International Journal of
Geo-Information

MDPI

Article

Identifying Urban Neighborhood Names through User-Contributed Online Property Listings

Grant McKenzie [1,*], Zheng Liu [2], Yingjie Hu [3] and Myeong Lee [4]

[1] Department of Geography, McGill University, Montréal, QC H3A 0B9, Canada
[2] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, USA; zliu1208@terpmail.umd.edu
[3] Department of Geography, University at Buffalo, Buffalo, NY 14260, USA; yhu42@buffalo.edu
[4] College of Information Studies, University of Maryland, College Park, MD 20742, USA; myeong@umd.edu
* Correspondence: grant.mckenzie@mcgill.ca

Received: 24 August 2018; Accepted: 22 September 2018; Published: 26 September 2018

Abstract: Neighborhoods are vaguely defined, localized regions that share similar characteristics. They are most often defined, delineated and named by the citizens that inhabit them rather than municipal government or commercial agencies. The names of these neighborhoods play an important role as a basis for community and sociodemographic identity, geographic communication and historical context. In this work, we take a data-driven approach to identifying neighborhood names based on the geospatial properties of user-contributed rental listings. Through a random forest ensemble learning model applied to a set of spatial statistics for all n-grams in listing descriptions, we show that neighborhood names can be uniquely identified within urban settings. We train a model based on data from Washington, DC, and test it on listings in Seattle, WA, and Montréal, QC. The results indicate that a model trained on housing data from one city can successfully identify neighborhood names in another. In addition, our approach identifies less common neighborhood names and suggestions of alternative or potentially new names in each city. These findings represent a first step in the process of urban neighborhood identification and delineation.

Keywords: neighborhood; neighborhood name; random forest; spatial statistics; housing; craigslist

1. Introduction

In 2014, Google published a neighborhood map of Brooklyn, the most populous borough in New York City, a seemingly harmless step in providing its users with useful geographic boundary information. The backlash was swift. Residents of Brooklyn responded angrily, many stating that a commercial company such as Google had no right to label and define boundaries within their city [1]. This was not a lone incident [2], as many mapping agencies, both government and commercial, have come to realize that regional boundaries and names are a contentious issue. Google and others are frequently placed in the difficult situation of publishing hard boundaries and definitive names for regions that are in dispute or poorly defined [3,4], often applying names to parts of the city that few residents have even heard before [5]. This poses a problem as the names assigned to neighborhoods are important for understanding one's identity and role within an urban setting. Names provide a bond between a citizen and a place [6]. In many cases, neighborhood names are much more than just a set of characters; they have a history that is situated in religious beliefs [7], gender identity [8] and/or race [9]. Neighborhood names evolve over time and are given meaning by the neighborhood's inhabitants. Applying a top-down approach to naming neighborhoods, a practice often done by municipalities and commercial agencies, can produces unforeseen, even anger-inducing, results.

Historically, neighborhood identification has also been predominantly driven through financial incentives. The term redlining, which describes the process of raising service prices or denying loans in selective neighborhood and communities based on demographics such as race, was coined in the 1960s [10] and is one of the foundational examples of neighborhood delineation driven by financial interests. In many ways, the neighborhood boundaries of many U.S. cities today are at least a partial result of these practices. Real estate companies still rely on neighborhood boundaries for comparable pricing [11], and being associated with a neighborhood name can significantly impact one's social capital [12], as well as mortgage rate [13]. Today, web-based real estate platforms such as Zillow, Redfin and Trulia curate their own neighborhood dataset [14]. These companies realize the immense value of these boundaries and names [15] and actively invest in promoting their brand's datasets (Zillow for example freely offers access to its neighborhood boundaries and real estate APIs.)

While commercial mapping companies and real estate platforms engage in the complex process of geographically splitting up a city into neighborhoods and labeling those regions, the inhabitants and citizens themselves often have their own understanding of the region in which they live. Their historically-rooted understanding of a neighborhood can sometimes be at odds with the neighborhood identification methods employed by these commercial entities. The urban historian, Lewis Mumford stated that "Neighborhoods... exist wherever human beings congregate, in permanent family dwellings; and many of the functions of the city tend to be distributed naturally—that is, without any theoretical preoccupation or political direction" [16]. That is to say that neighborhoods differ from other regional boundaries (e.g., city, census tract) in that they are constructed from the bottom-up by citizens, rather than top-down by governments or commercial entities. Any attempt to interfere with this bottom-up approach is met with resentment from residents of the neighborhoods, as evidenced by Google's Brooklyn neighborhood map. In fact, one of the goals of public participatory GIS has been to enable citizens to construct, identify and contribute to their communities and neighborhoods [17,18], thus defining the regions themselves.

Today, information is being generated and publicly disseminated online by everyday citizens at an alarming rate. While governments and industry partners have increased their involvement in public participatory GIS and engagement platform (See ArcGIS Hub and Google Maps Contributions, for example), the vast majority of content is being contributed through social media applications, personal websites and other sharing platforms, many of which include location information. Online classified advertisements are an excellent example of this recent increase in user-generated content. People post advertisements for everything from local services to previously used products and, most notably, rental properties. craigslist is by far the most popular online website for listing and finding rental properties in the United States, Canada and many other countries and is therefore a rich source of information for understanding regions within a city. As inhabitants, property owners or local rental agencies post listings for rental properties on such a platform, they geotag the post (either through geographic coordinates or local address) and provide rich descriptive textual content related to the property. Much of this content includes standard information related to the property such as square footage, number of bedrooms, etc., but other information is related to the geographic location of the listing, namely nearby restaurants, public transit, grocery stores, etc. Neighborhood names are also frequently included in rental listing descriptions. Those posting the rental properties realize that by listing the neighborhood name(s) in which the property exists, they are effectively situating their property within a potential renter's existing idea and understanding of the region. While the motivation and biases surrounding which neighborhoods are included in the textual descriptions of a listing are important (which will be discussed in Section 6.2), these data offer a novel opportunity to understand how citizens, property owners and local real estate companies view their urban setting and label and differentiate the neighborhoods that comprise the city.

Given our interest in both identifying and delineating neighborhoods, this work tackles the preliminary, but essential step of extracting and identifying neighborhood names. The specific contributions of this work are outlined in the five research questions (RQ) below. Each builds on the findings of the previous question, and direct references to these RQs can be found in the manuscript.

RQ1 Can neighborhood names be identified from natural language text within housing rental listings? Specifically, can spatially descriptive measures of geo-tagged n-grams be used to separate neighborhood names from other terms? A set of spatial statistical measures is calculated for all n-grams (An n-gram is a sequence of *n* items (often words) identified in text. For example 'kitchen' is a uni-gram, 'small kitchen' a bi-gram, etc.) in a set of listings and used to identify neighborhoods names.

RQ2 Does an ensemble learning approach based on spatial distribution measures more accurately identify neighborhood names than the spatial distribution measures alone? Given spatial statistics for each n-gram in a set of listings, we show that combining these in a random forest model produces higher accuracy than individual measures alone.

RQ3 Can an identification model trained on a known set of neighborhood names be used to identify uncommon neighborhood names or previously unidentified neighborhoods? Training a random forest model on spatial statistics of common neighborhood names within a city, we demonstrate that lesser known neighborhood names can be identified. In some cases, alternative names or other descriptive terms are proposed through the use of such a model.

