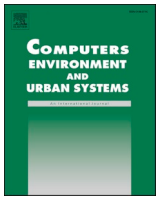
Computers, Environment and Urban Systems 88 (2021) 101658

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



The language of neighborhoods: A predictive-analytical framework based on property advertisement text and mortgage lending data

Elizabeth C. Delmelle *, Isabelle Nilsson

Department of Geography and Earth Sciences, University of North Carolina at Charlotte, United States of America

ARTICLE INFO

Keywords:
Text analysis
Real estate
Neighborhood typology

ABSTRACT

Real estate property listings use specific language to market properties to a target buyer – typically one that will garner the largest profit. As home-seekers have different preferences for house characteristics and neighborhood amenities, the words used to advertise homes are expected to vary according to the type of neighborhood and expected homebuyer. In this article, we develop a framework for extracting the key characteristics used to advertise properties according to the racial and income profile of home mortgage applicants in different types of neighborhoods. We perform an exploratory text analysis on words according to neighborhood types and use a binomial logistic regression model to determine the most discriminatory words for each type of neighborhood. Finally, we assess the ability of the property listing text to predict the type of neighborhood the property belongs to. Using a small, illustrative case study of listings from Charlotte, North Carolina, we find that the presence of specific neighborhood names holds more importance in neighborhoods with primarily White homebuyers. In gentrifying neighborhoods, unique property characteristics such as parquet flooring, and words associated with revitalization near the city center are common. Listings in neighborhoods with minority homebuyers are less likely to mention schools and feature traditionally suburban descriptors such as cars, garage, and roadways. We envision that this framework, using near real-time data sources, holds the potential to advance neighborhood prediction efforts, our understanding of amenity preferences and sorting patterns, and to illuminate less visible processes of change such as discrimination in the housing market.

1. Introduction

Processes of neighborhood change are largely driven by residential locational choices, or the in- and out- migration of people of different demographic, racial, and socioeconomic status. These decisions are made based on the attractiveness of the housing stock of a neighborhood, its available amenities, its racial and income makeup, the purchasing power of the home-seeker, and also, for many, structural constrains inherent in the housing market that enable or prevent them from realizing their ideal location or home. The passing of the Fair Housing Act of 1968 and its subsequent strengthening in 1988 made racial discrimination in the housing market illegal. Its passage offered optimism that the spatial patterns of segregation carved out through years of legal racist housing practices on behalf of the federal government, banks, and real estate agents would begin to dissipate. Contrary to these expectations, stark patterns of segregation by race and income remain throughout most cities across the United States (Connor, Gutmann, Cunningham, Clement, & Leyk, 2020; Sampson, 2012) and discrimination continues to be a common practice in multiple forms among real estate agents (Galster & Godfrey, 2005; Zhao, Ondrich, & Yinger, 2006).

Despite the durability of concentrated, race-based poverty in many neighborhoods across the United States, a fundamental shift in urban sorting is underway as historical patterns of urban, center city declines and suburban prosperity, largely byproducts of racist housing practices, have begun to reverse themselves as wealthy and White residents have increasingly returned to the urban core (Ehrenhalt, 2012; Florida 2016). Explanations behind this reversal have centered on the importance of urban cultural amenities in attracting wealth back to city centers (Borgoni et al., 2020). However, consumer preferences alone cannot fully explain residential sorting outcomes (Roscigno, Karafin, & Tester, 2009). Realtors, landlords, and financial institutions can serve as gatekeepers to neighborhoods through intentional or unintentional practices (Besbris & Faber, 2017; Roscigno et al., 2009).

While a body of research has investigated more deliberate acts of realtor discrimination, especially via residential steering practices

* Corresponding author.
E-mail address: edelmell@uncc.edu (E.C. Delmelle).

https://doi.org/10.1016/j.compenvurbsys.2021.101658
Received 14 December 2020; Received in revised form 11 May 2021; Accepted 13 May 2021
Available online 20 May 2021

(Besbris & Faber, 2017; Galster & Godfrey, 2005; Massey, 2005), in this article, we focus on the words used by real estate agents to market properties and we explore the potential of these advertisements for understanding the connection between the language used and the racial and income profile of home mortgage applicants in a neighborhood. Using a small, proof of concept dataset, we develop a framework for extracting the key amenities and home characteristics associated with different types of neighborhoods, and test how well these words predict neighborhood types. Specifically, our proposed approach uses two novel datasets whose temporal resolution enables changes to be understood in near real-time, rather than waiting for retrospective survey data to be released, as has been the norm in studies of neighborhood changes and residential sorting. We begin by classifying census tracts into five groups according to the racial and income profile of mortgage applicants and the 5-year change in these characteristics using data from the annually updated Home Mortgage Disclosure Act (HMDA) from 2013 to 2018. We then perform a text analysis on the words associated with each of the neighborhood classes using a 1-month sample (September, 2019) of property advertisements obtained from Zillow for the city of Charlotte, NC. Finally, we train a machine learning algorithm to determine how well the full text listing predicts a neighborhood's class.

Our results illustrate distinct language used according to the racial and income profile of homebuyers in certain neighborhoods. Words associated with an increasing share of White homebuyers in predominantly minority neighborhoods featured unique home characteristics (e.g. parquet flooring), and signal areas in high demand to "qualified" buyers, potentially pointing to restrictive language. Words like "welcoming" and "families" were negatively associated with this neighborhood type. Neighborhoods where the majority of homebuyers were either Black or Hispanic were much less likely to mention the specific neighborhood name, suggesting a lack of prestige compared to those with a majority of White homebuyers. They were also negatively associated with schools and more likely to mention words associated with more suburban traits. Our predictive results demonstrated the potential for these datasets and methods to capture the characteristics of the neighborhood based on the property listing text and thus hold promise as a predictive technique for understanding changes before residential movements have taken place, serving as an early warning system for neighborhood changes (Chapple & Zuk, 2016).

The remaining structure of this paper is as follows: we proceed in Section 2 with a background section on residential sorting, the shifting importance on urban amenities, and the role of real estate brokers in the sorting process. We then describe our study area, data and methods in Section 3 followed by our results in Section 4 and conclusions in Section 5.

2. Background

2.1. Residential sorting and amenities

Understanding how neighborhoods change according to their racial, demographic, socioeconomic, and housing characteristics has been a longstanding subject of inquiry. While theories abound on their causes and consequences, ultimately changes can be ascribed to either a shift in the characteristics of those moving into and out of neighborhoods or, less often, to a change in the characteristics of existing residents, such as improved socioeconomic circumstances (Van Criekingen & Decroly, 2003). Therefore, understanding residential mobility and location choice, or sorting patterns, is key to comprehending neighborhood changes.

One mechanism for understanding residential sorting patterns is through an amenity-based theoretical framework. There is an established body of literature that explains the propensity for households to sort themselves into homogenous clusters according to their preferences for certain amenities (Tiebout, 1956). Collectively, this research has been used to explain recent shifts in urban sorting patterns away from a poor and minority core and wealthy and White suburbs – a simplified dichotomy that described many post-war American cities. The more recent evidence points to a perceptible shift in preferences for vibrant, urban neighborhoods, especially among the younger, college-educated population over the past several decades (Glaeser & Gottlieb, 2006, Baum-Snow & Hartley, 2020, Lee, Lee, & Shubho, 2019; Couture & Handbury, 2017). The presence of creative-cultural amenities (bars, art galleries, retail, etc.) in walkable center-city neighborhoods are thought to serve as a harbinger for socioeconomic or demographic transformations (Bereitschaft, 2014, Nilsson & Reid, 2019) and their inclusion in property advertisements could be indicative of a target young, college graduate demographic who favors these sorts of non-tradable consumption amenities (Couture & Handbury, 2017).

