Low-Latency Gentrification Prediction with Airbnb Reviews

*By* Leo Kitchell\*

We propose a methodology for predicting neighborhood gentrification using high frequency, publicly available AirBnB data. Leveraging 3.8 million text reviews from Jan 2014 to Dec 2019 across 28 US cities, we find guest reviews and rental characteristics to be highly predictive of gentrification during the same period. Both structured features (e.g. number of listings) and unstructured features (e.g. word frequency in reviews) are found to be significant predictors. We then train a set of city specific random forest models and use review data collected between Jan 2020 and Oct 2021 to forecast current gentrification rates. These models are provided freely to enable rapid policy response and further research.[[1]](#footnote-1) †

\* Kitchell: Claremont McKenna College, (email: lkitchell21@cmc.edu). I would like to thank my reader, Professor Gelman, for his guidance, patience, and modeling knowledge upon which this paper is founded.

Gentrification is generally understood as a process of change whereby neighborhoods which have been historically under-capitalized experience an influx of new residents with high social capital. This demographic change corresponds with an increase in investment in the neighborhood, generally transforming the housing stock and character of the neighborhood. In theory, residents of gentrifying neighborhoods stand to benefit from increased retail and housing investment and the affiliated job opportunities and tax revenues. In practice, a gentrifying neighborhood’s original residents often find themselves forced out by a combination of passive market factors, namely increased rents, and active displacement mechanisms such as planned dilapidation, eviction, and reinvestment cycles. This widespread displacement of vulnerable populations is the principal concern of activists and is central gentrification’s contentiousness. Unfortunately, each of the three pillars of change (demographic, housing and retail stock, and cultural character) have characteristics which make quantifying their change difficult, and which consequently limits scholarship and effective policy intervention.

One of the principal data challenges is that detailed demographic data is infrequently collected, so researchers and policymakers are constantly fettered by data from the Census or the 5-year American Community Survey (ACS). This lag means that research must be conducted on sparse time series, limiting the research's power, while on the policy side the lag makes early identification of gentrification difficult, and dynamic monitoring impossible. Changes in the physical neighborhood real-estate composition are generally the easiest to detect, as home prices, building permits, and other measures of investment operate on shorter lag. The final pillar of change is also the most difficult to quantify. While the arrival of craft coffee shops, whole foods, and street art may be easily visible signs of cultural change for residents of a gentrifying neighborhood, this local knowledge does not get translated into traditional data sets. However, new data sources may solve some of these challenges.

Crowdsourced data from online platforms like Yelp, Zillow, and Airbnb are enabling researchers to investigate gentrification trends at a far more granular level and in real time. These data also supplement traditional metrics with novel features, such as user reviews and textual descriptions. These features represent an exciting frontier in gentrification research and may enable a better understanding of neighborhoods and their residents.

This paper utilizes 3.8 million text reviews from Airbnb listings in 28 U.S. cities to infer neighborhood characteristics which can be used to predict gentrification in real time. Our analysis is organized into 7 parts. Section 1 summarizes key empirical studies on gentrification with a focus on their specifications of gentrification and the geographies which are suitable for analysis. Section 2 introduces two recent studies which use low-lag data from Zillow and Airbnb, respectively, to predict measures of gentrification. Section 3 outlines our gentrification operationalization and two model specifications, with and without unstructured text data, which will be used to generate our results. Sections 4 and 5 present these results from the respective specifications. Finally, section 6 considers the implications of these findings and introduces an alternate model for near term gentrification prediction.

# I. Empirical Literature Review

The data challenges introduced in the previous section have hindered gentrification research, but a robust literature still exists, particularly when it comes to identifying areas that are ripe for gentrification. The term gentrification originates from the migration of London “gentry” into lower-income neighborhoods in the 1950’s and 1960’s (Glass 1964). Since the term’s coining, researchers have developed progressively more complicated methods of identifying gentrification, and considerable variation still exists in its operationalization. Among this backdrop of variation, resident income, education level, and race are commonly used measures of quantifying the first—demographic—pillar gentrification (Ding et al, Freeman 2005, Bates 2013, Goetz et al.). These variables generally describe the first pillar of gentrification, though significant changes in demographics may also signal a shift in the cultural character of a neighborhood. Although neighborhood demographic changes are necessary to establish gentrification, they are not sufficient on their own.