RQ4 Can a neighborhood name identification model trained on data from one city be used to identify neighborhood names in a different city? A random forest model constructed from neighborhood names in Washington, DC, is used in the identification of neighborhood names in Seattle, WA, and Montréal, QC.

RQ5 What are the biases associated with neighborhood names mentioned in rental property listings? Lastly, we report on the spatial distribution biases associated with craigslist rental listings in Washington, DC.

The remainder of this manuscript is organized as follows. Previous research related to this topic is discussed in Section 2, and an overview of the data is provided in Section 3. The spatial statistics and random forest methods are introduced in Section 4 including measures of accuracy. Section 5 presents the results of this work, which are then discussed in Section 6. Finally, conclusions, limitations and future work are the subjects of Section 7.

2. Related Work

Defining neighborhoods has been the subject of numerous research projects spanning many different domains. Understanding how neighborhoods are defined, as well as identifying characteristics that distinguish one neighborhood from another has a long history within geography, regional science, sociology and social psychology (see [19–21] for an overview). Many previous studies in the social sciences have contrasted citizen-defined neighborhoods against regions defined by government or commercial entities. Coulton et al. [22] provided an example of this type of research, having asked residents of a city to draw boundaries on a map, defining their version of neighborhoods within a city. This process inevitably results in some overlap between neighborhood names and boundaries, but also quite a few significant differences. These citizen-defined boundaries are then often compared to census or other government-designated areas [23,24]. An outcome of these works is a clear need to better understand what a neighborhood is and how it can be identified based on the people that inhabit it.

From a geographic perspective, a substantial amount of work has aimed at defining geographic areas of interest. While many researchers steer clear of the term 'neighborhood,' many of the methods employed focus on delineating a sub-urban region for its surrounding components based on some characteristic or spatial property. Many of these rely on analyzing user-contributed point data accompanied by names, categories or descriptive tags. For instance, Schockaert and De Cock [25] identified the spatial footprints of neighborhoods from geotagged content, while a number of studies [26,27] identified areas of interest based on user-contributed photograph tags. Tags have been used in the identification of vaguely-defined regions, as well. For instance, social media tags and text were used to differentiate Southern California from Northern California [28].

Recent work has focused on extracting functional regions based on human activities and category-labeled places of interest [29], while other work has identified thematic regions such as the bar district or shopping regions of the city based on the popularity of social media check-ins [30]. Though not explicitly labeled as neighborhoods, the characteristics and activities afforded by these regions often result in them being referred to colloquially as neighborhoods. The livehoods project [31] aimed to identify regions based on the similarities of geosocial check-in patterns in various cities around the United States. This project, however, did not involve naming the small livehood regions.

From a data source perspective, existing work has used housing posts to better understand, explore and, in some cases, define neighborhoods [32,33]. Chisholm and Cohen [34] developed The Neighborhood Project, a web map based on combining geocoded craigslist posts with neighborhood names extracted from text in the posts. The neighborhood names themselves, however, were determined by experts and user-contributed knowledge of the region. Hu et al. [35] used housing advertisements as training data for a natural language processing (NLP) and geospatial clustering framework that harvests local and colloquial place names in order to enrich existing gazetteers. Our work further enhances this approach, combining measures from a range of statistical techniques to extract sub-urban regional names specifically. Zhu et al. [36] explored the use of spatial statistics to differentiate geographic feature types and disambiguate place names [37]. In these works, they showed that different feature types and place names exhibit different spatial patterns, and it is through these individual patterns that geographic features can be compared (e.g., mountain tops to beaches).

While a considerable amount of previous work has focused on neighborhood boundary identification and delineation, far less work has focused on the extraction of neighborhood names. Brindley et al. [38,39] took a data-driven approach to mapping urban neighborhoods, using postal addresses to extract neighborhood names and boundaries. Specifically, the authors extracted commonly-found sub-district names from within postal addresses and used a kernel density function to estimate the geographic boundary. While similar to our work in their usage of publicly available geo-tagged content, their approach did not combine various spatial statistics with natural language text for the extraction of neighborhood names, nor did it produce a prediction model that could be learned from one city and applied to another.

Place name extraction has been an important topic within the geographic information retrieval community for some time. Jones et al. [40] focused on the extraction of place names and vague regions from natural language on websites, while others were able to extract spatial relations from natural language in web documents [41]. In that same thread, additional research has looked at the identification of place names based on their context within descriptive documents [42]. Further work has focused on disambiguation of terms within a geographic context. For example, Buscaldi and Rosso [43] used term similarity and context to disambiguate place names from one another. The rise of social media content has led to new sources of geotagged content that has been used for named geographic entity extraction [44,45]. Co-occurrence of place names and other geographic locations within natural language text has been shown to correspond with close spatial proximity [46]. Still, other research has proposed machine learning approaches to identify and disambiguate places within a document based on contextual terms [47,48]. The work presented in this manuscript continues with this leitmotif, proposing a novel approach to identifying neighborhood names based on the spatial distribution and content of rental property listings.

3. Data

Two sources of data are used in this work, namely rental property listings and curated lists of neighborhood names. Both sets of data were collected for three cities in North America. Further details on these data are described below.

3.1. Rental Property Listings

Rental property listings were accessed from the online classified advertisements platform craigslist (http://craigslist.org). Specifically, postings in the apts/housing for rent section of the subdomains for three cities, Washington, DC, Seattle, WA, and Montréal, QC, were accessed over a six-month period starting in September 2017. These three cities were chosen based on the availability of content and geographic locations (two different coasts of the United States and one bilingual city in Canada). The content collected for each city consists of rental housing property listings. At a minimum, each listing contains geographic coordinates, a title and unstructured textual content describing the rental property.

Table 1 presents an overview of the data collected for each of the cities. The first column, Listings, reflects the total number of rental housing listings collected in and around each city over the course of six months. The Unique Locations column lists the number of unique rental housing listings for each city after data cleaning. Cleaning involved removing duplicate entries and restricting posts to only those listed with a unique pair of geographic coordinates. This had to be done due to the fact that many posts were repeated for the exact same listing location, but with slightly different titles and content (presumably an advertising tactic to saturate the market). Those listings with no textual content were removed.

Table 1. Number of craigslist housing listings, unique housing locations, unique number of n-grams across all city listings and cleaned unique n-grams.