Empirical findings on amenity preferences and residential sorting do not necessarily point to a full reversal of sorting patterns – there continues to be strong demand among wealthy Whites for suburban housing (Frey, 2017). Lifecycle changes may explain this simultaneous demand as smaller, walkable urban locations are desirable among younger, childless residents, but once they settle down and start a family, movement to larger suburban homes with different amenities remains common (Beamish, Goss, & Emmel, 2001; Lee, 2021). The amenities available in newer suburban developments have been argued to serve as a means for establishing a more homogeneous (often Whiter and wealthier) population than would have occurred otherwise. In this exclusionary amenity situation, developers select amenities aimed at attracting a certain demographic and residents of the development are required to pay a fee for them in order to reside in the neighborhood (Strahilevitz, 2006). Golf courses are a common example of such an amenity.

Rather than a monocentric decline or ascent in socioeconomic status away from an urban core, the distribution of natural and man-made amenities around urban areas have given rise to a spatially fragmented pattern of wealth concentrated around cities (Delmelle, 2019; Florida & Adler, 2018; Lee, Irwin, Irwin, & Miller, 2021). However, the empirical knowledge on which amenities might drive sorting behaviors remains limited by the availability of high-resolution spatial datasets on all possible endogenous and exogenous amenities (Lee et al., 2019). Our approach circumvents this challenge by extracting the specific amenities targeted to homebuyers in advertisements thereby eliminating the need to collect and test various spatial variables.

2.2. Realtors and housing markets

The problem with relying solely on amenity-based theories for explaining residential sorting and subsequent neighborhood changes is that it ignores discriminatory practices in the housing market that serve to prevent some residents from fully realizing their preference for certain neighborhoods or amenities. Prior to the passage of the 1968 Fair Housing Act, denying access to housing in certain neighborhoods based on race was a legal and widespread practice performed by real estate agents (Massey, 2005).

While the 1968 Act legally banned this practice, new mechanisms for racial discrimination that sidestepped the law became common. For instance, audits on whether prospective homebuyers are shown different sets of properties in different neighborhoods according to their race have revealed that residential steering practices are common among realtors (Besbris & Faber, 2017; Galster & Godfrey, 2007; Zhao et al., 2006). Real estate agents have therefore been labeled as gatekeeps to certain neighborhoods, both responding to and shaping markets, and are charged with helping to perpetuate patterns of racial segregation that were anticipated to dissipate with the passage and strengthening of fair housing laws (Besbris & Faber, 2017; Pearce, 1979). Galster and Godfrey (2007) showed that minority home-seekers were disproportionately directed towards lower-class and by extension, lower-amenity neighborhoods with worse school quality and higher crime rates. These practices have undoubtably helped lead to the observed outcome that

predominantly Black, high-amenity neighborhoods are in short supply across the United States (Bayer & McMillan, 2005). Therefore, to live in high-amenity neighborhoods, Black homebuyers must normally live with a higher share of White residents.

There is a body of research that has investigated the role that real estate agents play in perpetuating discriminatory practices, especially in the form of residential steering practices, both quantitatively and qualitatively (Besbris & Faber, 2017; Galster & Godfrey, 2007; Zhao et al., 2006). The motivation behind these behaviors are generally not to advance cycles of metropolitan segregation, but rather to maximize realtors' profits. For this reason, realtors tend to cluster in space near higher-priced and high-demand local markets and avoid working in lower-income neighborhoods where commissions are lower. This is true regardless of the race of the realtor (Besbris & Faber, 2017). A lesser studied element of this process is the role that the listings themselves play in perpetuating these sorting patterns. A few studies have examined listings from different vantage points. Galster, Freiberg, and Houk (1987) analyzed print property advertisements in White, integrated, and Black neighborhoods and found that properties were advertised less favorably in Black neighborhoods. Pryce and Oates (2008) performed a linguistic analysis on listings suggesting that the style of writing varied over space and that there may be a local dialect in real estate listings. Delmelle, Nilsson, and Schuch (2020) performed an exploratory analysis on property advertisement text along a newly opened light rail line to help disentangle the role of rail from other advertised amenities in spurring adjacent neighborhood changes. Light rail was marketed most prominently in previously gentrified neighborhoods, close to the center city alongside other creative-cultural amenities.

Based on evidence on the shifting importance of amenities in residential sorting and the role of real estate brokers in potentially serving as guides in seeking to attract a particular type of homebuyer to a property, we can hypothesize that property listing texts will differ according to the racial and income profile of homebuyers and the type of neighborhood the property is listed in. We expect creative-cultural amenities associated with more urban living to be more prominent in gentrifying neighborhoods (where wealthier White homebuyers are purchasing properties in formerly minority and poorer neighborhoods). We expect differences in advertisements of more suburban-type homes and neighborhoods according to race and income with the potential for exclusionary amenities in wealthier and Whiter suburbs and a potential lack of amenities or signals of disinvestment in poorer and minority suburbs.

3. Case study, data, and methods

Our case study city for demonstrating our proposed framework is Charlotte, North Carolina, the largest city in the state and the fifth fastest growing city in the country over the past decade. According to Census estimates, the city's population grew 21% from 731,400 in 2010 to 885,700 in, 2020 (Chemtob & Off, 2020). Accompanying this growth are strong gentrification pressures in neighborhoods closest to the city's core, locally referred to as 'uptown' – the presence of a new light rail line, a walkable environment, breweries and restaurants have all helped to accelerate racial and income shifts in the population of these neighborhoods (Delmelle, Nilsson, & Schuch, 2020). However, rising demand and housing costs in and around the urban core has also led many to seek housing in the city's suburbs and outlying towns where population growth rates have been higher (Chemtob & Off, 2019). The dynamics has led to an at least partial suburbanization of poverty in the city (Delmelle, Nilsson, & Adu, 2020).

3.1. Data and methods

3.1.1. Defining types of neighborhoods

We begin by classifying neighborhoods (Census Tracts) according to the racial and income profile of loan applicants and the types of neighborhoods in which they are applying for a loan. To do so, we use data from the publicly available Housing Mortgage Disclosure Act (HMDA). We include the racial, ethnic, and income composition of loan applicants in 2018 (share of White, Black, and Hispanic applicants, and median income), and the change in these shares from 2013 to 2018. The HMDA data are available at the individual loan application level, but include information on the Census tract of the home in which the applicant is applying for a mortgage. Therefore, to derive Census-scale variables from this data source, we aggregated or grouped-by the Census tract to derive average race and income values. We also include the neighborhood's minority population in 2013 and change in minority population between 2013 and 2018. The minority share variables originally come from the US Census, but are included with the HMDA dataset.