A central tenant of gentrification is its disproportionate impact on vulnerable communities and its contrast with historical neighborhood characteristics. Following World War II, mortgage guarantee policies passed in the GI bill helped subsidize Americans to move from metropolitan areas to suburbs. However, actions taken by the Federal Housing Administration rendered these suburbs largely unavailable to black Americans (“Urban Displacement Project”). The resulting migration flows, known as white flight, spurred a trend of disinvestment in inner city communities. It is only in these previously disinvested neighborhoods where gentrification may occur. The origin of this disinvestment also informs our identification of such communities.

Ding et. al (2016) define neighborhoods as gentrifiable if their median household income is below the citywide median at the beginning of the analysis time period and gentrifying if it was gentrifiable at the beginning of the period and experienced a percentage increase in median gross rent or median home value above the citywide rate and an above citywide increase in percent of residents with a college degree. A similar specification by Meltzer (2017) includes only average household incomes to identify such neighborhoods. These specifications identify many neighborhoods which are gentrifying, but fail to capture the extent to which gentrifiable neighborhoods have been deprived of social capital and investment. To better capture these housing dynamics, Freeman (2005) includes the share of the housing stock built in the last 20 years and restricts gentrifiable neighborhoods to only include those whose share of new housing was below the 40th percentile of their metropolitan area. Freeman also restricts their sample to neighborhoods within the central city of metropolitan areas where real housing prices increased over the period of analysis.

Bates (2015) adds a final layer to this analysis by incorporating social capital measures, including the a tract’s proportion of owner occupied housing units, and the percentage of residents who identify as non-White. In both cases, a tract must have a larger proportion than the citywide rate to be considered gentrifiable. This additional further detail limits the universe of gentrifiable tracts, but better encapsulates the historical context of urban gentrification. It is this, more detailed, view which our study aims to emulate.

# II. Airbnb Text as Gentrification Predictors

Indented. Forthcoming

# III. Data

## ACS Gentrification Data

Following in the footsteps of many others (Meltzer 2017, Lester and Hartley 2014, Ellen and O’Regan 2008, McKinesh et al. 2010, Freeman 2008, Hwang 2014) our level of analysis will be the census tract. Census tracts contain approximately 4,000 residents and are frequently used to model neighborhood evolution. Demographic data is sourced from the American Community Survey (ACS) 5 year estimates via the Census API, which provide detailed information at the block group level.

As with the authors in sections two and three, our analysis of gentrification begins with identifying gentrifiable tracts. These will be defined as tracts with median household incomes below the county’s 50th%ile, educational attainment below the county’s 50th%ile, and median rent below the county’s 50th%ile. Our Airbnb data is already limited to the cores of metropolitan areas, so further geographic filtering by geography is not needed.

(2)

To measure gentrification, each gentrifiable tract will be given a score composed of an equally weighted average of the following variables, downloaded from the ACS: median household income, median monthly rent payment, percent of residents who identify as White and non-Hispanic, and percent of residents with a bachelors or higher degree. Changes in rent will be adjusted by the metropolitan statistical area’s consumer price index, as reported by the Federal Reserve. The percentage change a tract’s gentrification score at the outset of the study (2014) and its conclusion (2019) will be our dependent variable.

(3)

