Analysing Gentrification in American Cities with Zillow Economics Data

Cameron Brenke, Brian Kalinowski, Claudia Trafton

Department of Computer Science and Engineering, Syracuse University
Syracuse, New York

ABSTRACT

Gentrification is one of the hottest topics in current sociological research, yet it is not well understood nor is it well measured. The goal of this research is to provide a way to measure gentrification through the analysis of housing and rental data from Zillow in hopes of providing a way to better understand gentrification from a quantitative perspective. Additionally, this research seeks to leverage data mining methods to extract causes and trends in gentrified cities. The main data mining technique used in this analysis is Dynamic Time Warping (DTW). This analysis explains how DTW can be used to understand how gentrified a city is relative to other cities. It is the hope of this study to provide useful information to future studies to continue to create a better understanding of gentrification, including how and why it occurs.

INTRODUCTION

The term, gentrification, was coined in 1964 by sociologist Ruth Glass to refer to the reappraisal and use of old buildings by artists in London's Islington neighborhood. while Gentrification is officially defined as "the transformation of neighborhoods from low value to high value" by the Center for Disease Control and Prevention, [1] a quick Google search for the word links to articles about the displacement of people of color, rising housing prices, and decreased neighborhood crime.

Even though gentrification is an important and well-studied topic, it is still a vague and politically-charged word, lacking any standard quantitative measures. [2] Additionally, the research on gentrification has led to even more convoluted results from being performed on varying geographic scales and with unstandardized data from a mix of sources. Most research has been performed at the neighborhood level, but only in cities. A neighborhood is defined as a geographically localized community within a city, town, etc, and so without including smaller population demarcations such as towns, research has inadvertently lead to a vague understanding of gentrification at a larger scale. It is not necessarily the case that the summation of city based neighborhood gentrification metrics accurately portray the gentrification of the city as a whole.

This body of work seeks to create clarity around gentrification at the city level by analyzing housing and rental data within a given city to generate a quantitative gentrification score. It is the hypothesis of this research that data mining methods can be used to extract a numerical score that indicates the level of the gentrification a city experiences relative to other cities.

DATASET AND DATA PROCESSING

A. Dataset

According to research done by the Urban Displacement Project, some of the biggest indicators of a gentrifying area are lack of affordable housing, lack of market-rate housing and sharp increases in housing prices over a short period of time. [3] For this reason, this research focused on housing and rental price metrics. The data was sourced from the Zillow Economics Data, with the City Time Series tables[1] for both rental and owned properties as our core datasets.[4] The datasets contains metrics about both rented and owned properties within a given city,

[1] City_Time_series.csv can be found at the following location:
https://www.kaggle.com/zillow/zecon#City_time_series.csv

including home values, selling prices, estimated market rates, price-rent ratios, value indices and more over the period of 1996 to 2017.[2] Altogether, the dataset had 81 total columns/features.

B. Feature Selection

According to a study from Pace University, demand factors that are commonly associated with gentrification such as population increase and more amenities like shopping and dining are correlated with increases in rental prices. [5] In another study done in New York City, an area considered to have highly gentrified neighborhoods, it was found that between 2000 and 2014, rental prices increased by a whopping 53% city-wide. [6] This evidence so clearly stated the correlation between increasing housing/rental prices and gentrification, which is why these were chosen for our core metrics in this study.

Since the analysis was done on time series data, which tracks the trends in data over time, the normal ebb and flow of seasonal rental changes had to be taken into account. Therefore, the core features selected were Zillow Rental Indices (ZRI) for rental metrics and Zillow Housing Value Indices (ZHVI) for the housing metrics. These indices represent seasonally adjusted market rate metrics on rental properties. Therefore, the analysis of ZRIs and ZHVIs in the context of time series is not sensitive to seasonal changes. [4][3]

C. Data Pre-processing

In order for the data to be useful, the data needed to be preprocessed and cleaned. Data preparation and cleaning was performed in Python using the Pandas library. Since the dataset was so large and spanned over twenty years, there were several tasks that had to be performed to make the data ready for analysis. These included removing empty entries, aggregating rows, and labeling data. This section will describe each data processing task.

Data preprocessing tasks

The Zillow city time-series dataset contained 81 columns (79 measures, date column, and the city a name column) and roughly 3.76 million records. Each individual row of the dataset is all 79 measures for a single city during a year/month period. The total date range is from April 1996 to December 2017, with 16,636 unique cities. The most apparent issue with the raw dataset was the high amount of null values across all of the measure columns. Simply dropping all rows containing null values drastically reduced the number of cities we would be able to research and would severely limit the validity of any results we could produce. Instead of removing all rows containing null values, we implemented a 4-step approach to fill null values where possible, and aggregate metrics in a way that maximized the number of cities we could retain.