We used a k-means classification procedure on these 10 variables for the 233 census tracts in the county. All variables were normalized prior to running the k-means algorithm by creating a z-score with a mean of zero. This is a necessary step in the clustering algorithm as the Euclidean distance is first computed between all variables to construct a difference matrix. Therefore, all variables need to be on the same measurement scale as to not give disproportionate weight to variables on different scales. The algorithm was run for $k = 2$ to $k = 25$ clusters, and their results were compared initially using a set of fit statistics. These statistics suggested that two clusters were optimal in maximizing similarity within and minimizing overlap between cluster solutions. However, based on our in-depth knowledge of the study area, this separation essentially divided the county along White and Black racial divisions, overlooking the nuances of the housing market that we are aware of. This becomes more evident following the text analysis that identifies specific neighborhoods associated with each neighborhood type, as will be discussed in the results section. Tradeoffs between the parsimony of fit statistics and level of detail in resulting clusters have been acknowledged in the neighborhood geodemographic literature before (Kang, Rey, Wolf, Knaap, & Han, 2020). As a result, we landed on a five-clustering solution to segment neighborhoods in the county.

3.1.2. Text analysis

Our sample property description dataset is obtained from the website Zillow of all properties listed for sale in Charlotte, NC in September of 2019. September tends be a month that sits between the summer peak and the slowdown that occurs in the real estate market during the winter months (NAR, 2019). In Charlotte, NC, which has been a sellers' market for the past decade with a relatively low supply of housing compared to demand (Childress Klein Center for Real Estate, 2020), the median days on market as well as the median list to sold price during the quarter July-September 2019 stayed relatively constant (Realtor.com, 2021) as well as the number of homes sold (Childress Klein Center for Real Estate, 2020). However, it might be that different types of homes are listed during the fall than during the peak of summer. Our small, one-month sample therefore serves to illustrate our proposed approach and may not be a truly representative sample of Charlotte's housing market.

The addresses of the collected listing were geocoded using ESRI's ArcPro to assign property records a neighborhood typology. The resulting dataset contained 8448 property records. The text property description data were cleaned and analyzed using the TidyText package in R (Silge & Robinson, 2017). All stopwords and words generic to properties were removed such as 'bathroom' or 'bedroom'. Several words were combined to improve the interpretation of results – for instance, 'light' and 'rail' were merged to 'lightrail' and names of popular neighborhoods or landmarks were also combined in the text editing phase. We ran the models with and without city specific neighborhood, street, and amenity names in the data. The purpose of running the model with city specific neighborhood, street and amenity names was to determine whether our segmentation and predictions make sense in terms of what neighborhood and street names appear as important predictors in each cluster. To evaluate whether mentioning

neighborhoods, street names, and certain amenities differs according to neighborhood type, we substitute the specific name with a generic 'neighborhoodname' and 'streetname' placeholder. For city specific amenities such as names of certain restaurants, cafés, or entertainment establishments, we substitute the specific names with placeholders such as 'restaurant', 'cafe', 'entertainment', etc. Our hypothesis is that the marketing of the neighborhood by name will vary from cluster to cluster and be of particular importance for "up-and-coming" and higher income neighborhoods.

There are several popular machine learning methods that have been applied to text for the purpose of prediction. For this case study, we opted to fit a binomial logistic regression model combined with a LASSO regularization method for selecting variables, also referred to as glmnet in R (for lasso and elastic-net generalized linear models). Our selection was guided by the desire both to understand the predictive power of words in describing neighborhood types and to determine the relative importance of each word in discriminating between types. This latter consideration made glmnet preferable to methods that do not indicate the importance of individual words. Furthermore, the choice of a binomial function rather than a multinominal specification was motivated by the desire to know what words distinguishes one neighborhood type from all other neighborhood types (rather than a base case). In terms of algorithm accuracy for text classification, research has suggested that glmnet performs similarly to Support Vector Machines and Random Forests, two other popular approaches (Jurka, Collingwood, Boydstun, Grossman, & van Atteveldt, 2013). The glmnet package in R is used to estimate the model using a penalized maximum likelihood approach (Hastie & Qian, 2014).

Our estimations began by splitting the data into a training and testing set, using a 75% vs. 25% split. We opted for this larger testing set given the rather small sample size of 8448 property listings. A binomial logistic regression with LASSO regularization was fitted for each class (or cluster) using the training sample. Using the parameters of the fitted model and the testing sample, we made predictions with regards to neighborhood type.

4. Results

4.1. Types of neighborhoods

Basic descriptive statistics of the initial unnormalized variables used in the *k*-means procedure are shown in Table 1, below.

The resulting five cluster solution to our *k*-means analysis is described in Fig. 1 and is summarized as follows:

- *White-Higher-Income*: The 2nd highest share of White applicants (75%) and the 2nd highest median income ($105,530). These neighborhoods had the second lowest minority population in 2013.
- *White Homebuyers-Minority Neighborhoods*: In 2018, 68% of mortgage applicants were White in these neighborhoods, but in 2013, they were predominantly minority neighborhoods (77% minority). The share of Black applicants dropped by 8% between 2013 and 2018 and incomes rose.
- *Increasing Black-Minority Neighborhoods*: The largest share of Black applicants in 2018 (39%), an increase of 8.5% from 2013 to 2018. The share of White applicants declined by 12% during that time.
- *White-Increasingly High Income*: Where the first cluster was second with respect to income and percent Whites, this one is first. Highest share of Whites, very few Black or Hispanics. The highest median incomes, and the greatest increase in median incomes.
- *Hispanic Homebuyers-Minority Neighborhoods*: 35% of applicants in 2018 were Hispanic. These neighborhoods have the lowest median incomes and the highest minority population (80%).

The names of the neighborhood types attempts to highlight the main characteristics of each group – the inclusion of 'increasing' signifies that change was a defining characteristic in our interpretation, but does not necessarily imply that the other groups are static.

Table 1
Descriptive statistics of raw variables used in K-means clustering.

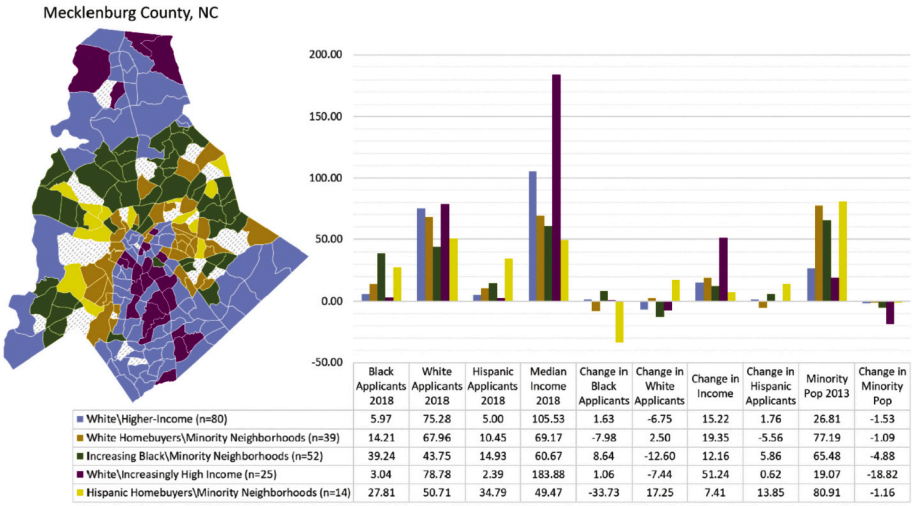
Variable	Source	Mean	S.D.	Min	Max
% Black Applicants 2018	HMDA[1]	0.17	0.17	0	0.73
% White Applicants 2018	HMDA	0.64	0.18	0.20	1.00
% Hispanic Applicants 2018	HMDA	0.10	0.11	0	1.00
Median HH Income 2018 (In thousands)	HMDA (US Census)[2]	93.25	46.86	32	335
Percent Minority[3] Residents in Tract, 2013	HMDA (US Census)	48.52	27.93	3.23	98.39
% Change in Black Applicants 2013–2018	HMDA	−0.84	15.97	−100	50
% Change in White Applicants 2013–2018	HMDA	−4.96	16.85	−63.63	85.71
% Change in Hispanic Applicants 2013–2018	HMDA	2.08	9.08	−42.85	42.85
Change in Median Income 2013–2018	HMDA	1895.95	1878.18	−4400	8700
% Change in Tract Minority Population 2013–2018	HMDA (US Census)	−4.26	18.34	−111.21	54.59