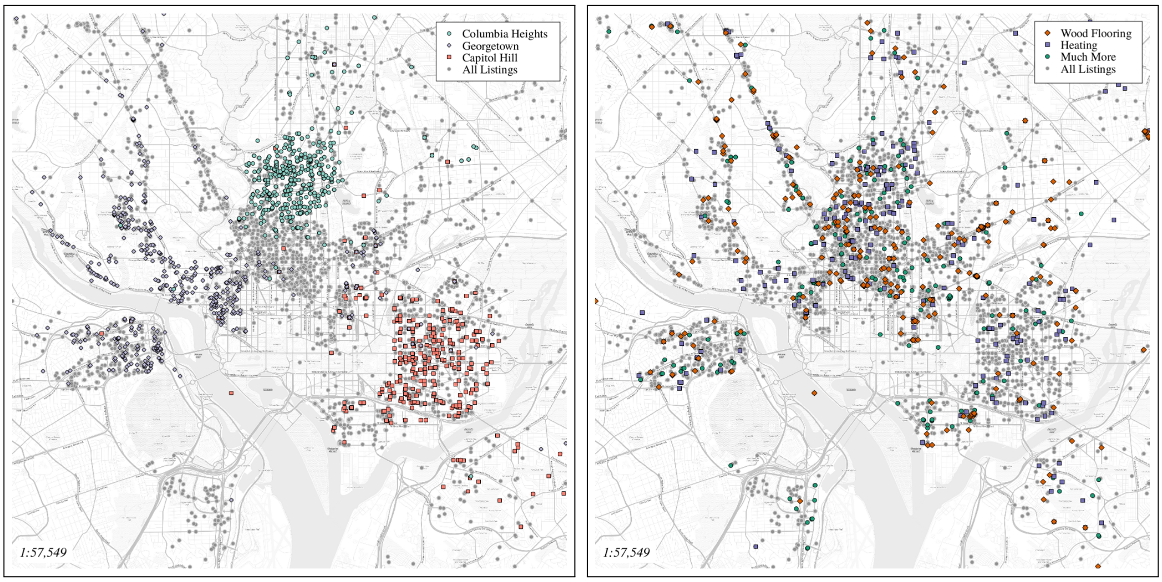
City	Listings	Unique Locations	Unique n-Grams	Cleaned n-Grams
Washington, DC	60,167	13,307	1,294,747	3612
Seattle, WA	68,058	17,795	1,053,297	5554
Montréal, QC	10,425	4836	571,223	2914

3.1.1. n-Grams

All the textual content, including titles, for each listing in a city were combined into a corpus, and the Natural Language Toolkit [49] was employed to tokenize words in the corpus and extract all possible n-grams (to a maximum of three words). The total number of unique n-grams per city is shown in Table 1. The frequency of occurrence within the corpus was calculated for each n-gram, and those with frequency values above four standard deviations from the mean were removed, as well as all n-grams that occurred less than 50 times within each city. Furthermore, all n-grams consisting of less than three characters were removed. The removal of the exceptionally high frequency n-grams was done to reduce computation given that it is highly unlikely that the most frequent words are neighborhood names. For example, the top five most frequent, greater than two character words in each of the cities are and, the and with. Similarly, the removal of n-grams occurring less than 50 times was done to ensure robustness in our neighborhood identification model and elicit legitimized neighborhood names. Given the long tail distribution of n-gram frequencies, this latter step removed most of the n-grams including single occurrence phrases such as included and storage, throughout painted and for rent around.

3.1.2. Geotagged n-Grams

Provided the reduced set of n-grams for each city, the original geo-tagged listings were revisited, and any n-grams found in the textual content of the listings were extracted and assigned the geographic coordinates of the listing. This resulted in a large set of <latitude, longitude, n-gram > triples for each city. These geo-tagged n-grams were intersected with the 1-km buffered boundaries for each city to remove all listings that were made outside of the city. The buffers were added to account for listings that described neighborhoods on city borders (e.g., Takoma Park on the District of Columbia-Maryland border). Figure 1 shows two maps of geo-tagged n-grams in Washington, DC. Figure 1a depicts the clustering behavior of neighborhood names (three examples shown in this case). Figure 1b shows a sample of three generic housing-related terms.



(**a**) N-grams of three neighborhood names

(**b**) N-grams of three non-neighborhood names

Figure 1. n-grams mapped from rental property listings in Washington, DC. (**a**) shows the clustering behavior of three neighborhood names, while (**b**) visually depicts the lack of clustering for a sample of generic housing terms.

3.2. Neighborhood Names and Boundaries

Since neighborhoods in the United States and Canada are neither federally-, nor state-/province-designated geographical units, there is no standard, agreed upon set of neighborhood names and boundaries for each city. In many cases, neighborhood boundaries are arbitrarily defined, and there is little agreement between neighborhood data sources. Zillow, for example, provides a freely available neighborhood boundaries dataset for large urban areas in the United States that is heavily based on property values. Platforms such as Google Maps also contain neighborhood boundaries for most cities in the United States. However, Google considers these data proprietary and does not make them available for use in third-party applications. There are numerous other sources of neighborhood or functional region boundaries available for specific cities, but few of these sources offer boundaries for more than a single urban location. Table 2 lists four sources of neighborhood names and boundaries along with the number of neighborhood polygons available for each city. Notably, the number of neighborhood names and polygons ranges substantially between data sources. Washington, DC, for example, consists of 182 neighborhood boundaries according to Zetashapes compared to 46 listed on DC.gov.

To build a training set for our machine learning model, we attempted to match each of the neighborhood names in each of the sources and exported those names that occurred in the majority of the sources. We label these our common neighborhoods and use them as the foundation on which to build the identification model.

Table 2. Neighborhood names and boundary sources including polygon counts for each city. The * indicates that this source assigns many neighborhood names (comma delimited) to larger than average neighborhood regions. Note that Zillow and Zetashapes do not provide neighborhood names outside of the United States.

Source	Washington, DC	Seattle, WA	Montréal, QC
Wikipedia	129	134	73
Zillow	137	115	N/A
Zetashapes/Who's On First	182	124	N/A
City Government/AirBnB	46 *	106	23
Common Neighborhoods	*95*	*79*	*23*

4. Methodology

In this section, we first give an overview of the various spatial statistics used to describe the n-grams spatially. This is followed by assessing the prediction power of each spatial statistic *predictor* in identifying neighborhood names and finally describing how the predictors are combined in a random forest ensemble learning model. Figure 2 depicts a flowchart of the process, with example data, from data-cleaning to the random forest model.

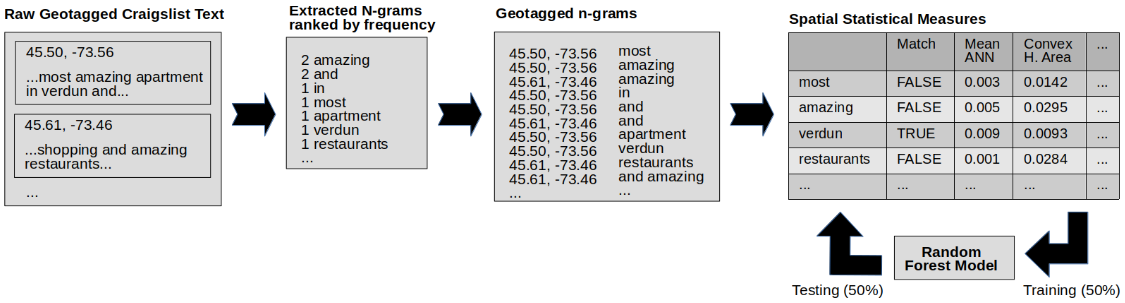

Raw Geotagged Craigslist Text
45.50, -73.56
...most amazing apartment in verdun and...
45.61, -73.46
...shopping and amazing restaurants...
...
Extracted N-grams ranked by frequency
2 amazing
2 and
1 in
1 most
1 apartment
1 verdun
1 restaurants
...
Geotagged n-grams
45.50, -73.56 most
45.50, -73.56 amazing
45.61, -73.46 amazing
45.50, -73.56 in
45.50, -73.56 and
45.61, -73.46 and
45.50, -73.56 apartment
45.50, -73.56 verdun
45.61, -73.46 restaurants
45.61, -73.46 and amazing
... ...
Spatial Statistical Measures
Match | Mean ANN | Convex H. Area | ...
most | FALSE | 0.003 | 0.0142 | ...
amazing | FALSE | 0.005 | 0.0295 | ...
verdun | TRUE | 0.009 | 0.0093 | ...
restaurants | FALSE | 0.001 | 0.0284 | ...
... | ... | ... | ... | ...
Random Forest Model
Testing (50%)
Training (50%)

Figure 2. A flowchart showing the process and example data for the methodology in this work. Note that the data are simplified/rounded for example purposes.