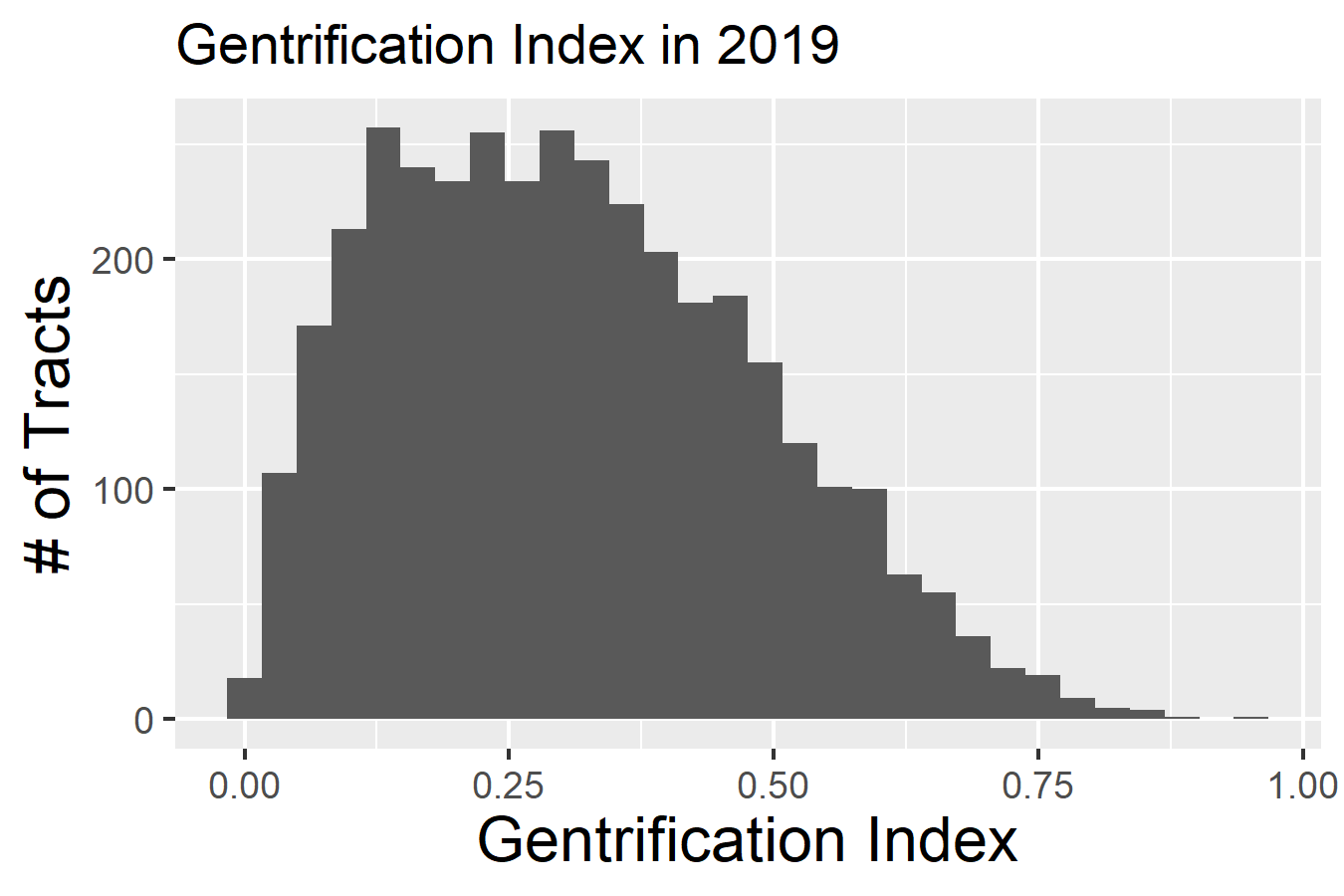
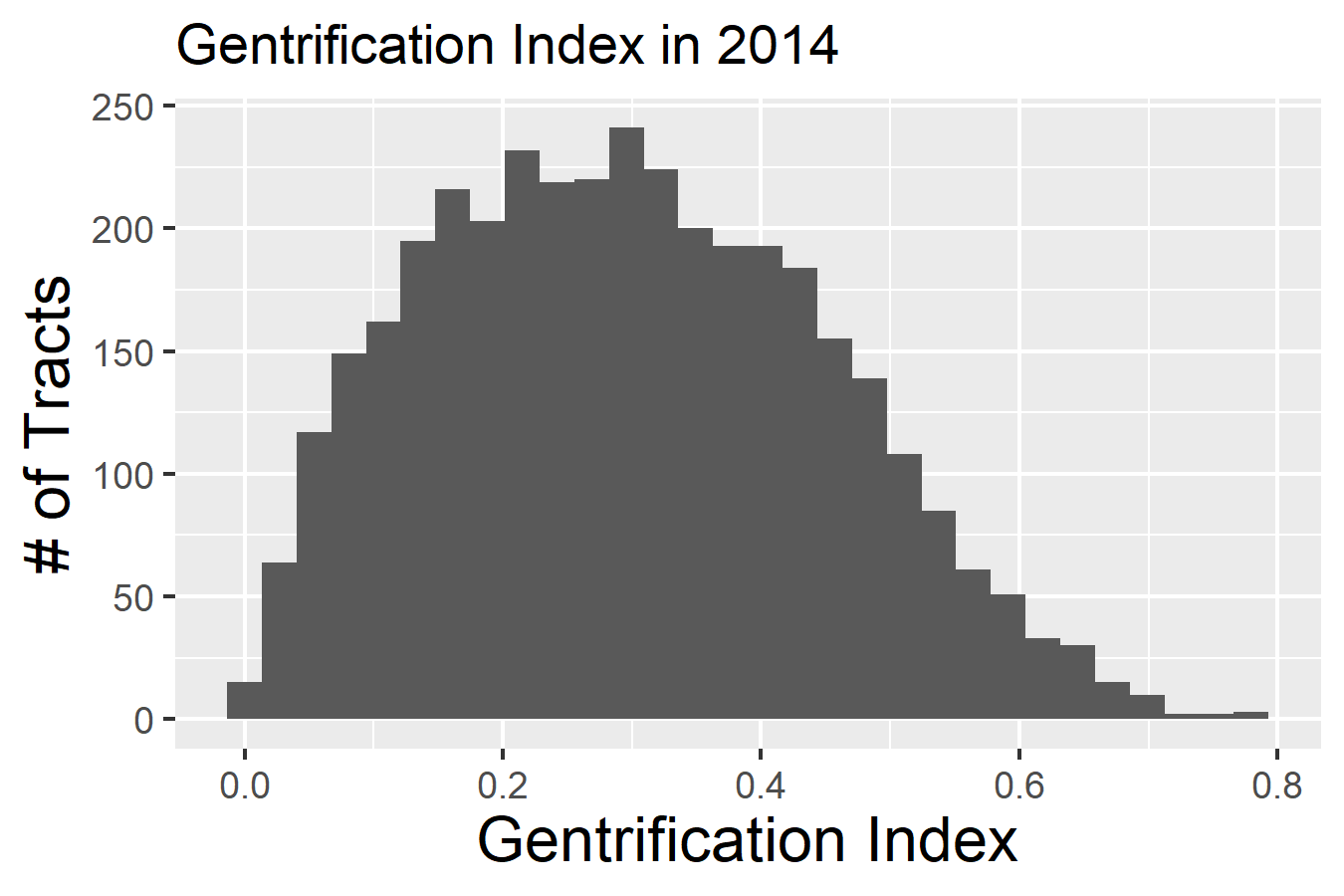