Step 1: Removing Null Values

The raw data was filtered to only include the feature columns within its group (rental, core housing, sale housing), along with the Date and City/State identifier columns. Rows that contained null values for all feature columns were removed outright, as there was no valid information for that city during the given time period.

Step 2: Data Aggregation by City and Backfilling

Next, the data was grouped by each unique State and City combination, with the date column being set as the index, as is common with most time-series analysis. Then, looking at each individual State/City grouping, any metrics containing null values for that individual city were filled with the back-fill method. The rationale behind this was all the null values in the time-series were due to lack of data available to Zillow for a particular city. Once Zillow started receiving/collecting data from a city, there were no pockets or gaps of null values. Zillow started receiving data for rental properties in 2010. All those rental metrics contained null values from 1996 to mid-2010. Once Zillow started recording rental data for a city, there were no instances of null values in say 2014, with valid data in 2013 and 2015.

Since our goal is to measure change over time, backfilling previous null values with future metrics was the neutral solution. Filling these values with averages, or 0's would falsely create positive or

[2] The full data dictionary can be found at the following location: https://www.kaggle.com/zillow/zecon#DataDictionary.csv

[3] For a list of the core rental and housing metrics, see Appendix I.

negative spikes, thus compromising the change over time. Since the backfilling only took place within each individual city, there was no risk of populating one cities metrics with a different city.

Step 3: Further removal of aggregated rows

The back-filling method only works if there is at least one value to propagate. If every value in a column is null, it will remain so after the processing in step 2. After attempting to backfill by each individual city, any cities with metrics still containing null values were removed from our population. The majority of the cities removed in this step were smaller cities/townships in mostly rural areas.

Step 4: Aggregate into yearly averages by city

Of the cities that were retained after the null value filling, each metric was aggregated into a yearly average per city. This helped minimize the amount of inter-city date variance, and since data was only represented once for each month within each year, the overall trend per city was preserved.

These steps were performed for each grouping of features we chose to analyze. We were able to retain 10,132 cities for the rental group, 9,771 for the core housing group, and 1,344 for the housing sale group. Since the filling method was performed with each individual metric, we discovered that segmenting the raw data into distinct groupings of similar metrics, allowed us to retain the maximum number of cities for each group. The downside to this is that not all cities were represented in each grouping, some cities in one group may not be represented in another and vise versa. This makes a comparison of cities within each group difficult due to the varying group sizes.

After cleaning and filling the null values and averaging each yearly metric, the final step in our preprocessing was to normalize the measurement features for each grouping with the min-max scaling method. This scaling allows us to compare values like rental/home prices which are in the thousands to hundreds of thousands range to other values like a price-to-rent ratio that is represented as a float value between 0.0 and 1.0. Scaling was performed using the Python sklearn library's MinMaxScaler class. [12] This scaling is also required for accurate results when being ingested into various machine learning models like clustering.

Models and Methods

This research used the following models to generate a quantitative score that represents the level of the gentrification of a given city. A common misconception with gentrification is if a particular city has a high cost of living, it must mean it's been gentrified. However, this is not entirely accurate. Because gentrification is a process that occurs over time, The current housing or rental prices need to be high in the context of their historical value. What may be considered a high rent or home value in one area, could be completely normal or even below average for another area. According to the most recent Zillow report from April 2019, the median home value in the US is $227,000, and the median rent price is $1,477.[16] Depending on one lives, these numbers may seem very low or maybe even higher than expected. A quantifiable gentrification index can't be determined from an arbitrary value of high or low housing costs. It needs to come from historical costs for an individual region and any dramatic changes it has gone through over time.

A. Dynamic Time Warping

The primary model used was the Dynamic Time Warping (DTW) algorithm. DTW calculates a non-linear mapping between two curves to calculate a similarity score based on the sum of all the distance vectors calculated between points on the curves. This research utilized the tslearn library for Python, which calculates each distance vector using Euclidean distance.[6] A higher score indicates less similarity between the curves, while a lower score indicates that the curves follow a similar pattern. Essentially, DTW is able to perform the task of looking at the two signals or time series curves and detecting any patterns or similarities. [7]

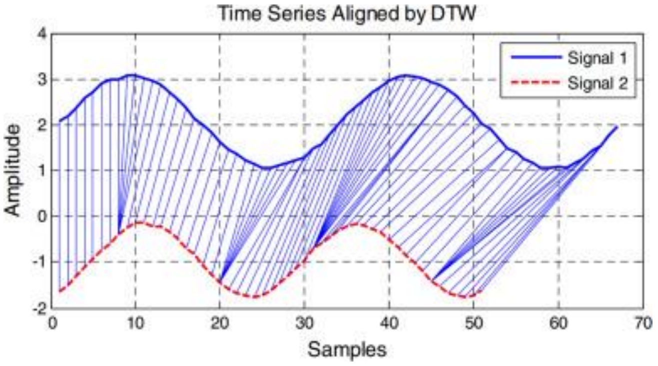


Fig. 2 An example of Dynamic Time Warping being applied to two Time Series. The lines between the curves are the distances that aggregated after all have been calculated to determine the DTW score.