[1] HMDA = Home Mortgage Disclosure Act.
[2] Indicates that the data originated from the US Census, but was included with the HMDA data.
[3] HMDA defines minority as all races other than Whites and Whites of Hispanic or Latino origin.

With respect to the spatial distribution of the neighborhood groups shown in the map in Fig. 1, historically, the city of Charlotte (whose core is directly in the center of the Mecklenburg county map) has often been characterized by a dichotomous wedge of wealth, and crescent of poverty – this wedge can be depicted from the map in Fig. 1 as neighborhoods to the southeast of the center fall in either the White-Higher-Income or White-Increasingly High-Income category. The traditional crescent was concentrated around the city center in neighborhoods primarily to the west, however, recent gentrification trends can be depicted from the map as the White Homebuyers-Minority Neighborhood class now describes neighborhoods immediately south of the city center - this follows the first light rail line constructed in the city where significant redevelopment has taken place (Delmelle, Thill, Furuseth, & Ludden, 2013) and continues in a middle ring around to the edge of the eastern-most side of the wedge. These are places where gentrification pressure is strongest. Increasingly, Black-Minority Neighborhoods are now concentrated away from the core in older suburban neighborhoods, while Hispanic Homebuyers-Minority Neighborhoods tend to occupy transitional spaces between gentrifying and predominantly Black neighborhoods.

4.2. Text analysis

We begin with an exploratory analysis on the most frequently occuring words obtained from the property advertisements in each neighborhood type, shown in Fig. 2. From the figure, we can first observe that the neighborhood name placeholder features prominantly in all neighborhood types when simply examining the raw count of words. This is also the case for certain adjectives – "beautiful", for example is a top word for all neighborhood categories. However, note the difference in the relative frequency in the neighborhood name and the next most commonly used word (e.g., hardwood in White-Higher



	Black Applicants 2018	White Applicants 2018	Hispanic Applicants 2018	Median Income 2018	Change in Black Applicants	Change in White Applicants	Change in Income	Change in Hispanic Applicants	Minority Pop 2013	Change in Minority Pop
White\Higher-Income (n=80)	5.97	75.28	5.00	105.53	1.63	-6.75	15.22	1.76	26.81	-1.53
White Homebuyers\Minority Neighborhoods (n=39)	14.21	67.96	10.45	69.17	-7.98	2.50	19.35	-5.56	77.19	-1.09
Increasing Black\Minority Neighborhoods (n=52)	39.24	43.75	14.93	60.67	8.64	-12.60	12.16	5.86	65.48	-4.88
White\Increasingly High Income (n=25)	3.04	78.78	2.39	183.88	1.06	-7.44	51.24	0.62	19.07	-18.82
Hispanic Homebuyers\Minority Neighborhoods (n=14)	27.81	50.71	34.79	49.47	-33.73	17.25	7.41	13.85	80.91	-1.16

Fig. 1. Description and spatial distribution of HMDA-derived types of neighborhoods. The y-axis in the graph is in percentages for all variables except median income which is in hundreds of thousands of dollars.

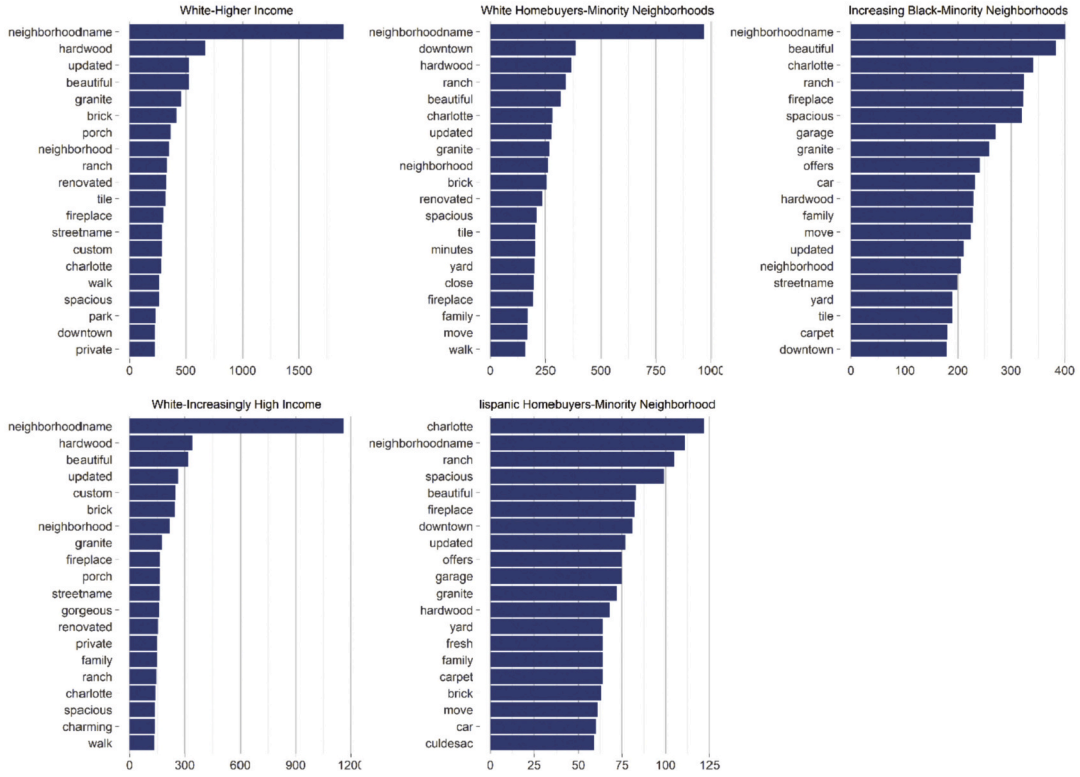


Fig. 2. Most frequently occurring words by neighborhood type.

Income neighborhoods). For property listings in neighborhoods with a majority or increasing share of White homebuyers, the neighborhood name is used approximately three times as often as the next most common word. On the other hand, in listings in neighborhoods with an increasing share of minority homebuyers, the difference in frequency between neighborhood name and the next most common word is very small. This points to a comparatively higher importance of mentioning specific neighborhood names among neighborhoods attracting White homebuyers as compared to minorities.