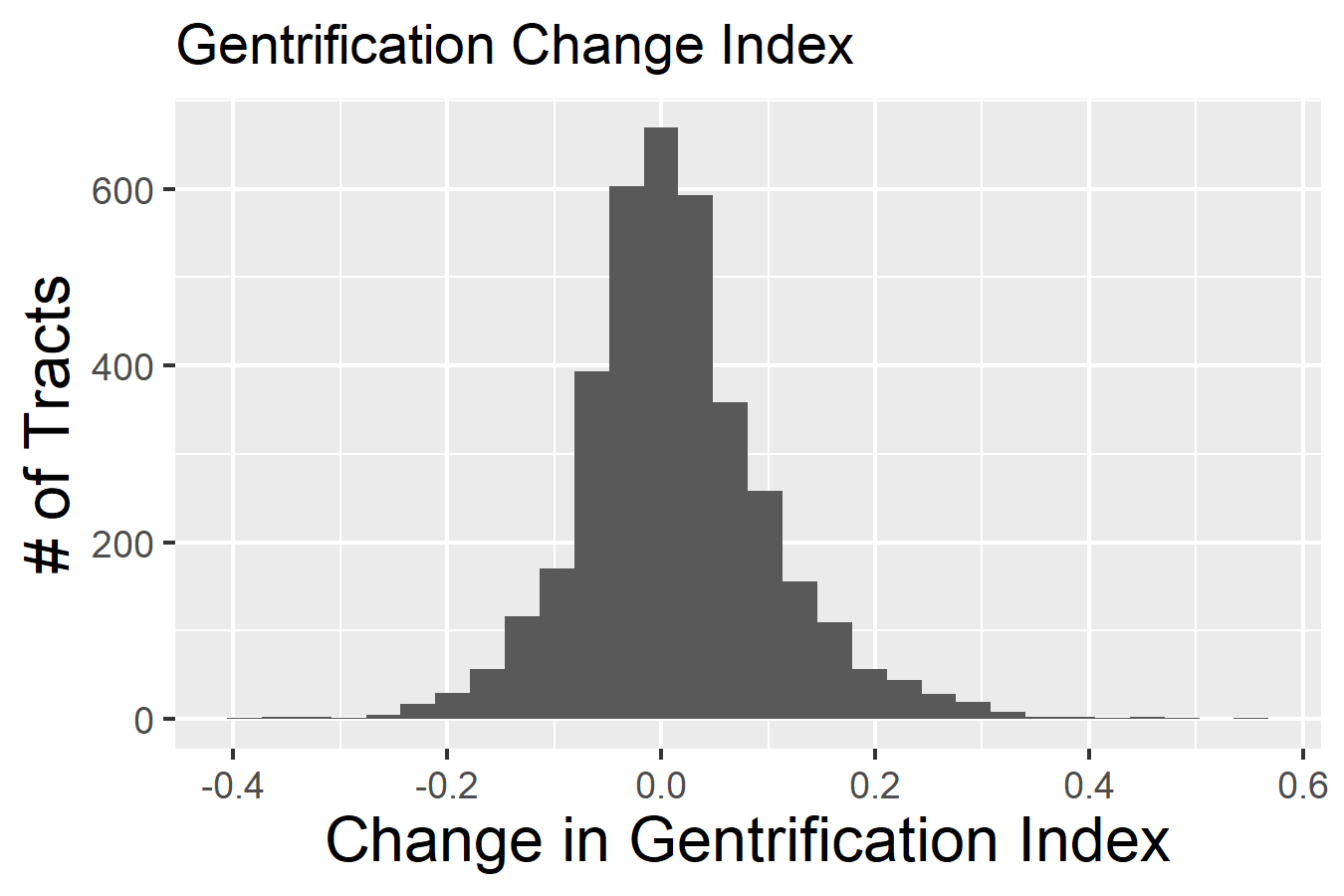