4.1. Spatial Statistics

The fundamental assumption in this work is that different categories of words can be described by an array of statistics associated with the locations of their use. We hypothesize that neighborhood names exhibit unique spatial statistical patterns, which can be used to identify and extract these neighborhood names from other terms specifically. With this goal in mind, we identified a few foundational spatial statistics that can be applied to representing point data in space. In total, 24 different spatial statistics measures, roughly grouped in to three categories, are used in describing each of the n-grams in our dataset. To be clear, we do not claim that this list of spatial statistics is exhaustive, but rather intend to show what is possible with a select set of measures.

4.1.1. Spatial Dispersion

Nine measures of spatial dispersion were calculated for each n-gram in our datasets. *Standard Distance*, a single measure representing the dispersion of points around a mean centroid, was calculated along with average nearest neighbor and pairwise distance. We hypothesize that neighborhood names will be identified by this measure as neighborhood n-grams are likely to display a unique spatial dispersion pattern, different from most other non-geographic terms. Standard distance is shown in Equation (1) where x and y are individual point coordinates, $\bar{X}$ and $\bar{Y}$ are the mean centroid coordinates and n is the total number of geographic coordinates associated with the n-gram.

$$\mathit{Standard\ Distance} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} + \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}} \tag{1}$$

Within the category of average nearest neighbor (ANN), we calculated the mean and median for each point's closest n-gram neighbor (NN1), second nearest (NN2) and third nearest (NN3), resulting in six unique measures. Finally, we computed the mean and median pairwise distance, or the distance between all pairs of points assigned to a single n-gram. ANN and pairwise calculations were done using the `spatstat`package in R [50]. Similarly to *Standard Distance*, we hypothesize that the average spatial distance between the closest (2nd closest and 3rd closest) n-grams that describe the same neighborhood will be unique for neighborhoods, thus allowing us to include this measure in our approach to neighborhood name identification.

4.1.2. Spatial Homogeneity

The spatial homogeneity of each geo-tagged n-gram was calculated through a binned approach to Ripley's L or variance stabilized Ripley's K [51,52]. Ripley's L measures the degree of clustering across different spatial scales. Specifically, our approach split the resulting Ripley's L clustering function into ten 500-m segments and averaged the range of clustering values for each n-gram within each segment. Figure 3 shows the binned Ripley's L approach for two n-grams in Washington, DC, one a neighborhood name (Columbia Heights) and the other what should be an a-spatial term (wood flooring). From a conceptual perspective, one might expect that most neighborhood names will show a higher than expected degree of clustering around a certain distance mark. Higher than expected clustering at a small distance might identify landmarks, while clustering at a large distance might be useful for the identification of metro stations. Ripley's L allows us to assess clustering vs. expected clustering across these different distances. This approach of binning spatial homogeneity functions has been employed successfully in differentiating point of interest types (e.g., bars vs. police stations) [53].

In addition to the ten binned relative clustering values, the kurtosis and skewness measures for each Ripley's L function over 5 km were recorded for each n-gram. The kurtosis and skewness provide overall measures of the Ripley's L function instead of a single measure based on binned distance.

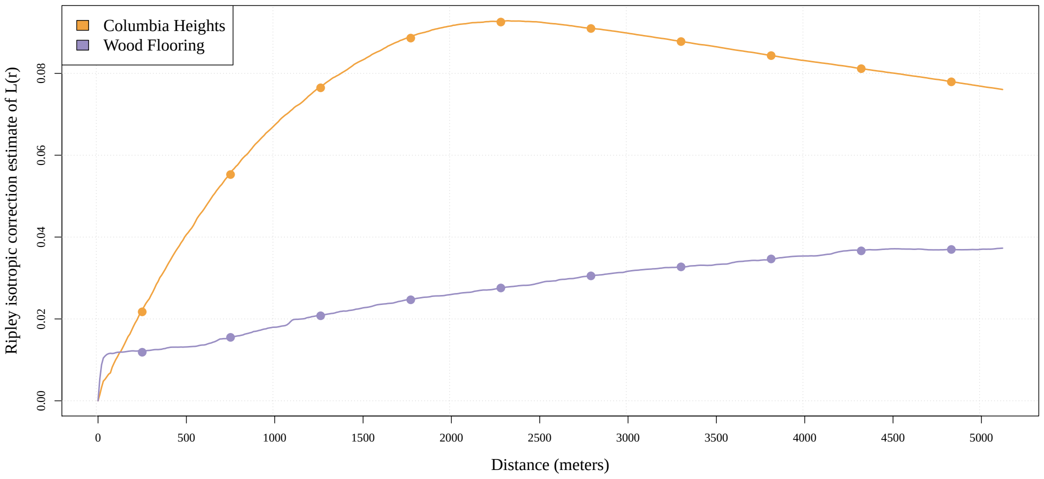


Figure 3. Ripley's L function over 5 km for two n-grams, Columbia Heights (a neighborhood name) and wood flooring. The points show the averaged 'binned' values over 500 m.

4.1.3. Convex Hull

The convex hull [54] is the smallest convex set (of listings in this case) that contains all listings. Using the *chull R package*, we computed the area of the convex hull for all geo-tagged n-grams in our dataset, as well as the density of the convex hull based on the number of listings in the set divided by the area. These two measures offer a very different description of the property listings as they represent the maximum area covered by all listings. Convex hull area simply assigns a numerical value for the region covered by all listings. This measure is heavily impacted by outliers (e.g., random mention of a neighborhood across town), as one occurrence can drastically alter the area of the convex hull. Conceptually, the density of points within the convex hull is useful for comparing n-grams, as we would expect to find a higher than average density of points within a region identified as a neighborhood, compared to an a-spatial term such as wood flooring.

4.1.4. Spatial Autocorrelation

As part of our initial exploratory analysis for this project, spatial autocorrelation was investigated as a meaningful spatial feature due to its potential relatedness to neighborhood names. This form of measurement, however, is substantially different from many of the other measures mentioned here, as there is really no way to report spatial autocorrelation through a single value per n-gram. As with other measures of correlation, this inference statistic involves interpreting the results through a number of values, not least of which are p-values and Z-scores. Running Moran's I across our set of geo-tagged n-grams, we found the results inconclusive overall. At least half of the values for global Moran's I were not of a high enough significance including what many would consider 'prototypical' neighborhood names in Washington, DC, such as Georgetown and Capital Hill. For these reasons, we elected to leave Moran's I after the exploratory phase of analysis and did not use it in the final random forest models.