Digging deeper into the graphs reveals additional distinguishing characteristics between words according to neighborhood groups. For instance, in the White Homebuyers-Minority Neighborhoods group, words alluding to revitalizing neighborhoods close the urban core are included such as “uptown”, “updated”, “renovated”. The suburban nature of the Increasingly Black and Hispanic Homebuyers Minority Neighborhoods is also apparent with words such as “garage”, “car”, and “culdesac” whereas “walk” is the only mobility-oriented mention in the predominatly White homebuyer neighborhoods.

We next turn to the results of our binomial regressions for each type. We first run the models including city specific neighborhood, street and amenity names in the text, and thereafter with generic placeholders such as ‘neighborhoodname’, ‘streetname’, ‘restaurant’ and ‘entertainment’ included instead of the actual names of neighborhoods, streets, and places. This step serves to help validate the results of the k – means

grouping from local knowledge and also aids in an initial exploratory step for identifying specific names that should be grouped into more generic categories of amenities, landmarks, or property features. Fig. 3 shows the variables (i.e., words) associated with the 15 largest coefficients in each direction, respectively, for each neighborhood cluster. These are the words that increase or decrease the probability of a property listing to belong to a certain cluster the most.

In this set of models, names of specific neighborhoods play a large role in distinguishing neighborhood clusters from one another. For example, the near-center-city neighborhood of Villa Heights has

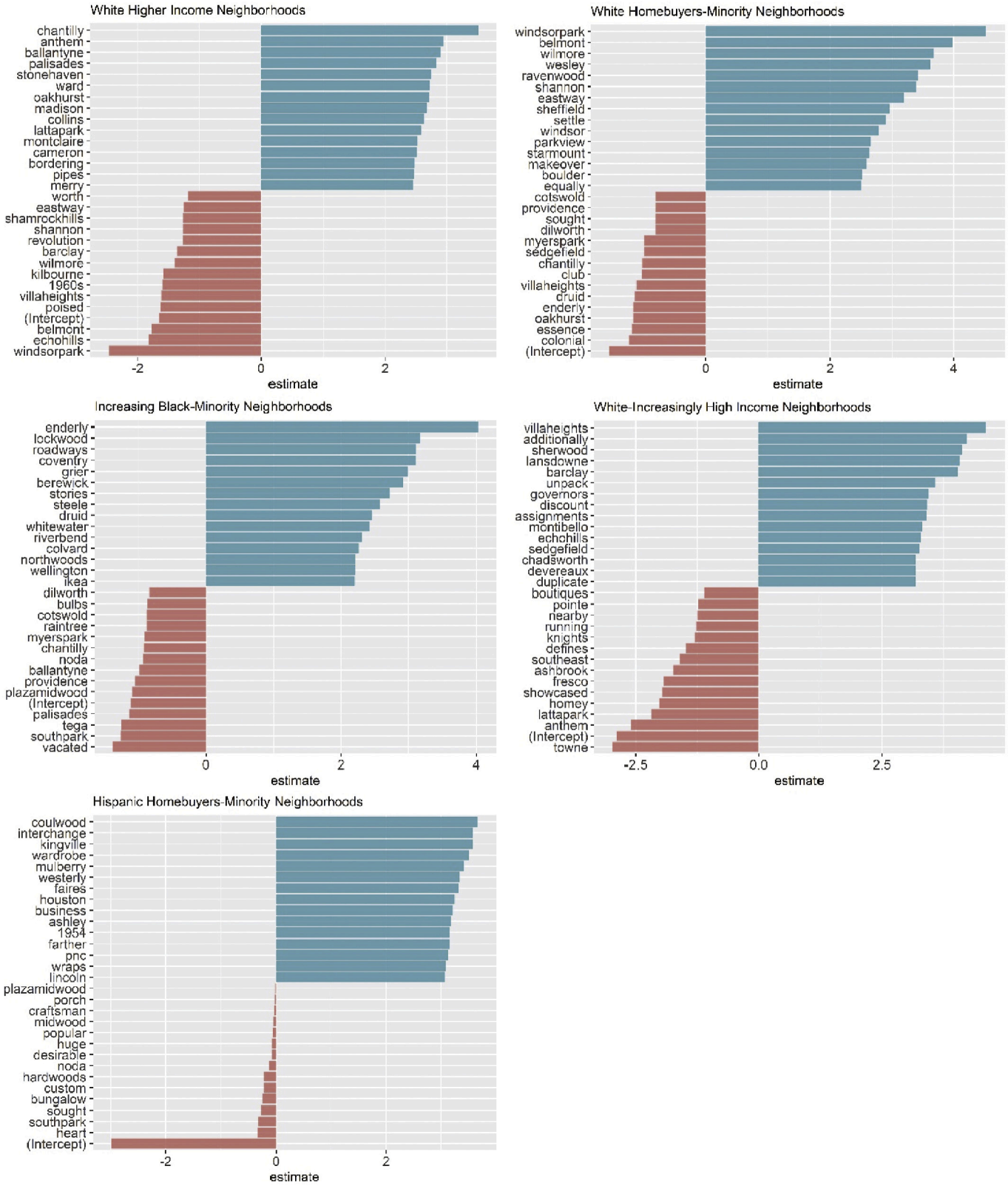


Fig. 3. Words with largest coefficients for each neighborhood type.

experienced an influx of White, higher-earning millennials in the past decade and is today considered gentrified (Logan, 2018) – it appears first on the list for the White Increasingly High-Income type. Names of neighborhoods that are most immediately facing gentrification pressures, those that still have a considerable minority population, but are experiencing an increasing share of White residents such as Belmont, Wesley Heights, and Wilmore top the list of the White Homebuyer-Minority Neighborhood class (Dunn, 2017). Ballantyne, the wards constituting the uptown area, and Dilworth where Latta Park is located, are wealthier neighborhoods that are often cited as the 'best places to live' in Charlotte (Hill & Hoover, 2020). As for the neighborhood names associated with Hispanic homebuyers in minority neighborhoods, these are suburban neighborhoods in the city's far northwest and northeast parts such as Coulwood, Westerly Hills, and Harris Houston. The minority neighborhoods associated with an increasing share of Black applicants include traditionally minority neighborhoods near the center city such as Enderly Park, Lockwood, and Grier Heights which has, as noted in the popular press, experienced increasing gentrification pressures recently (Clasen-Kelly, 2017; Dunn, 2017; Lambert, 2019). However, suburbs at the very edge of the city such as Steele Creek, Colvard Park, and Northwoods, also show up among the top 15 (positive) predictors. At a minimum, including neighborhood names affirms that the procedure – in terms of both the classification using the HMDA data and the text analysis can accurately connect neighborhoods with the currently known characteristics of these places.