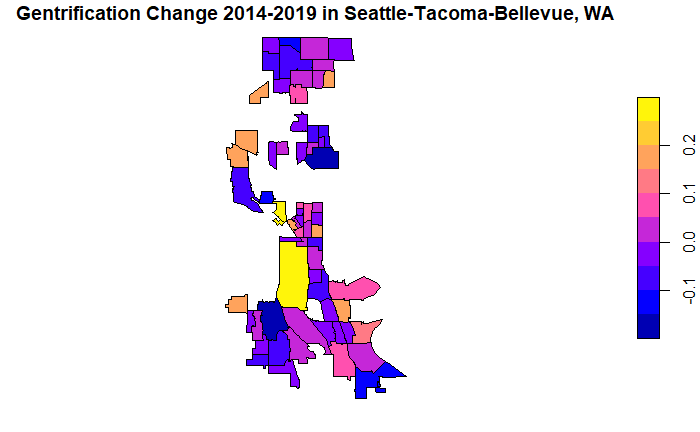
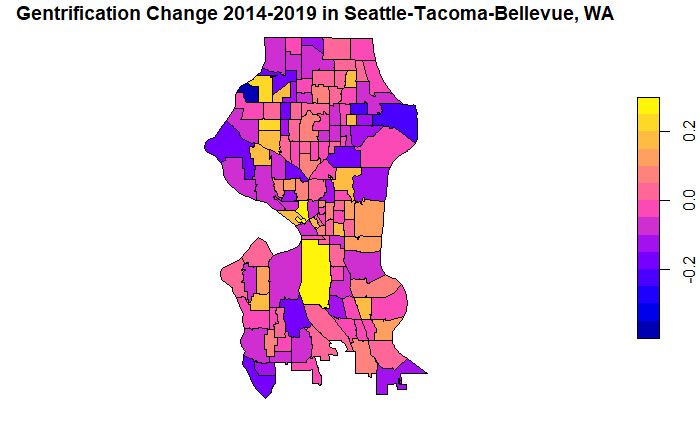
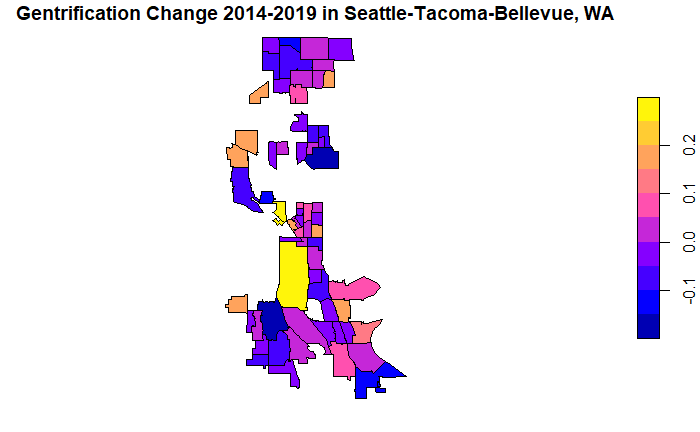
Figure 1. distribution of gentrification index in gentrifiable neighborhoods