For the case of gentrification, this research theorized that the DTW scores generated for two highly gentrified cities would be low, and the DTW score between a gentrified city and a non-gentrified city would be high. DTW scores were obtained using the tslearn Python library. A DTW score was calculated for every city's time series against a gentrification control time series, described in the following section. To further classify the cities, a label was assigned to each city to indicate the general level of gentrification experienced by that city with the following thresholds:

Label	Low Threshold	High Threshold
HIGH_GENT	0.0	1.0
MID_GENT	1.0	2.0
LOW_GENT	2.0	3.0
NO_GENT	3.0	4.0

Fig. 3 The thresholds for each gentrification label. The thresholds represent the DTW score threshold, with the Low Threshold, showing the lower limit (inclusive) value of the given label and High Threshold showing the upper limit (exclusive) of the given label.

Since DTW recognizes contextual changes and patterns over time, we hypothesized that the DTW score would be able to distinguish between the regular ups and downs of the housing market and meaningfully sharp increases and decreases.

Control Time Series

So that scores for each city could be generated against a uniform baseline, we generated a control time series for the housing and rental features selected. This control time series qualitatively represents the average trends in the features selected of cities considered to be gentrified. The first step to generating the control time series was to identify cities known to be experiencing gentrification. The list of cities was compiled based on two sets of research. The first, by realtor.com, identified gentrified areas based on increases in educational attainment, median household income, and median home value between the years 2000 and 2017.[13] The second study, performed by GoBankingRates identified gentrified cities based on one and five-year increases in home values and median household income for 500 cities. [14] We then combined the lists from the aforementioned research into a single, control list for cities known to be gentrified.[4]

A yearly average for each feature of all cities in the list of gentrified cities was calculated for each city in the list, indexed by city and state.

When looking at the cities included in our gentrification control, we can see sharp increases in housing and rental prices over short periods of time (usually 1-2 years) and then prices maintaining their value with small fluctuations, but never returning to their values before the sudden spike. This trend was mostly observed within the rental data. For housing data, the main identified trend was after recovering from the 2008 housing market crash the housing values for gentrified cities returned to its value from late 2006 - early 2007, or even surpassed it by a sizable amount.

D. Time Series Clustering

To further explore how these metrics may predict levels of gentrification, we used Time Series K-Means clustering on ZHVI and ZRI time series data sourced from the Zillow portal. While these were separate datasets from the main dataset used, they contained the same metrics. Therefore, we

[4] See Appendix II for the list of cities used in the construction of the control time series.

hypothesized that cities with similar DTW scores would end up in the same cluster.

Time Series KMeans was chosen over a more typical implementation of KMeans so that the temporal nature of the data would be accounted for. Since time series data represents a curve as opposed to a scatter plot, there are extra steps needed to make time series data clusterable.

First, the curves are translated into a series of horizontal line segments that represent the amplitudes at those given points in time. The points in the segments are then placed into a higher dimensional space, where the data points from the segments can be clustered using KMeans as it is typically understood.

In this research, data points were assigned to one of five clusters, each representing a different level of gentrification, ranging from no gentrification to high levels of gentrification.

The results from the clusters were used as a comparison against the DTW scores assigned. It was our hypothesis that cities with the same gentrification level label assigned by their DTW scores would be clustered together. Clustering was also performed as a way to explore the shape of the data set.

Results

DTW Scores for Rental Metrics

The results from the DTW scoring revealed some expected results, such as cities with sharper increases in ZRI over time had lower DTW scores. This is consistent with the aforementioned prior research.