4.2. Data Setup

In setting up the data for input to a prediction and identification model, we calculated the above statistics for each n-gram in our dataset. These values were then combined into a single data table, one for each city with rows as n-grams and columns as statistical measures. From this point on, we will refer to the spatial statistic values as predictor variables. The n-grams in the common neighborhood names dataset (see Section 3.2) were programmatically compared against all n-grams in the merged dataset, and matches were recorded. While in an ideal world, this would have resulted in clean matches, a number of matches were not made due to slight misspellings, abbreviations (e.g., Columbia Hts. for Columbia Heights) and variations of n-grams that include the neighborhood names (e.g., to Dupont or Logan Circle from). These neighborhoods were identified and matched manually by two researchers, and disagreements were resolved by a third person. Again, manual matching was only based on the common neighborhood names, not all potential neighborhood names. As a result of this process, all n-grams were given a binary value, either identified as neighborhood matches or not.

4.3. Individual Predictors

Having calculated spatial statistics values for each of our n-grams based on the methods described in the previous sections, we now turn to RQ1, examining how well each individual statistic performs at identifying a neighborhood from within the set of descriptive n-grams. All predictor variables were normalized to between 0 and 1 based on their minimum and maximum values to allow for simpler comparison. The Pearson's correlation matrix of all predictors and neighborhood matches is shown in Table A1, Appendix A. A single star (*) indicates $p < 0.1$ and three stars (***) no significance, and all other values are significant to $p < 0.01$. Notably, there tends to be a negative correlation between the mean and median nearest neighbor values and neighborhood match and a positive correlation to all binned Ripley's L variables.

Each of the individual variables was then used to predict which of the n-grams was a neighborhood name, and the accuracy of each prediction was recorded. The F_{score} (Equation (2)), the harmonic mean of precision and recall, was used to assess prediction power. Accuracy measures were recorded at 0.05 threshold increments, meaning the first time the model was run, any predictor variable value above (and including) 0.05 was considered a match, and everything below was not.

The threshold value that produced the best F_{score} for each predictor variable was identified. The best scores are reported in Section 5.1.

$$F_{score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

4.4. Random Forest Ensemble Learning

In addressing RQ2, we now combine our predictor variables and take a machine learning approach using a random forest [55,56] ensemble learning method to identify neighborhood names within our n-gram dataset. Random forest models have proven quite successful in numerous other classification and regression tasks involving geographic components [57–59]. Random forest models are touted as being better suited to lower dimensional problems (as opposed to support vector machines for example) and correct for over-fitting, which tends to happen when training decision trees. Random forest is a supervised learning algorithm that approaches the problem of learning through a combination of decision trees. Multiple decision trees are built and merged together to get a more accurate and stable prediction. This follows the idea that en masse, these individual learners will combine to form a learner better than its individual parts. In this work, we used the R *randomForest* package ported from Breiman et al.'s original Fortran code (https://cran.r-project.org/package=randomForest).

4.4.1. Training and Testing

The first random forest model was trained with a randomly selected 50% of the n-grams in the Washington, DC, dataset (both neighborhood matches and non-neighborhood matches) and tested for accuracy against the remaining 50% of the data. This combination of training and testing was done 100 times in order to produce robust measures for the results, each time training on a different randomly selected 50% of the Washington, DC, data. When each model was trained, it was applied to the testing data in order to predict which n-grams were neighborhoods and which were not. This was done using a probability method of prediction with the resulting probability for each n-gram bounded between 0 (not a neighborhood) and 1 (definitely a neighborhood). The F-scores (Equation (2)), were recorded at 0.05 probability increments every time the prediction model was run, and the probability threshold that produced the best mean F-score was identified. By way of comparison, we also randomly selected n-grams as neighborhood matches in our dataset and trained a separate random forest model on these data. The purpose of this was to provide a baseline on which our true neighborhood matching model could be compared. The same number of n-grams was chosen, at random, so as to provide comparable results.

The purpose of this research is not only to show that spatial features can be used to identify existing neighborhoods, but also can be used in the identification of less common and previously unidentified neighborhoods (RQ3). To this end, the model probability threshold should be adjusted to alter the precision as we want to identify those false positives that may actually be neighborhood names, but may either not have been found in our dataset, were not matched to a common neighborhood name or are more colloquial, or unofficial, neighborhood names. After computing the optimal threshold value (based on F-score), we manually examined the false positives, those that the model identified as neighborhoods, but were not considered matches to our common neighborhood list. Through manual inspection, we discovered a number of interesting false positives. Many were neighborhood names that appeared in one or more of the curated neighborhood lists, but did not appear in enough sources to be considered part of the common neighborhood set. Provided these new neighborhood matches, we build a subsequent random forest model, this time with the addition of those newly identified false positive n-grams that are in fact neighborhood names. The resulting accuracy of both of these models is reported in Section 5.

4.4.2. Variable Importance

The random forest models described in the previous section were constructed with 500 tries and 4 variables tried at each split. As a result of these splits, the model produced a ranking of the predictor variables based on their contribution to the overall accuracy of the model. Figure 4 shows the importance of these variables by way of the mean decrease in the Gini index of node purity. What this demonstrates is that some variables are more useful than others at splitting mixed label nodes into pure single-class nodes. The higher the value, in this case, the more important the predictor variable is to the model's accuracy. We see here that larger bin distances of Ripley's L are substantially more important to the success of the model than the mean nearest neighbor measures, for example. To some extent, this mirrors the ranking of individual predictor accuracy that is reported in Section 5.1.

4.5. Evaluation

Having trained two random forest models based on housing rental n-grams from Washington, DC, we next turn our attention to RQ4, namely evaluating the accuracy of such a model using data from two other North American cities, Seattle, WA, and Montréal, QC. As described in Section 4.2, the predictor variables for each n-gram were merged into city-specific datasets and matched against existing common neighborhood lists for their respective cities. Manual inspection and matching were done as before, and those n-grams that matched neighborhood names were marked as matches, while all others were not. The random forest model trained on the Washington, DC, data was then tested against the geo-tagged Seattle and Montréal n-grams independently using the highest performing probability threshold value from the Washington, DC, testing results.