*Notes:* These are the notes applicable to the figure. The style is named Figure Notes.

Table 1—Summary Statistics of Census and Airbnb variables

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Minimum | Median | Mean | Maximum | Std. Dev. |
| Gentrification Change | -0.37 | 0.01 | 0.02 | 0.50 | 0.09 |
| Initial Population | 173.00 | 3931.00 | 4142.00 | 28827.00 | 1796.71 |
| Population Change | -0.61 | 0.02 | 0.04 | 1.17 | 0.15 |
| # of Listings | 1.00 | 5.00 | 10.73 | 783.00 | 21.25 |
| # of Reviews | 1.00 | 182 | 512.3 | 28041 | 1092.00 |
| Avg. Listing Price | 19.00 | 107.00 | 133.23 | 9900 | 222.79 |
| Avg. Listing Rating | 1.00 | 4.76 | 4.63 | 5.00 | 0.51 |
| Avg. # of Bedrooms | 1.00 | 1.37 | 1.56 | 6.00 | 0.67 |
| Avg. Location Rating | 1.00 | 4.74 | 4.67 | 5.00 | 0.31 |

*Notes:* n = 3129 for all variables.



Low Gentrification High Gentrification

Figure 2. Gentrification Change in Seattle   
all tracts (left) and gentrifiable tracts (right)

*Notes:* Gentrification is prevalent in South Central Seattle and in pockets of North Seattle including Ballard and Wallingford. Color differences between maps at left and right is due to graphing parameters; gentrification change values are not recalculated or adjusted after removing neighborhoods with high initial household incomes.

## B. AirBnb Data

To explain tract variation, we propose XXX classes of random forest models: a baseline model without Airbnb data, one with structured Airbnb data and one with the unstructured added. Independent variables in the baseline will be limited to initial population and population growth. The structured Airbnb data model will include the tract’s population, number of Airbnb’s, the number of Airbnb Reviews, the Airbnb star rating and Airbnb location rating. We expect that a higher density of Airbnb’s will correlate with changes in neighborhood characteristics that make a tract more appealing to gentrifiers. The more appealing these characteristics are, the higher we expect the Airbnb location rating to be, thus we expect both the number of listings and the location rating to be positively correlated with gentrification.

In the second specification we include a matrix of unique words used in a given tract’s Airbnb listing reviews and their prevalence. We also hope to include a measure of review sentiment, although the feasibility of this piece is still being determined. These covariates will be added to the first model specification. We predict that the textual understanding of a neighborhood, as given by user reviews, contains useful information about the neighborhood demographics, and thus expect the matrix to increase the predictive power of our model. We do not have hypotheses at present about which sorts of words will have high explanatory power

(1) Equation – The style is named Equation,

where … The style is named Normal No Indent when you want a non-indented paragraph, especially after an equation. Place an extra return after the equation for spacing.

(2) Equation – The style named Equation.

The style is named Normal after an equation when you want an indented paragraph. Place an extra return after the equation for spacing.

### *Third Level Heading*.—This is an example of a third level heading. The style is named Heading 3. A period and em dash should separate the heading from the paragraph text. When applying this heading, additional formatting can be manually added if preferred, but it is not required. Italics will automatically be applied in production.