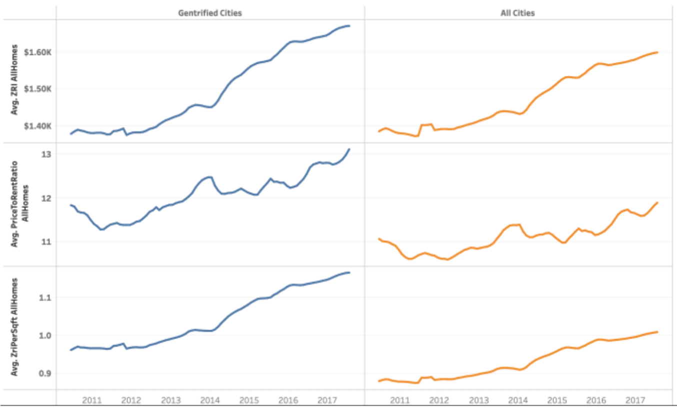
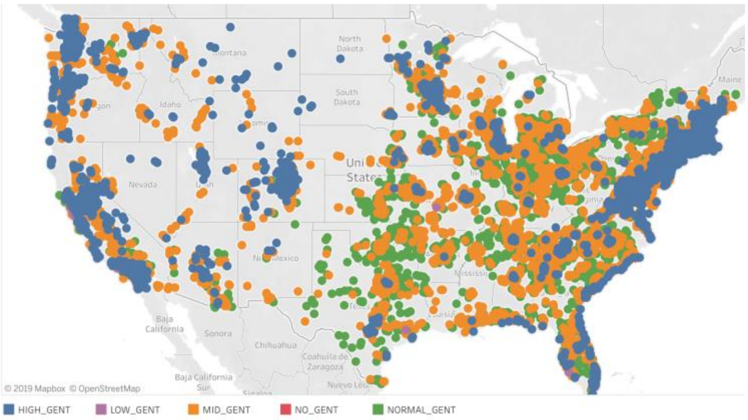


Fig. 4: Comparison of the Time Series for each metric between the gentrified cities and all cities from the dataset. The comparisons show that the gentrified cities experience much sharper increases across all metrics when the housing market experiences increases.

Figure 4 illustrates that our initial hypothesis was correct: Rental prices in gentrified cities experience sharper increases when the housing market experiences increases. This is especially prominent in the Avg_ZRIAllHomes comparisons: in 2014, the gentrified cities and all cities both had Average ZRI's for all homes near $145,000.00. However, in 2015, the Average ZRI for all homes in gentrified cities was approximately $158,000.00 whereas the average for all cities was only 150,000.00. Another interesting trend that can be identified from the figure is that gentrified cities appear to have recovered from the 2008 housing crisis at higher Price per square foot values and price to rent ratios.

Another interesting trend that is observed from the data is geographic distribution. It appears from the dataset that highly gentrified areas are concentrated in specific metro areas, with a great majority in larger metropolitan areas and coastal cities. This trend was observed in both the housing and rental metrics DTW scores.

Figure 5: A heatmap showing the geographical distribution of gentrified cities. Most cities identified to have a high level of gentrification are concentrated in



coastal areas and large, attractive metropolitan areas, such as Denver, CO, and Minneapolis, MN.

Most of these cities are part of larger metropolitan areas including the San Francisco Bay area, Wine Country, and Los Angeles. Most of the control cities were in California, which may partially account for this result; however, California is notorious for high rates of urban displacement which is also considered a major marker of gentrification. [15]

The results also showed that the DTW score was able to distinguish consistently expensive cities from gentrified cities. This is illustrated by the resulting scores for the cities with the lowest levels of gentrification:

	State	City	dtw_score	dtw_value	dtw_label
0	FL	Jupiter Island	3.905696	4.0	NO_GENT
1	CA	Atherton	2.849645	4.0	NO_GENT
2	CA	Beverly Hills	2.413206	3.0	LOW_GENT
3	CA	Belvedere	2.314848	3.0	LOW_GENT
4	FL	Palm Beach	2.299241	3.0	LOW_GENT
5	CA	Stinson Beach	2.118444	3.0	LOW_GENT
6	CA	Portola Valley	2.034014	3.0	LOW_GENT
7	CA	Hillsborough	2.028716	3.0	LOW_GENT
8	CA	Los Altos Hills	2.010883	3.0	LOW_GENT
9	FL	Gulf Stream	1.990482	3.0	LOW_GENT
10	CA	Malibu	1.964989	3.0	LOW_GENT

Figure 6: The cities with the lowest levels of gentrification according to the DTW scores for rental metrics. The city with the least amount of gentrification according to the DTW score is Jupiter Island, FL. Other notables in this list are Beverly Hills, CA and Palm Beach, FL.

Jupiter Island, FL had the highest DTW score, meaning the greatest distance from the gentrification control time series, despite having an average ZRI for all homes of $19,190.50 and an Average ZRI per square foot of over 4.0. Beverly Hills, CA is also on the list of least gentrified areas, yet has a median home price of over three million dollars. [17] However, Beverly Hills has been the home of celebrities and wealth since the 1920s and Jupiter Island was noted to be the most expensive place in the country in 1999. [18]

The implications of these findings are that the DTW scores calculated on these time series are able to distinguish between areas that are gentrified and those that are consistently expensive rental prices. In terms of finding cities that are experiencing gentrification, the list of the most gentrified cities was also consistent with the research:

	State	City	dtw_score	dtw_value	dtw_label
0	GENT_CONTROL	GENT_CONTROL	0.000000	0.0	HIGH_GENT
1	CA	Santee	0.064712	0.0	HIGH_GENT
2	CA	Rohnert Park	0.065260	0.0	HIGH_GENT
3	CA	Pinole	0.068630	0.0	HIGH_GENT
4	CA	Oceanside	0.070289	0.0	HIGH_GENT
5	MN	Excelsior	0.074732	0.0	HIGH_GENT
6	CA	Valinda	0.080531	0.0	HIGH_GENT
7	CA	Lakeside	0.080739	0.0	HIGH_GENT
8	CA	Cotati	0.081757	0.0	HIGH_GENT
9	CA	Concord	0.083737	0.0	HIGH_GENT
10	CA	El Cajon	0.087704	0.0	HIGH_GENT
11	MA	Wilmington	0.088586	0.0	HIGH_GENT

Figure 7: The cities with the highest levels of gentrification according to their low DTW scores. The GENT_CONTROL represents the baseline curve used and is not a city.

Consistent with Figure 5, the most gentrified areas according to the rental metrics show a majority in popular metropolitan areas in California. Santee, the most gentrified city shown in Figure 7 is a suburb of San Diego (as is El Cajon) which experienced a population growth of about 250,000 from 2010 to 2019. [19] These findings are shown to be consistent with our hypothesis since rapid population growth is also considered by social scientists to be a sign of gentrification.

DTW Scores for Housing Metrics

The Housing Sales Metrics were not able to be used, as there were too many nulls and pockets of nulls that could not be filled with backfilling. Regardless, the other metrics were able to provide DTW score calculations that were consistent with our hypothesis and were somewhat consistent with the results for rental metrics. There were some cities that appeared to be in the most and least gentrified areas in both the rental and housing metric calculations.

	State	City	dtw_score	dtw_value	dtw_label
0	CA	Atherton	5.424433	4.0	NO_GENT
1	CA	Los Altos Hills	4.585154	4.0	NO_GENT
2	CA	Hillsborough	4.393282	4.0	NO_GENT
3	CA	Portola Valley	3.905214	4.0	NO_GENT
4	NY	Water Mill	3.758229	4.0	NO_GENT
5	CA	Belvedere	3.557124	4.0	NO_GENT
6	CA	Woodside	3.312129	4.0	NO_GENT
7	CA	Beverly Hills	3.307819	4.0	NO_GENT
8	CA	Hidden Hills	3.251623	4.0	NO_GENT
9	CA	Monte Sereno	3.166060	4.0	NO_GENT
10	CA	Stanford	3.030621	4.0	NO_GENT

Fig. 8: The cities with the highest DTW scores for housing metrics. Consistent with the findings for rental, the least gentrified areas were consistently expensive areas in California.

The results of the DTW score calculations for the housing metrics also show the least gentrified areas to be cities that have always been expensive, such as Atherton, CA, a suburban area on the outskirts of San Francisco and Water Mill, NY, a popular vacation spot for the well-to-do on the south fork of Long Island. A noted difference between housing and rental data is the number of areas that are not considered to be gentrified. For rental metrics, there were only two cities considered to have no gentrification, whereas, with housing, there were nineteen. This appears to be a product of much higher DTW scores than the rental metrics, with a difference in about 1.5 between Jupiter Island, FL for rentals, and Atherton, CA for housing.

The cities found to be the most gentrified in terms of housing were concentrated mostly in California and Massachusetts suburbs. It should be noted that consistent with Figure 5, While the California cities were spread across various metropolitan areas, the Massachusetts suburbs identified were consistently close to Boston. Furthermore, these cities identified are all home to at least one public train stop. There were also two cities in Fairfax County, Virginia, a wealthy suburb of Washington, DC. The county is also home to multiple public train stops. These findings are consistent with studies that indicate new public transportation projects, including those in San Diego and Washington, DC being a cause of gentrification. [20]

	State	City	dtw_score	dtw_value	dtw_label
0	GENT_CONTROL	GENT_CONTROL	0.000000	0.0	HIGH_GENT
1	MA	Medford	0.088412	0.0	HIGH_GENT
2	MA	Stoneham	0.088817	0.0	HIGH_GENT
3	CA	Oceanside	0.092052	0.0	HIGH_GENT
4	CA	Santee	0.097922	0.0	HIGH_GENT
5	VA	Centreville	0.100390	0.0	HIGH_GENT
6	VA	Lorton	0.102779	0.0	HIGH_GENT
7	CA	El Cajon	0.103188	0.0	HIGH_GENT
8	MA	Woburn	0.104216	0.0	HIGH_GENT
9	CA	Rohnert Park	0.108622	0.0	HIGH_GENT
10	CA	West Carson	0.113171	0.0	HIGH_GENT
11	MA	Saugus	0.114105	0.0	HIGH_GENT

Fig. 9: The cities with the lowest DTW scores for housing metrics.