Next, we look at the predictive performance of the models, summarized in Table 2 at a 0.8 threshold (i.e., test observations with predicted probabilities greater or equal to 0.8 are assigned to the cluster being modeled). The estimated binomial models across all neighborhood types tends to underestimate the probability of a neighborhood belong to a certain cluster ($y_i = 1$) and overestimate the number of neighborhoods belong to other clusters ($y_i = 0$). This is reflected in the trade-off between precision (correctly predicted 1 s as a share of the total number of predicted 1 s) and recall (correctly predicted 1 s as a share of the total number of actual 1 s) across all models, with high precision but lacking recall. The high accuracy (correctly predicted 0 s and 1 s as a share of total records predicted) mainly comes from the high share of true negatives (correctly predicted 0 s). The main cause for the underperformance of the predictive model is likely due to the relatively small sample where the testing dataset only contains 2112 observations and the training dataset 6336, split across five clusters. This makes our sample (1) have a rather skewed (or unbalanced) distribution for each cluster model, and (2) noisy and sensitive to outliers. This is particularly evident for the cluster with the smallest number of observations, Cluster 5, were the binomial distribution between 0 s and 1 s is highly skewed and the recall of the model is very low. However, splitting the data in favor of a larger testing dataset would have given the model a relatively small set to train on and hence introduced a lot of variance/noise into the model estimates. Hence, we believe that predictive performance will increase with a larger sample. This should lower the importance of city specific characteristics as well as outliers. In sum, even with this rather limited dataset, the results demonstrate the promise of this empirical approach to predict neighborhood type based on words from property advertisement descriptions.

Finally, we estimate the models without specific neighborhood or street names in the data substituting these for generic placeholders. Fig. 4 shows the coefficients that increase or decrease the probability of a property listing to belonging to a certain cluster. Turning to the White Homebuyers-Minority Neighborhoods class first, we see that "parquet" flooring is a key advertised home feature in these currently gentrifying neighborhoods. Other words such as "demand" signals a hot home market that a realtor is likely to capitalize on to garner a premium price. "Possibility" and "cosmetic" indicate that the properties are being advertised to an individual seeking a home to fix-up or possibly flip for a profit in a high demand market. "Qualified" is a word that may allude to some discriminatory practices as though only certain homebuyers qualify for these homes in historically minority neighborhoods. Finally, "families" is one of the most negatively associated words with these types of neighborhoods, giving credence to the notion of a city increasingly divided by age – with families in the suburbs and child-less households in the center.

For the Black Homebuyer -Minority Neighborhood type, we see that the generic neighborhood name placeholder is the most negatively associated term – this is also true for the Hispanic Homebuyer neighborhoods suggesting that mentioning the specific neighborhood the property is in carries more prestige in neighborhoods where Whites are the predominant homebuyer. We also see the home type for the Black homebuyer neighborhood reflects newer subdivision (multi-storied) homes with a lack of historical character (as identified by "stories" and a negative relationship with "bungalow"). Schools are also negatively associated with Black and Hispanic homebuyer neighborhoods. White-Increasingly High-Income neighborhoods feature very expensive home attributes ("subzero" appliances, "quartzite", "cherry") and attribute choice words of "exquisitely" versus "nice".

The predictive performance of this set of models are presented in Table 3. The number in parentheses is the percentage point change in the different metrics from the models in Table 2.

Across all models, the recall is significantly reduced from the models that included the specific neighborhood names. The reduction in accuracy and recall is the highest for clusters 1, 2, and 4, the White and Higher Income neighborhoods. However, while clusters 1 and 4 experienced the largest reduction in precision, cluster 2's precision increased. The accuracy of clusters 3 and 5, the minority neighborhoods with increasing shares of Black and Hispanic homebuyers remains similar to the models where neighborhood and street names were included. The better performance of the models in the first set of predictions (Table 1) with the inclusion of city-specific neighborhood, street, and amenity names is expected given that these become very distinguishing features for each set of neighborhoods (and will have high differential probabilities). This is especially true given the clustering of neighborhood types in space, meaning that proximity to certain neighborhoods or places become common characteristics for each type. What is noteworthy is that this appears to have a greater influence on the White-Higher Income, White Homebuyers in Minority Neighborhoods, and the White-Increasingly High-Income neighborhoods. With a large enough dataset covering multiple cities and time periods, we expect such city-specific characteristics should play a less important role.

Table 2
Performance metrics for models (threshold = 0.8).

Binomial model for cluster:	Accuracy	Precision	Recall
1: White Higher Income	75.4%	95.4%	21.0%
2: White Homebuyers-Minority Neighborhoods	81.3%	89.7%	16.8%
3: Increasing Black-Minority Neighborhoods	78.5%	86.8%	10.3%
4: White-Increasingly High Income	86.7%	81.1%	20.6%
5: Hispanic Homebuyers-Minority Neighborhoods	92.3%	100%	3.1%

5. Conclusions

In this article, we analyzed how the text used to advertise real estate properties varied according to the racial and income profile of homebuyers in different types of neighborhoods. Drawing from amenity-based theories of residential location choice and historical practices of real estate agents serving as neighborhood gatekeepers as they seek to maximize profits to attract particular homebuyers to neighborhoods, we hypothesized that the vernacular of the advertisement would vary by neighborhood. Using a small, 1-month sample property listing dataset as a proof of concept for Charlotte, North Carolina, we developed a framework for analyzing words by neighborhoods and for predicting neighborhood type from property listings.