#### **Fourth Level Heading:** This is an example of a fourth level heading. The style is named Heading 4. A colon should separate the heading from the paragraph text. When applying this heading, additional formatting can be manually added if preferred, but it is not required. Bold will automatically be applied in production.

## B. Footnotes

Insert footnotes as necessary using the Insert | Footnote menu item in Word 2003 or the References | Insert Footnote ribbon item in Word 2007.[[2]](#footnote-2) The footnote number in the text is formatted automatically. An area for the text for the footnote appears at the bottom of the page and is formatted automatically.

## C. List Styles

Examples follow for bulleted and numbered lists.

* This is a bulleted list. Only one level is generally used.
* The style is named List Bullet.
* This is a bulleted list.

1. This is a numbered list. Only one level is generally used.
2. The style is named List Number.
3. This is a numbered list.

## D. References

A sample reference is shown below.

# I. First Level Heading – The Style is Named Heading 1

Type the text here. The style is named Normal for regular paragraphs. Type the text here. The style is named Normal for regular paragraphs. Type the text here. The style is named Normal for regular paragraphs.

## A. Second Level Heading – The Style is Named Heading 2

Type the text here. The style is named Normal for regular paragraphs. Type the text here. The style is named Normal for regular paragraphs. Type the text here. The style is named Normal for regular paragraphs.

(1) Equation – The style is named Equation,

where … The style is named Normal No Indent when you want a non-indented paragraph, especially after an equation. Place an extra return after the equation for spacing.

(2) Equation – The style named Equation.

The style is named Normal after an equation when you want an indented paragraph. Place an extra return after the equation for spacing.

### *Third Level Heading*.—This is an example of a third level heading. The style is named Heading 3. A period and em dash should separate the heading from the paragraph text. When applying this heading, additional formatting can be manually added if preferred, but it is not required. Italics will automatically be applied in production.

#### **Fourth Level Heading:** This is an example of a fourth level heading. The style is named Heading 4. A colon should separate the heading from the paragraph text. When applying this heading, additional formatting can be manually added if preferred, but it is not required. Bold will automatically be applied in production.

## B. Footnotes

Insert footnotes as necessary using the Insert | Footnote menu item in Word 2003 or the References | Insert Footnote ribbon item in Word 2007.[[3]](#footnote-3) The footnote number in the text is formatted automatically. An area for the text for the footnote appears at the bottom of the page and is formatted automatically.

## C. List Styles

Examples follow for bulleted and numbered lists.

* This is a bulleted list. Only one level is generally used.
* The style is named List Bullet.
* This is a bulleted list.

1. This is a numbered list. Only one level is generally used.
2. The style is named List Number.
3. This is a numbered list.

## D. References

A sample reference is shown below.

REFERENCES (this style is called Reference HeADING)

Last, First, and First A. Last. 2003. “Article Title” In Publication Name, ed. First Last. 335–405. City: Publisher. The style named is References.

# II. Sample Figure

A sample figure is at the end of this document. At publication, your figure must fit within the width of one printed journal page (5 inches or 12.7 cm).

If you choose to place a figure indicator in the text in the approximate location where the figure should go as shown below, the style of this indicator is called Figure Placeholder. You may choose to place the figures themselves at the end of the document.

[ Insert Figure 1 Here – The style is named Figure Placeholder ]

# III. Sample Table

See the sample table at the end of this document. Your table may take up the full page width (5 inches or 12.7 cm) and must have no more than 9 columns.

You may choose to place a table indicator in the text in the approximate location where the table should go as shown below. The style of this indicator is named Table Placeholder. You may choose to place the tables themselves at the end of the document.

[ Insert Table 1 Here – The style is named Table Placeholder ]

You may enter your figures on this page of the template. Your figures may take up the width of one printed journal page, which is 5 inches or 12.7 cm.