One interesting finding regarding the cities in figure nine is an inconsistency in population growth. For example, Centreville, VA experienced a population growth of 4.81% from 2009 to 2019, while Medford, MA only experienced a population growth rate of 0.46% in the same span of time.[21]

A. DTW KMeans

The results were consistent with the hypothesis that the labeled data would be clustered together. However, the shapes discovered were surprising.

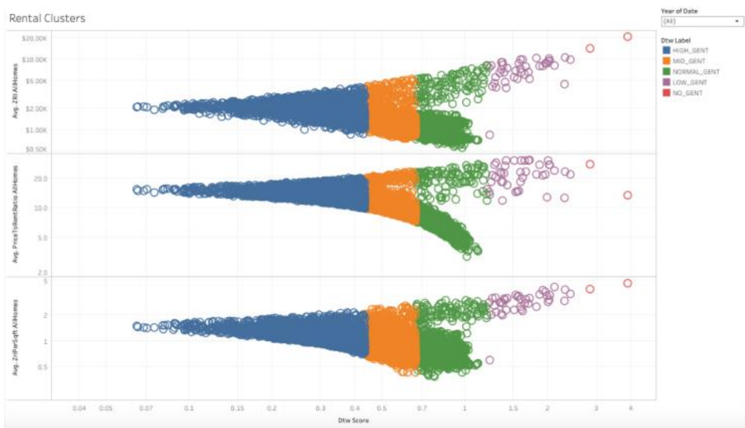


Fig. 9: The DTW clusters for rental data. The HIGH_GENT cities

The shapes of clusters found using DTW KMeans on both the housing and rental features revealed band-like structures for all metrics. For all metrics in both cases, the HIGH_GENT and MID_GENT cities had the densest clusters, while the LOW_GENT and NO_GENT clusters were much sparser. In the case of NORMAL_GENT, a prominent horizontal partition appeared horizontally. As DTW scores got larger and smaller, the clusters became denser. This is to be expected, as higher and lower scores indicate more extreme cases of gentrification or areas where there is none occurring at all.

DISCUSSION

The results achieved confirmed a strong correlation between sharp increases in housing costs and gentrification across both rental and housing metrics. These results were consistent with the prior research that suggested the correlation. The cities that were identified as highly gentrified by the algorithm were correctly classified according to the previous research and all exhibited hallmarks of gentrification. However, in the case of both housing and rental metrics, the top cities all exhibited different hallmarks of gentrification with little consistency. For example, San Diego, whose suburb, Santee appeared in both experiments has experienced both rapid population growth and increases in public transportation availability. However, Pinole, CA, a suburb of San Francisco which appeared as one of the top gentrified areas in both experiments only experienced a growth of 0.30% since 2009. [21] Additionally, Centreville, VA which appeared only in the list for housing, has experienced rapid population growth, only appeared in the list for housing metrics. Despite the very different markers of gentrification for the top cities, the prior research indicates that the algorithm accurately predicts levels of gentrification, especially among highly gentrified areas and non-gentrified areas. The inconsistency of non-housing factors could indicate that sharp increases in rental and housing prices are the result of many possible underlying factors that contribute to gentrification. Essentially, these sharp increases in market values could be the end result of a culmination of these many factors including population growth, transportation, jobs, amenities, and even Starbucks coffee shops. [22]

While the DTW scores accurately predicted gentrification of cities according to the prior research, cities are large and contain smaller communities within which all may experience gentrification differently. While our methodology paints a picture of the net gentrification of a city, it is limited in the sense that it cannot look at the housing behavior of smaller communities within cities, such as the boroughs of NYC. Were the results more granular, more information about the underlying factors contributing to the increased housing and rental prices could be more apparent.

Despite the more granular causes not being very apparent, there were some large-scale trends in cities that were highly gentrified, such as being coastal and highly urbanized. One less obvious feature is being an economic center for technology. As an example, many of the California cities marked as HIGH_GENT cities are in the San Francisco Bay area and the Massachusetts cities within commuting distance to Boston. Both of these areas are two of the top economic areas for technology and were consistently ranked more gentrified by the DTW scores. This may indicate that a quick influx of technology workers may be a strong indicator of gentrification and show that investment in a city’s technology sectors may create an attractive incentive to move there, though possibly at the expense of the current residents.

Ultimately, the goal of the research to discover a way to quantify gentrification was met. The scores generated are significant in that they provide a way to measure and compare gentrification at the city level in multiple housing markets and allow for the extrapolation of possible underlying factors that contribute to the higher housing and rental prices. These scores provide an excellent starting point for further research into more granular studies on gentrification and show that temporal analysis of changes in a community is essential to gaining insight about gentrification.