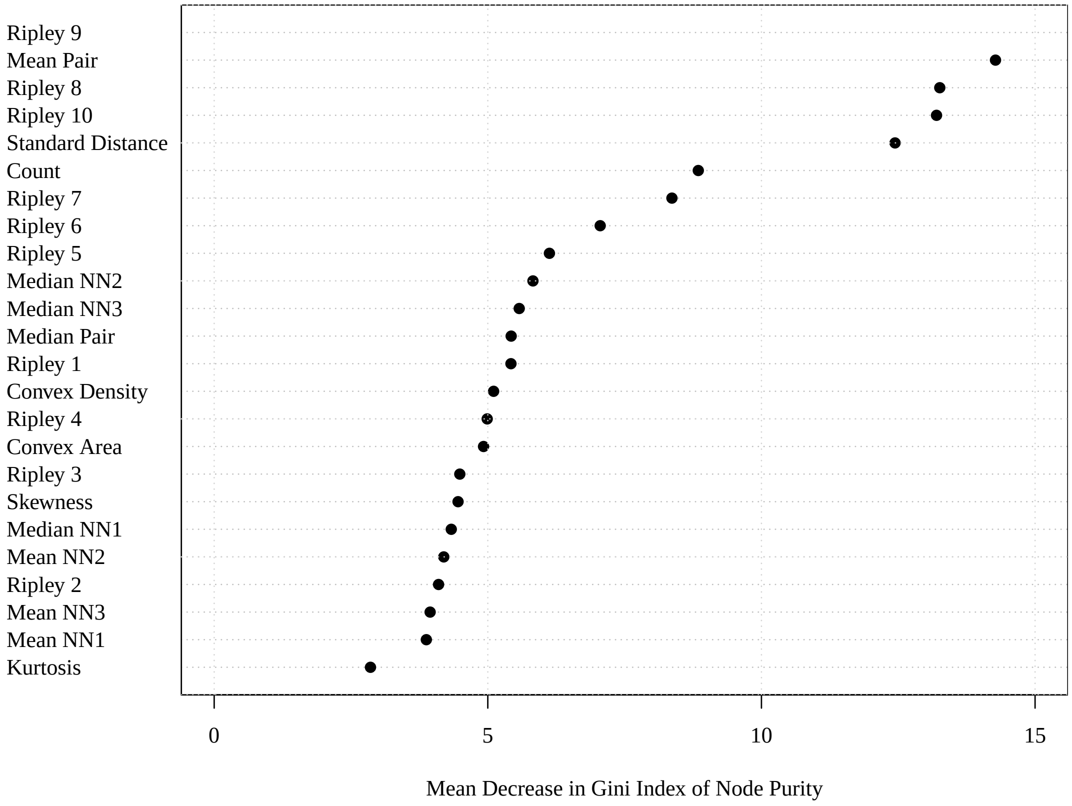


Figure 4. Mean decrease in the Gini index of node purity due to shuffling of values within the predictive variable. NN, n-gram neighbor.

5. Results

In this section, we present the results of the methods presented in the previous section. Specifically, we focus on the accuracy values of the individual predictors, as well as the combined random forest model.

5.1. Individual Predictors

The maximum F_{score} accuracy values for the individual predictor variables are shown in Table 3. On average, the accuracy of each predictor variable independently was not high. However, the binned Ripley's L variables produced the best predictive results with the 4500-m bin (Ripley 9) producing the best $F_{score}$of 0.633 with a recall and precision of 0.724 and 0.562, respectively. These results demonstrate that even without employing a more complex ensemble learning technique, a single spatial statistics measure could perform reasonably well at differentiating neighborhoods from non-neighborhoods. Notably however, not all spatial statistics were useful for this endeavor independently. Next, we explore combining these individual predictor variables with the purpose of improving neighborhood name identification.

Table 3. Max F-scores for individual predictor variables trained and tested on data from Washington, DC.

Measure	Max F-Score
Standard Distance	0.047
Count	0.083
Mean NN1	0.047
Mean NN2	0.047
Mean NN3	0.047
Med.NN1	0.050
Med. NN2	0.050
Med. NN3	0.050
Mean Pair	0.047
Med. Pair	0.047
Ripley 1	0.099
Ripley 2	0.279
Ripley 3	0.405
Ripley 4	0.500
Ripley 5	0.548
Ripley 6	0.570
Ripley 7	0.587
Ripley 8	0.624
Ripley 9	0.633
Ripley 10	0.624
Kurtosis	0.101
Skewness	0.047
Convex Area	0.047
Convex Density	0.144

5.2. Ensemble Learning

The first step in matching common neighborhoods to n-grams (both programmatically and manually) resulted in 59 neighborhood names, out of 95, being identified in the 3612 unique n-grams in Washington, DC. Of these, 30 were direct matches, with 29 indirect, manually-identified matches. There are a number of reasons why not all common neighborhood names were found in our dataset, which will be discussed in Section 6.

The first random forest model was trained on the predictor variables of n-grams tagged as either common neighborhoods or not. The resulting averaged F_{score} is shown in Table 4. This value is based on a prediction probability threshold of 0.35. This is a high F-score given the noisiness of the user-generated content on which the model was constructed. The recall value indicates how well the model did at identifying known neighborhoods, whereas the precision tells us how well the model did at identifying neighborhoods n-grams as neighborhoods and non-neighborhood n-grams as such. As mentioned in Section 4, these results allowed us to re-examine our dataset and uncover neighborhood names that were not previously identified, i.e., those that did not appear in our common set, but rather one of the individual neighborhood sources such as Wikipedia. Through manual inspection, we increased the number of neighborhood/n-gram matches in our dataset and trained a new random forest model on the data. The results of this second random forest model are shown in the second row of Table 4. The F_{score} improved as had both the precision and the recall with the largest increase occurring in the recall value.

As a base-line, we also included the F_{score} results of a random forest trained on randomly-assigned matches (not necessarily neighborhood names). As expected, the results were considerably lower than the previous two models with an accuracy of roughly 0.05.

Table 4. F-score, precision and recall values for two random forest models trained and tested on listings from Washington, DC. Accuracy values for a model built on random assignments are also shown for comparison.

Model	F-Score	Precision	Recall
Common matched neighborhoods	0.807	0.845	0.777
Common + secondary matches	0.872	0.863	0.882
Randomly assigned matches	0.047	0.063	0.037

5.3. Identifying Neighborhoods in Other Cities

Equipped with the best performing random forest model trained and tested on the Washington, DC, n-grams, we then tested it against our two other North American cities, as outlined in RQ4.

5.3.1. Predicting Seattle Neighborhoods

The first row of Table 5 shows the results of the random forest model trained on Washington, DC, n-grams. This first model used the common Seattle neighborhoods as matches. As was reported in the previous section, the results of the first RF model prediction led to an investigation of the precision of the model resulting in the identification of a number of neighborhoods that were not previously identified as such. This was rectified, and the model was run again, producing the values shown in the second row of the table.

The third row of Table 5 presents the results of a random forest model trained on half of the Seattle data rather than the Washington, DC, n-grams, and tested on the other half of the Seattle data. These results indicate that while the DC-trained RF models do perform well at predicting Seattle neighborhoods, a model trained on local data still performs better.

Table 5. F-score, precision and recall values for two random forest models trained on listings from Washington, DC, and tested on listings from Seattle, WA (the first two rows). The last row shows the results of a model trained and tested on listings from Seattle, WA.

Model	F-Score	Precision	Recall
Common matched neighborhoods	0.671	0.625	0.724
Common + secondary matches	0.733	0.702	0.767
Trained on Seattle (common)	0.786	0.782	0.791

5.3.2. Predicting Montréal Neighborhoods

In many ways, Seattle, WA, is very similar to Washington, DC. Both are major metropolitan, predominantly English-speaking cities. Both host populations of roughly 700,000 and have similar population densities, median age and median income. To test the robustness of the DC-based random forest model, we chose to test it against a very different city, namely Montréal, Quebec, in Canada. Montréal is a bilingual French/English speaking island city, boasting French as its official language. Montréal has a population of roughly two million (on island) residents. craigslist rental housing listings in Montréal are written in either French or English and often both. In addition to all of this, the city has a historically unique rental market with the majority of leases beginning and ending on 1 July [60]. Given the data collection dates, far fewer rental postings were accessed for the city compared to both Washington, DC, and Seattle, WA. These factors combined, this city offers a unique dataset on which to test our model.

