however the model output does not report p-values, as is shown in the Appendix. A T-value with absolute value greater than 2 is a rough indicator that the covariate is significant. The modeling uses the lme4 package in R, which allows for multilevel modeling.

## 8.3 Table of Coefficients for Midtown Greenway Model

| **Variable Name** | **Parameter Estimate** | **Standard Error** | **T-statistic** |
|---|---|---|---|
| Intercept | 15.64588 | 0.63967 | 24.459 |
| Year | 0.02217 | 0.00471 | 4.711 |
| Distance (miles) | -0.12108 | 0.05251 | -2.306 |
| Within one mile | 0.0413 | 0.09659 | 0.428 |
| Direction South | 0.05514 | 0.05076 | 1.086 |
| log(Acres) | 0.42024 | 0.02776 | 15.139 |
| Median Household Income | 0.08643 | 0.04181 | 2.067 |
| sqrt(Diversity Index) | -0.14633 | 0.02881 | -5.081 |
| sqrt(Median Age) | -0.21427 | 0.09029 | -2.373 |
| Distance:Within one mile | 0.24612 | 0.08751 | 2.812 |
| Within one mile:Direction South | -0.24521 | 0.05595 | -4.383 |
| Year:Direction South | -0.00438 | 0.00218 | -2.006 |
| Year:sqrt(Diversity Index) | -0.00285 | 0.00065 | -4.351 |