FUTURE RESEARCH

As stated, the algorithms and scores generated provide an excellent starting point for further research about gentrification at a more granular level. Since many studies have been done on neighborhoods, future work could use DTW on time series data at the neighborhood level. A study at the neighborhood level compared to this study could provide insight into whether gentrification occurs in the same way at a smaller scale.

Another possible experiment may be to continuously track and update the DTW scores found here in the next ten or twenty years to see how they change. As technology jobs grow and become more common, it may be possible that this becomes less of a contributing factor. One observation from the data revealed that many houses are currently priced the same or higher than before the 2008 recession. It may be interesting to see how the rates of gentrification change if another recession occurs.

CONCLUSION

The research ultimately achieved its goal of discovering a way to quantify gentrification in American cities. This score was produced using Dynamic Time Warping (DTW) to discover the net difference between a given city's housing and rental metrics time series and a control time series comprised of cities known to be gentrified. In addition to providing a score that was accurate based on the previous research, the scores provided insight into some of the underlying causes of gentrification, most prominently being sharp increases in housing or rental prices caused by a number of possible factors including population growth, public transportation, and technology jobs. The scores and algorithms used in this research provide a very useful starting point for continued research into gentrification by providing a quantified measure of gentrification.

REFERENCES

[1] Department of Health and Human Services, "Health Effects of Gentrification",Center for Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/healthyplaces/healthtopics/gentrification.htm.

[2] R. FLORIDA, "No One's Very Good at Correctly Identifying Gentrification", *https://www.citylab.com*, 15 Dec. 2014. [Online]. Available: https://www.citylab.com/equity/2014/12/no-ones-very-good-at-correctly-identifying-gentrification/383724

[3] T. MISRA, "Mapping Gentrification and Displacement in San Francisco", *https://www.citylab.com*, 31 Aug. 2015. [Online]. Available: https://www.citylab.com/equity/2015/08/mapping-gentrification-and-displacement-in-san-francisco/402559/.

[4] "Zillow Economics Data - Turning on the lights in housing research.", *https://www.kaggle.com*, 24 Oct. 2017. [Online]. Available: https://www.kaggle.com/zillow/zecon.

[5] G. Ryan, "The Impact of Rapid Gentrification on Rents", Bachelor's Thesis, Department of Economics, Pace University, New York, 2016. [Online]. Available: https://www.researchgate.net/publication/318298803_The_Impact_of_Rapid_Gentrification_on_Rents.

[6] "tslearn.metrics.dtw", *https://tslearn.readthedocs.io*, 2017. [Online]. Available: https://tslearn.readthedocs.io/en/latest/gen_modules/metrics/tslearn.metrics.dtw.html?highlight=dtw.

[7] Y. Bun, "Zillow Rent Index: Methodology", *https://www.zillow.com*, 12 Mar. 2012. [Online]. Available: https://www.zillow.com/research/zillow-rent-index-methodology-2393/.

[8] Zillow, "Zillow Home Value Index: Methodology", *Zillow Research*, 3. Jan 2014, [Online]. Available: https://www.zillow.com/research/zhvi-methodology-6032/.

[9] D. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series", Information Systems Department, Stern School of Business, New York University, New York, AAAI Technical Report WS-94-03, 26 Apr. 1994, [Online]. Available: https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf.

[10] B. Robinson, "Residential Home Types: 17 Clarifying Home Definitions", *www.zillow.com*, 2 Jan 2018. [Online]. Available: https://www.zillow.com/blog/residential-home-types-definitions-223701.

[11] C. Fuller, "Affordability Tears", *www.zillow.com*, 2 May 2015. [Online]. Available: https://www.zillow.com/research/home-affordability-by-price-tier-9547.

[12] "sklearn.preprocessing.MinMaxScaler", *https://scikit-learn.org*, n.d., [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[13] Y. Pan, "The U.S. Cities That Are Gentrifying the Fastest—You'll Never Guess No. 1", *www.realtor.com*, 23 Jan. 2017. [Online]. Available: https://www.realtor.com/news/trends/10-surprising-cities-that-are-gentrifying-the-fastest.

[14] C. Measom, "New Study Shows the Huge Impact Gentrification Is Having on Residents in These Cities", *www.gobankingrates.com*, 13 Feb 2019. [Online]. Available: https://www.gobankingrates.com/making-money/economy/most-gentrified-cities-in-us.