As shown in Table 6, the first random forest model built from the DC n-grams produces an F_{score} of roughly 0.4. Upon examining the results of this model, additional non-common neighborhoods were identified, and a second model was run, resulting in a slightly higher F-score. While clearly not as high as the Seattle results, these values are still substantially higher than a model built on randomly-matched n-grams. As was the case with Seattle, a model built on local Montréal data produced the best results with an F-score of 0.655 and notably a recall inline with that of Seattle's. A set of n-grams identified as neighborhoods by this model is presented in Appendix B.

Table 6. F-score, precision and recall values for two random forest models trained on listings from Washington, DC, and tested on listings from Montréal, QC. The last row shows the results of a model trained and tested on listings from Montréal, QC.

Model	F-Score	Precision	Recall
Common matched neighborhoods	0.397	0.353	0.453
Common + secondary matches	0.483	0.412	0.583
Trained on Montréal (common)	0.655	0.559	0.792

6. Discussion

The results presented in this work offer evidence as to how neighborhoods can be identified by the spatial distribution of rental housing advertisements. These findings demonstrate that identification of a sample of common neighborhood names with spatial distribution patterns can be used to predict additional, less common neighborhood names within a given city accurately. Furthermore, we find that an array of spatial distribution measures from neighborhoods identified in one part of North America can be used to train a machine learning model that can then be used to identify neighborhoods on another part of the continent accurately. While rental housing data from local listings produces a more accurate model, we find that this model can also span linguistic barriers, admittedly producing less accurate, but quite significant, results. In this section, we further delve into the nuanced results of using such a machine learning approach and identify unique aspects and biases within the dataset.

6.1. False Positives

The F-score values presented in Tables 4–6 depict an overall view of the accuracy of the model, but omit the nuances of the actual on-the-ground data and neighborhoods. Specifically some regions of the city are better represented by the dataset than others, and this is reflected in the analysis results. The size, dominance and popularity of a neighborhood all impact the probability of a neighborhood being identified in the n-gram datasets. For example, many of the historic neighborhoods in Washington, DC (e.g., Georgetown, Capital Hill, Brightwood), were clearly represented in the original data, thus resulting in high accuracy results. These prevalent neighborhoods then had a much larger impact in contributing to the construction of the neighborhood identification model. This often meant that smaller and less dominant neighborhoods, e.g., Tenleytown, were less likely to be identified through the machine learning process and that other, non-neighborhood regions were more likely to be identified.

While the model performed well provided training data from within the city, there was an expected set of false positives (see Table 7 for examples). Further examination of these false positives allows us to categorize them into six relatively distinct groupings. Landmarks such as the Capitol Building or the White House were falsely identified as neighborhoods given the importance of these landmarks within Washington, DC. Many housing rental listing specifically mentioned a proximity to these landmarks, thus resulting in spatial distribution measures similar to those of neighborhoods. Similarity, some important streets, academic institutions and popular transit stations were labeled as neighborhoods given their dominance within a region of the city. This reiterates the argument from the introduction of this paper that neighborhoods are simply regions with distinct characteristics that are given a descriptive name by inhabitants and visitors. It therefore follows that many neighborhood names come from important streets (e.g., George Ave.), transit stations (Union Station) and Universities (Howard). While many of these n-grams identified as neighborhoods by our model were labeled as false positives, there is an argument to be made that the n-grams do exist as neighborhood names.

Though many of these false positives can be explained given knowledge of the region, spatial dominance of a certain term or prevalence of the geographic feature, a small portion of the false positives appeared to be non-spatially related. For example, terms such as concierge and du vieux appear to not be related to any geographic feature or place within a city and rather are n-grams within the data that happen to demonstrate spatial distribution patterns similar to neighborhoods. In addition to these, a number of real-estate company names were falsely identified as neighborhoods in our initial models given that many real estate companies are focused specifically on one region of a city. These real estate company related n-grams were removed early in the data cleaning process.

Table 7. Examples of n-grams falsely identified as neighborhood names split by city (columns) and category (rows).

Category	Washington, DC	Seattle, WA	Montréal, QC
Landmarks	Capitol Building	Space Needle	Place Jacques-Cartier
Academic Institution	Catholic University	University of Washington	McGill University
Streets	Wisconsin Ave.	Summit Ave.	Cavendish Blvd.
Broader Regions	National Mall	Waterfront	Saint-Laurent River
Transit Stations	Union Train Station	King Street Station	Jolicoeur Station
Companies	Yes, Organic	Amazon	Atwater Market
Misc.	blvd	concierge	du vieux

6.1.1. Washington, DC

Washington, DC, is a particularly interesting city, arguably representative of many east coast U.S. cities, namely in the way that many populated regions run into one another. Washington, DC, itself is part of the larger Metro DC area, which includes cities in the neighboring states of Virginia and Maryland. Since rental housing listings were clipped to the buffered boundary of Washington, DC; this meant that some neighborhoods were identified by the model that do not appear in the common DC neighborhood set, as they technically exist outside the district boundary. Examples of such neighborhoods identified by our model are Alexandria and Arlington in Virginia and Silver Spring and College Park in Maryland.

Within the district boundaries, a number of neighborhoods were identified through the machine learning model that did not originally exist in the common neighborhoods set for the district such as Cleveland Park and University Heights, both labeled as neighborhoods on Wikipedia. Moreover, alternative or secondary names for neighborhoods were identified in the results, such as Georgia Ave., a secondary name for Petworth, and Howard, the name of a university that has taken on a colloquial reference to a sub-region within or overlapping the Shaw neighborhood. While many of the false positives were smaller than a typical neighborhood area (e.g., Capitol Building), the ensemble learning model also identified a number of larger regions, such as the National Mall, an important tourist attraction within Washington, DC, and the broader Northeast region of the district. Notably, Washington addresses are divided into quadrants based on intercardinal directions. As stated previously, a few major streets were identified, namely Wisconsin Ave., Connecticut Ave. and Rhode Island Ave., all major thoroughfares leading from outside of the district to the city center. As demonstrated with Georgia Avenue, many street names have taken on neighborhood-like statuses being used to describe regions of similar socioeconomic status, demographics or other characteristics.

6.1.2. Seattle, WA

Further qualitative discussion of the n-gram neighborhood identification results in Seattle expose some unique aspects of the city. As was the case in Washington, DC, investigation of false positives exposed a number of neighborhood names that did exist as neighborhoods in one of the neighborhood datasets (e.g., Wikipedia), but not in the common neighborhood set. Examples of these are Columbia City, Lake Union and Wallingford. Neighborhoods outside the Seattle city boundary such as Bothell or Mercer Island were also identified, as were neighborhoods such as Lincoln Park, a large park that has given rise to a new neighborhood name, and Alki Beach, a sub-neighborhood within West Seattle along the waterfront. While popular streets, e.g., Summit and Anderson, were labeled as neighborhoods, the biggest difference in false positives compared to Washington, DC, is an increase in company/foundation names identified as neighborhoods. Amazon.com Inc, The Bill and Melinda Gates Foundation and Microsoft (outside of Seattle) were each clearly identified as neighborhoods, and the first Starbucks location (in Pike Place Market) was initially identified as a neighborhood when the model was built on local training data.