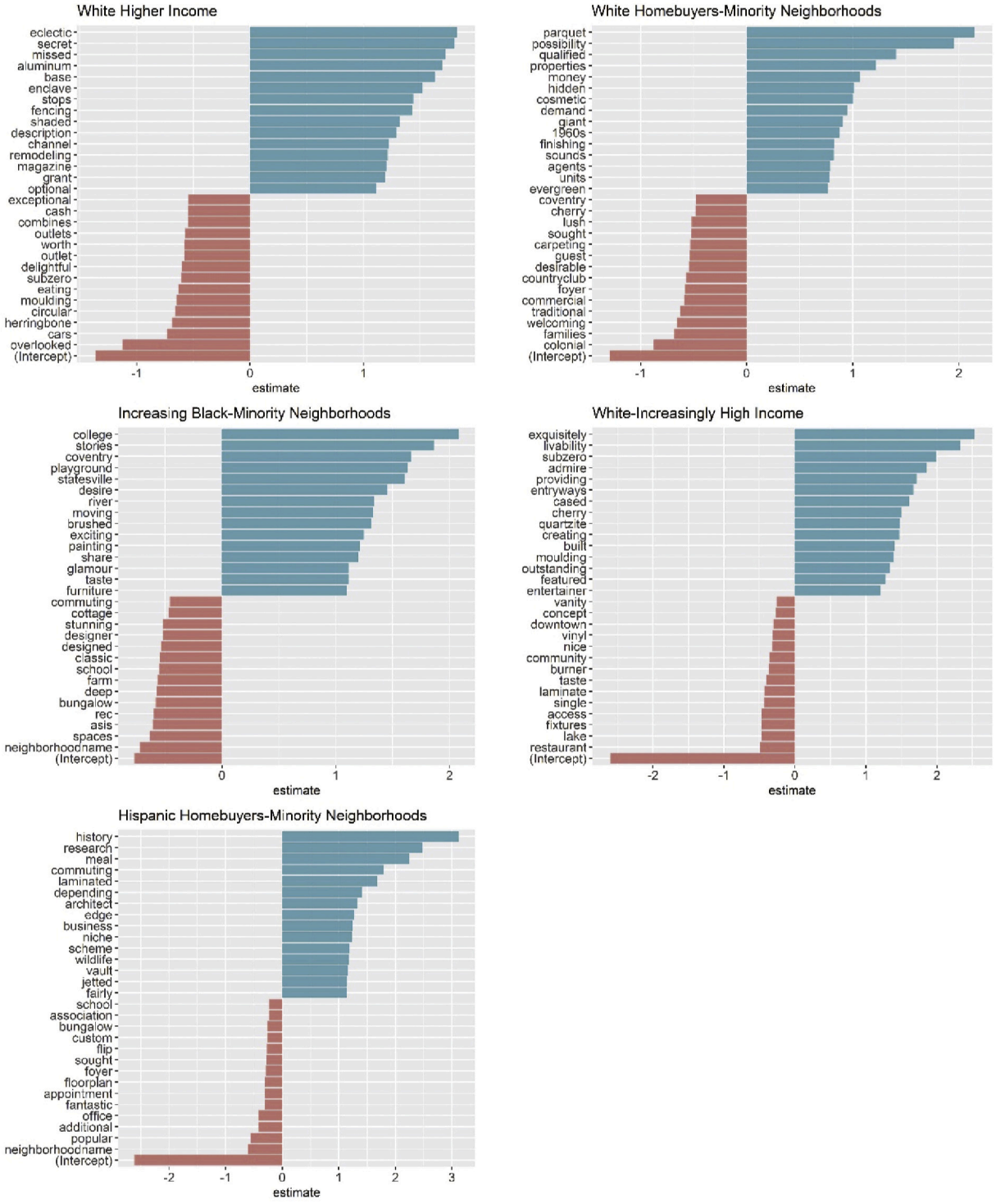


Fig. 4. Words with largest coefficients for each neighborhood cluster using generic neighborhood, street, and amenity names.

We expected neighborhoods currently undergoing a process of gentrification – identified in our analysis by a higher share of White home mortgage applicants in predominantly minority neighborhoods – would possess a larger share of consumer-cultural urban amenities to draw in younger, educated, and child-less homebuyers. We found some evidence of this, specifically in references to proximity to the urban center and a negative association with the word 'families'. Unique home characteristics (parquet flooring) and indicators that properties were prime for remodeling to certain "qualified" buyers, potentially exclusionary language used in changing neighborhoods, were also predictors of this neighborhood type. A more explicit examination of these property listings as well as those for rental housing for discriminatory language is an area of future research we plan to pursue. We did not find as many indicators of creative-cultural amenities as we had expected as was the case in a preliminary analysis of words along a new light rail corridor in Charlotte (Delmelle, Nilsson, & Schuch, 2020). To better extract these types of urban amenities, our future research will seek to develop ways of improving and automating the classification of specific city-specific names of restaurants, breweries, galleries, etc., into generic categories to facilitate analyses of larger datasets across multiple cities.

Table 3
Performance metrics for models estimated on data with generic neighborhood, street and amenity names (threshold = 0.8).

Binomial model for cluster:	Accuracy (diff)	Precision (diff)	Recall (diff)
1: White Higher Income	70.6% (−4.8)	73.3% (−22.1)	1.9% (−19.1)
2: White Homebuyers-Minority Neighborhoods	78.4% (−2.9)	100% (10.3)	0.1% (−16.7)
3: Increasing Black-Minority Neighborhoods	76.3% (−2.2)	81.8% (−5.0)	1.9% (−8.4)
4: White-Increasingly High Income	82.3% (−4.4)	63.6% (−17.5)	2.0% (−18.6)
5: Hispanic Homebuyers-Minority Neighborhoods	93.3% (1.0)	100% (0)	0.1% (−3.0)

We also expected the name of the neighborhood to hold more prestige in wealthier and Whiter and neighborhoods, and found this to be the case as our generic placeholder for neighborhood name was very negatively associated with neighborhoods with larger shares of Hispanic and Black homebuyers. These neighborhoods were also negatively associated with schools and associated with more suburban traits – two-story homes, cars, and garages, and lacking in historical character thus reinforcing the racial inversion of center cities and older suburbs and linking back to prior work on real estate agent steering practices where Black residents were disproportionately shown homes in neighborhoods with worse schools (Galster & Godfrey, 2007). These preliminary results thus underpin the idea that Black, high, amenity neighborhoods are in short-supply. We plan to further investigate this finding in future research.

With respect to the predictive ability of the approach, the estimated models across all neighborhood types tended to underestimate the probability of a neighborhood belonging to a certain type (true positives) and overestimate the number of neighborhoods belonging to other types (true negatives). This is reflected in the trade-off between precision and recall across all models, with a high precision but lacking in recall. While the accuracy of the binomial models for each neighborhood type is fairly high given the small sample size, it mainly comes from the high share of true negatives. The small sample size makes our sample have a rather unbalanced distribution as well as noisy and sensitive to outliers. We believe that the predictive performance will increase with a larger sample including more properties across multiple cities. Regardless, even with this small sample, the method shows the potential to predict neighborhood types based on words from property advertisements.

Aside from those state above, we see several future research directions that can be built upon by this illustrative example. First, given the predictive promise of determining a neighborhood's racial and income composition in near real time (derived from annually-updated HMDA data), one next step of expansion is to build upon the predictive power of this approach to serve as an early warning system of neighborhood change (Chapple & Zuk, 2016). If a large share of listed properties predicts a neighborhood type that differs from its current one, it is suggestive that real estate agents are seeking out a different type of homebuyer and may indicate that transformations are underway, even before demographic shifts have actually taken place. Second, our approach relied on a small, cross-sectional sample of real estate listings, but larger, longitudinal datasets from different types of metropolitan areas should help tease out more urban amenities and their potentially evolving role in advertising to various segments of the population and hence contribute to knowledge on residential sorting.

Overall, this illustrative example has served to highlight the potential in exploiting spatially-based text analysis of property listings for understanding urban dynamics. These data may prove especially useful in illuminating the less visible processes at play before residential sorting has played out: they can highlight subtle terms that may be used to deter or attract certain homebuyers to neighborhoods. The evaluation of rental property listings may further contain more explicit, legal discriminatory language (e.g. restrictions on housing vouchers, setting credit limits, restrictions on former evictions and criminal records) that serve to reinforce and explain perpetual patterns of segregation by race and income.

Acknowledgements

This research was supported in part by a faculty research grant from the University of North Carolina at Charlotte. The authors thank the two reviewers for their very helpful comments that helped us to improve this manuscript.