[15] Urban Displacement Project, "Urban Displacement Project - Executive Summary", University of California Berkeley, Berkeley, CA, Dec. 2015. [Online]. Available: https://www.urbandisplacement.org/sites/default/files/images/urban_displacement_project_-_executive_summary.pdf

[16] "Local Market Reports", *www.zillow.com*, n.d. [Online]. Available: https://www.zillow.com/research/local-market-reports/

[17] "Beverly Hills, California", *https://www.bestplaces.net*, n.d. [Online]. Available: https://www.bestplaces.net/cost_of_living/city/california/beverly_hills

[18] R. Nolin, "The Nation's Most Costly Real Estate", Sun Sentinel , July 6, 1999. [Online]. Available: https://web.archive.org/web/20170403004704/http://articles.sun-sentinel.com/1999-07-06/news/9907050396_1_median-home-price-worth-magazine-jupiter-island

[19] Z. Self, "San Diego sees large population growth from 2010 to 2018", ABC News 10, April 22, 2019. [Online]. Available: https://www.10news.com/news/local-news/san-diego-sees-large-population-growth-from-2010-to-2018

[20] J. Turrentine, "When Public Transportation Leads to Gentrification", *www.nrdc.org*, 1 June 2018. [Online]. Available: https://www.nrdc.org/onearth/when-public-transportation-leads-gentrification

[21] Open Data Network, *www.opendatanetwork.com*, n.d. [Online]. Available: https://www.opendatanetwork.com

[22] E. Glaeser and H. Kim and M. Luca, "Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change", Working Paper 18-077, Harvard Economics Department and Harvard Business School, Cambridge, MA, c.a. 2018. [Online]. Available: https://www.hbs.edu/faculty/Publication%20Files/18-077_a0e9e3c7-eceb-4685-8d72-21e0f518b3f3.pdf

APPENDIX

I: CORE METRICS

Core Rental Metrics

ZRI (All Homes)	A smoothed seasonally adjusted measure of the median estimated market rate rent across a given region and all housing types.
ZRI (Single Family Homes)	A smoothed seasonally adjusted measure of the median estimated market rate rent across a given region and single-family homes.* ** A single-family home is a standalone house built for one family. If a structure includes more than one collection of living spaces with separate entrances and privacy, it's a multi-family home. [10]*

ZRI Per-Square ft.	Median of the estimated monthly rent price of all homes, per square foot. This is calculated by taking the estimated rent price for a home and dividing it by the home's square footage.
Price-To-Rent Ratio	This ratio is first calculated at the individual home level, where the estimated home value is divided by 12 times its estimated monthly rent price. The median of all home-level price-to-rent ratios for a given region is then calculated.

Core Housing Metrics

ZHVI (All Homes)	A smoothed seasonally adjusted measure of the median estimated home value across a given region and all housing types.
ZHVI Per Square ft.	Median of the value of all homes per square foot. This number is calculated by taking the estimated home value for each home in a given region and dividing it by the home's square footage.

ZHVI (Tiers)	Zillow defines 3 housing affordability tiers (Bottom, Middle, and Top), and provides separate ZHVI indexes for each tier group.* **Zillow calculates affordability with rentals and home values with data that was not included in this data set but does provide their methodology for determining the affordability tiers within a particular location. [11]*

(Continued on next page)

Sales (Seasonally Adjusted)	A seasonally adjusted measure of the number of homes sold during a given month in conservatively defined consumer-to-consumer transactions adjusted for latency.
Inventory (Seasonally Adjusted)	A seasonally adjusted measure of the median weekly snapshot of for-sale homes within a region for a given month.
Percentage of Homes Increasing in Value	The percentage of homes in a given region with values that have increased in the past year.

Percentage of Homes Decreasing in Value	The percentage of homes in a given region with values that have decreased in the past year.

II: Gentrification Control Cities

City	State
Asheville	NC
Nashville	TN
Oakland	CA
Charleston	SC
Anaheim	CA
Berkeley	CA
Seattle	WA
Austin	TX
LosAngeles	CA
SanDiego	CA
Midland	TX
Washington	DC
Portland	OR
Sacramento	CA
NewYork	NY
RoyalOak	MI
Bentonville	AR
CostaMesa	CA
SanMarcos	CA
AnnArbor	MI
JerseyCity	NJ
Somerville	MA
Thornton	CO
Vista	CA
LongBeach	CA
Pittsburgh	PA
Quincy	MA
Napa	CA
Hillsboro	OR
Denver	CO
Hayward	CA

III: TABLEAU VISUALIZATIONS

Visualizations for Housing Metrics:
https://public.tableau.com/profile/brian.kalinowski#!/vizhome/Zillow_Housing_Viz/DTWClusteringMetricsOverTime

Visualizations for Rental Metrics:
https://public.tableau.com/profile/brian.kalinowski#!/vizhome/Zillow_Rentals_Viz/CitiesRentalMetricsMap