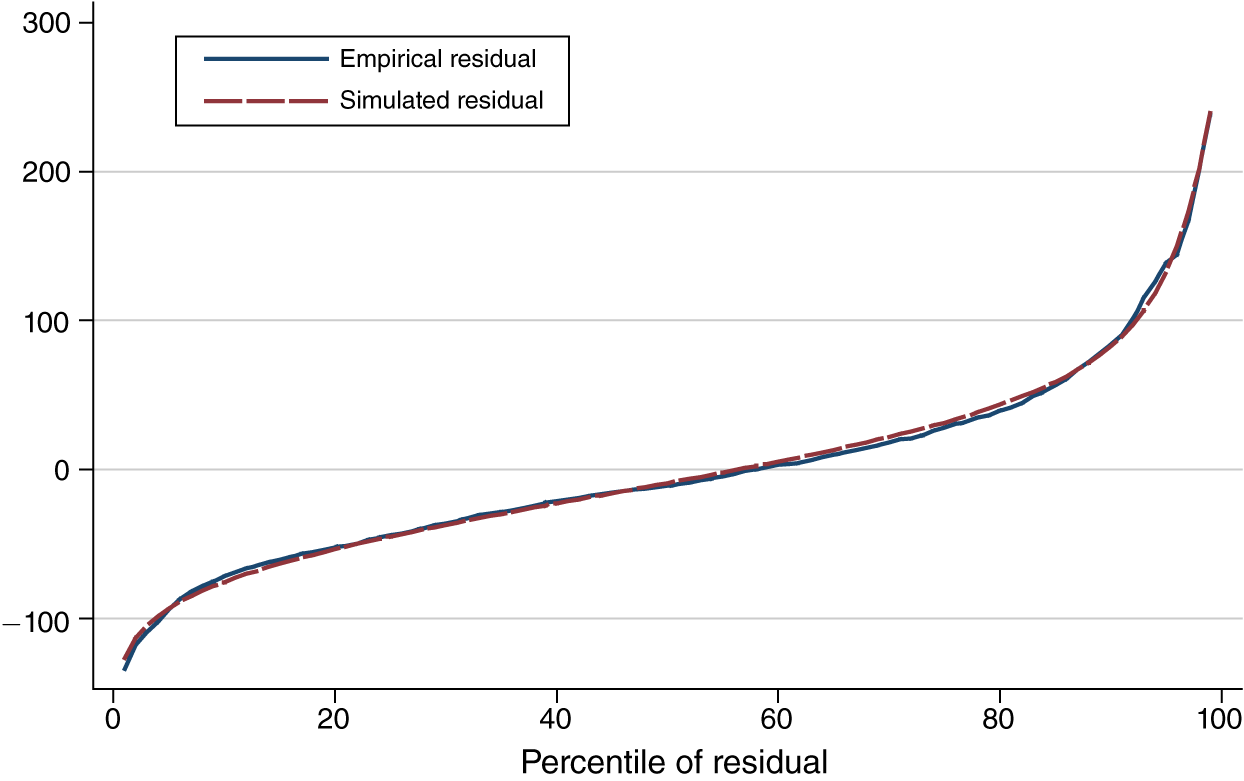


Figure 1. The Title of the Figure  
The Style is named Figure Title

*Notes:* These are the notes applicable to the figure. The style is named Figure Notes.

You may choose to type your tables on this page of the template. Your tables may take up the width of one printed journal page, which is 5 inches or 12.7 cm. It should contain no more than 9 columns, which includes any row headings.

Table 1—The Title of the Table, The Style is Named Table Title

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1948–2007a | 1948–1972 | 1973–1994 | 1995–2007 |
| Panel A. The style is named Table Text |  |  |  |  |
| Tangible | 11.4\* | 11.2 | 12.3 | 10.4 |
| Intangible | 8.6 | 5.9 | 9.2 | 12.8 |
| Panel B. Share of capital input |  |  |  |  |
| Tangible | 76.2 | 82.6\*\* | 74.8 | 66.1 |
| Intangible | 23.8 | 17.4 | 25.2\*\*\* | 33.9 |

*Notes:* These are the notes applicable to the table. The style is Tables Notes.

*Source:* Author calculations. These are more table notes. The style is Table Notes.

a Applicable to the whole period. The style is named Table Footnote.

\*\*\* Significant at the 1 percent level. The style is Table Notes.

\*\* Significant at the 5 percent level.

\* Significant at the 10 percent level.

# VII. Appendix

1. † Models, replication data, and code can be accessed at https://github.com/LeoKitchell/SeniorThesis/ [↑](#footnote-ref-1)
2. Type the footnote here. It is automatically formatted to the correct style. [↑](#footnote-ref-2)
3. Type the footnote here. It is automatically formatted to the correct style. [↑](#footnote-ref-3)