References

Baum-Snow, N., & Hartley, D. (2020). Accounting for central neighborhood change, 1980–2010. *Journal of Urban Economics, 117*, 103228.
Bayer, P., & McMillan, R. (2005). *Racial sorting and neighborhood quality (No. w11813)*. National Bureau of Economic Research.
Beamish, J. O., Goss, R. C., & Emmel, J. (2001). Lifestyle influences on housing preferences. *Housing and Society, 28*, 1–29.
Bereitschaft, B. (2014). Neighbourhood change among creative–cultural districts in mid-sized US metropolitan areas, 2000–10. *Regional Studies, Regional Science, 1*(1), 158–183.
Besbris, M., & Faber, J. W. (2017). Investigating the relationship between real estate agents, segregation, and house prices: Steering and upselling in New York State. *Sociological Forum, 32*(4), 850–873.
Borgoni, R., Michelangeli, A., & Pontarollo, N. (2018). The value of culture to urban housing markets. *Regional Studies, 52*(12), 1672–1683.
Chapple, K., & Zuk, M. (2016). Forewarned: The use of neighborhood early warning systems for gentrification and displacement. *Cityscape, 18*(3), 109–130.
Chemtob, D., & Off, G. (2019). Charlotte jumps in rankings of largest U.S. cities, surpassing Indianapolis. In *Charlotte Observer*. May 24 https://www.charlotteobserver.com/news/business/biz-columns-blogs/development/article230790609.html (retrieved 12/4/20).
Chemtob, D., & Off, G. (2020). Charlotte growth pushes it past San Francisco to become 15th biggest city in the US. In *The Charlotte Observer*. *https://www.charlotteobserver.com/news/business/biz-columns-blogs/development/article242897996.html* (retrieved 12/2/20).
Childress Klein Center for Real Estate. (2020). *2020 the state of housing in Charlotte report*. University of North Carolina at Charlotte. https://realestate.uncc.edu/research/state-housing-charlotte-report (retrieved 3/22/21).
Clasen-Kelly, F. (2017). 'We can't be bought.' Can this Charlotte neighborhood stop investors from moving in?. Charlotte Observer, June 30 https://www.charlotteobserver.com/news/local/article159093699.html.
Connor, D. S., Gutmann, M. P., Cunningham, A. R., Clement, K. K., & Leyk, S. (2020). How entrenched is the spatial structure of inequality in cities? Evidence from the integration of census and housing data for Denver from 1940 to 2016. *Annals of the American Association of Geographers, 110*(4), 1022–1039.
Couture, V., & Handbury, J. (2017). *Urban revival in America, 2000 to 2010 (No. w24084)*. National Bureau of Economic Research.
Delmelle, E., Thill, J. C., Furuseth, O., & Ludden, T. (2013). Trajectories of multidimensional neighbourhood quality of life change. *Urban Studies, 50*(5), 923–941.
Delmelle, E. C. (2019). The increasing sociospatial fragmentation of urban America. *Urban Science, 3*(1), 9.
Delmelle, E. C., Nilsson, I., & Adu, P. (2020). Poverty Suburbanization, Job Accessibility, and Employment Outcomes. *Social Inclusion, 9*(2), 166–178.
Delmelle, E. C., Nilsson, I., & Schuch, J. C. (2020). Who's moving in? A longitudinal analysis of home purchase loan borrowers in new transit neighborhoods. *Geographical Analysis*. https://doi.org/10.1111/gean.12234.
Dunn, A. (2017). In Charlotte's trendy neighborhoods, a culture clash of Black and White, rich and poor. Charlotte Agenda, July 20 https://www.charlotteagenda.com/97973/charlottes-trendy-neighborhoods-culture-clash-Black-White-rich-poor/.
Ehrenhalt, A. (2012). The great inversion and the future of the American city. *Vintage*.
Florida, R. (2017). The new urban crisis: How our cities are increasing inequality, deepening segregation, and failing the middle class-and what we can do about it. *Basic Books*.
Florida, R., & Adler, P. (2018). The patchwork metropolis: The morphology of the divided postindustrial city. *Journal of Urban Affairs, 40*(5), 609–624.
Frey, W. H. (2017). City growth dips below suburban growth, census shows. *Brookings, 30*, 2017.
Galster, G., Freiberg, F., & Houk, D. L. (1987). Racial differentials in real estate advertising practices: An exploratory case study. *Journal of Urban Affairs, 9*(3), 199–215.
Galster, G., & Godfrey, E. (2005). By words and deeds: Racial steering by real estate agents in the US in 2000. *Journal of the American Planning Association, 71*(3), 251–268.
Glaeser, E. L., & Gottlieb, J. D. (2006). Urban resurgence and the consumer city. *Urban Studies, 43*(8), 1275–1299.

Hastie, T., & Qian, J. (2014). *Glmnet Vignette, 9*(2016), 1–30. Retrieved June.
Hill, L., & Hoover, H. (2020). *Here are the best places to live if you're moving to Charlotte, NC*.
Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2013). RTextTools: A supervised learning package for text classification. *R Journal, 5*(1).
Kang, W., Rey, S., Wolf, L., Knaap, E., & Han, S. (2020). Sensitivity of sequence methods in the study of neighborhood change in the United States. *Computers, Environment and Urban Systems, 81*, 101480.
Lambert, L. (2019). America's 10 fastest-gentrifying neighborhoods. Realtor.com, April 15. https://www.realtor.com/news/trends/the-10-fastest-gentrifying-neighborhoods-in-america/?identityID=5b32653e27287d5f621833cd&MID=2019_0419_WeeklyNL&RID=5389840302&cid=eml_promo_Marketing_NonPRSL_WeeklyNL_cons.10824802_2019_0419_WeeklyNL-blog_1_10fastestgentrify-blogs_trends&fbclid=IwAR0DdpU3lMoJb_2IvyP_vbbgXMzdNKf9QhhXfPtV7dLihBABoNPRJN8BTqU.
Lee, H. (2021). Are millennials leaving town? Reconciling peak millennials and youthification hypotheses. *International Journal of Urban Sciences*, 1–19.
Lee, J., Irwin, N., Irwin, E., & Miller, H. J. (2021). The role of distance-dependent versus localized amenities in polarizing urban spatial structure: A Spatio-temporal analysis of residential location value in Columbus, Ohio, 2000–2015. *Geographical Analysis*. https://doi.org/10.1111/gean.12238. In Press.
Lee, Y., Lee, B., & Shubho, M. T. H. (2019). Urban revival by Millennials? Intraurban net migration patterns of young adults, 1980–2010. *Journal of Regional Science, 59*(3), 538–566.
Logan, L. (2018). Like an ex, Villa Heights isn't the neighborhood we used to know. Charlotte Observer, July 9 https://www.charlotteobserver.com/charlottefive/c5-people/article236117218.html.
Massey, D. S. (2005). Racial discrimination in housing: A moving target. *Social Problems, 52*(2), 148–151.
NAR – National Association of Realtors. (2019). Seasonality in the Housing Market. https://www.nar.realtor/blogs/economists-outlook/seasonality-in-the-housing-market (retrieved 3/22/21).
Nilsson, I., & Reid, N. (2019). The value of a craft brewery: On the relationship between craft breweries and property values. *Growth and Change, 50*(2), 689–704.
Pearce, D. M. (1979). Gatekeepers and homeseekers: Institutional patterns in racial steering. *Social Problems, 26*(3), 325–342.
Pryce, G., & Oates, S. (2008). Rhetoric in the language of real estate marketing. *Housing Studies, 23*(2), 319–348.
Realtor.com. (2021). *Charlotte*. NC Housing Market. https://www.realtor.com/realestateandhomes-search/Charlotte_NC/overview (retrieved 3/22/21).
Roscigno, V. J., Karafin, D. L., & Tester, G. (2009). The complexities and processes of racial housing discrimination. *Social Problems, 56*(1), 49–69.
Sampson, R. J. (2012). *Great American city: Chicago and the enduring neighborhood effect*. University of Chicago Press.
Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach: " O'Reilly Media, Inc."*.
Strahilevitz, L. J. (2006). Exclusionary amenities in residential communities. *Virginia Law Review, 92*, 437.
Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64*(5), 416–424.
Van Criekingen, M., & Decroly, J. M. (2003). Revisiting the diversity of gentrification: Neighbourhood renewal processes in Brussels and Montreal. *Urban Studies, 40*(12), 2451–2468.
Zhao, B., Ondrich, J., & Yinger, J. (2006). Why do real estate brokers continue to discriminate? Evidence from the 2000 Housing Discrimination Study. *Journal of Urban Economics, 59*(3), 394–419.