6.1.3. Montréal, QC

Examination of the n-gram results in Montréal produced some interesting insight into how a machine learning model such as this is actually quite language independent, at least as it relates to English and French rental listings. Importantly, though a single rental listing may contain both French text and English translation, the neighborhood names in Montréal are either in French or in English, not both, at least according to the reference datasets we employed. This means that each neighborhood does not have two names (one in each language) and implies that a model does not have to be adjusted for sparsity in the labels, but rather can be run as is.

As in the previous two cities, non-common neighborhoods were identified through the model such as Mile End and Quartier Latin, as well as academic institutions such as Loyola College/High School. Colloquial references to existing neighborhoods such as NDG for Notre-Dame-de-Grâce were also identified, as were many important street names in Montréal such as Crescent or Ste.-Catherine. Interestingly, since these street names were referenced either in French or English, the n-gram, which includes the generic type, e.g., street or rue (in French), is often not identified as a neighborhood, only the specific name. This is notably different than the other two English-language-dominant cities.

6.2. Listing Regional Bias and False Negatives

In the previous section, we discussed a number of the false positives and examined some possible explanations. Here, we investigate instances where our model did not correctly identify common neighborhoods, as well as some of the potential reasons for this. Data from Washington, DC, in particular are the subject of further examination, and Figure 5 presents a good starting point for this discussion.

The regions represented in purple in this figure are neighborhoods in our common neighborhood set that were correctly identified in the initial RF model. The regions shown in orange are those neighborhoods that did not appear in the common neighborhood set, but did appear in at least one of the source-specific neighborhood datasets (government-defined neighborhoods in this case). These are the neighborhoods that were successfully identified by the first iteration of the RF model that were then properly tagged as neighborhoods for input into the second RF model (for use in training a model for other cities). Green regions of the map depict those neighborhoods that were never identified (false negatives), or did not exist, in the n-grams from the craigslist data. Dark gray regions can be ignored as they represent uninhabitable space such as the Potomac and Anacostia rivers, Rock Creek Park, Observatory Circle and Joint Base Anacostia-Bolling (military controlled). In observing Figure 5, there is a clear geographic bias between the true positives (blue and orange) and unmentioned or false negatives (green). The green regions are predominantly in the east-southeast region of Washington, DC, east of the Anacostia River in what is municipally defined as Wards 7 and 8 (Washington, DC's, planning department splits the district into eight wards). In referencing the 2015 American Community Survey data, we find that Wards 7 and 8 contain the largest number of residences in the district living below the federal poverty line. In addition, the neighborhoods in Wards 7 and 8 contain a mean of 232.3 (median 290) public housing units (housing provided for residents with low incomes and subsidized though public funds). By comparison, neighborhoods in all other wards list a mean of 173.2 (median 13) public housing units.

Further investigation into the neighborhood names in Wards 7 and 8 shows that none of the names or reasonable partial matches of the names occur in the rental listing-based craigslist dataset. Either listings did not occur in those neighborhoods, were too few and thus removed from the dataset during cleaning or the neighborhood names themselves were not stated in the listings. The mean number of listings per square kilometer or neighborhoods in Wards 7 and 8 is 0.0063 (median 0.0054, SD 0.0035), whereas the rest of the neighborhoods showed a mean of 0.0526 (median 0.0352, SD 0.0539), suggesting that the lack of n-gram neighborhood identification was due to the lack of listings, not necessarily missing names in the text or false negatives. This bias in rental listings related to poverty supports existing research in this area [61].

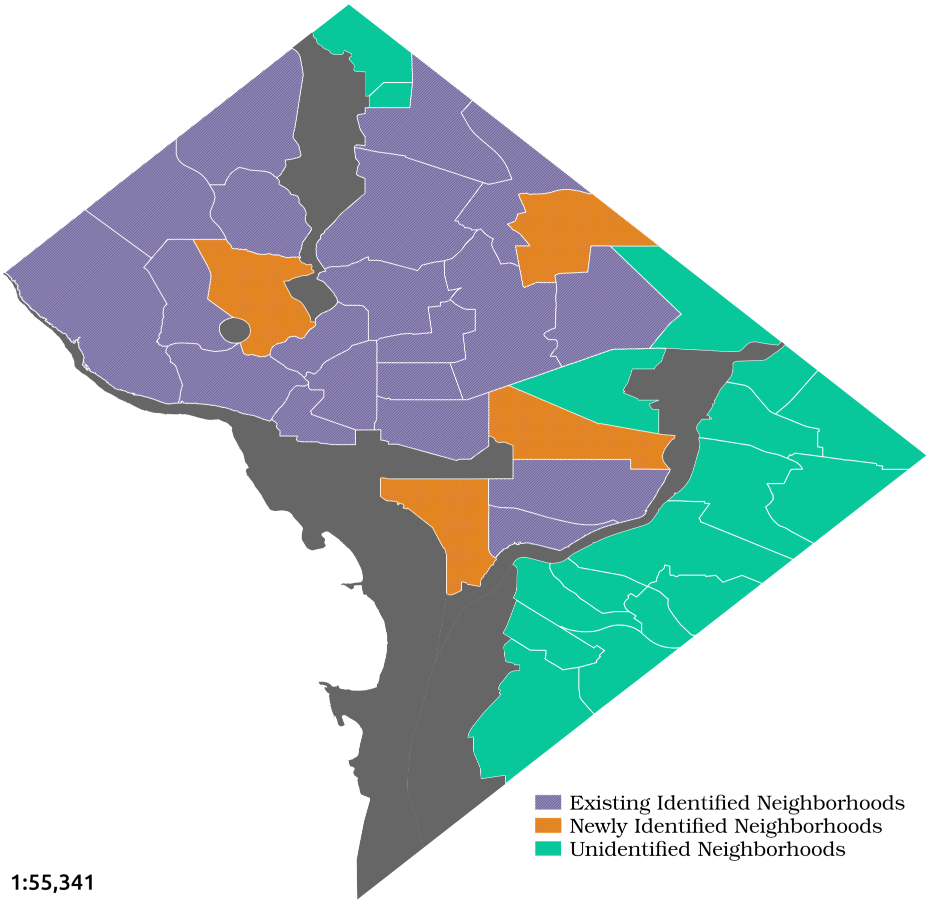

Existing Identified Neighborhoods
Newly Identified Neighborhoods
Unidentified Neighborhoods
1:55,341

Figure 5. Identified and unidentified neighborhoods in Washington, DC.

7. Conclusions and Future Work

Neighborhoods are an odd concept related to human mobility and habitation. They are difficult to quantify and, within the domain of geographical sciences, have been historically ill-defined. Neighborhoods are given meaning by the people that inhabit a region based on a set of common or shared characteristics. Part of the problem is that a top-down approach to defining a neighborhood is fraught with problems, and the resulting names and boundaries are often at odds with the citizens that live and work within them. In this work, we take a bottom-up and data-driven approach to identifying neighborhood names within urban settings. Using geotagged rental property listings from the popular classifieds platform, craigslist, we demonstrate that neighborhood names can be identified from natural language text within housing rental listings (RQ1). Using an ensemble learning approach based on spatial descriptive statistics, we demonstrate that it is possible to differentiate neighborhood names from other descriptive natural language terms and phrases (RQ2). Three unique cities within North America are used as focal study sites with listings from one (Washington, DC) being used to train a model that is tested on the other two (Seattle, WA, and Montréal, QC). The results of this approach demonstrate that neighborhood names can successfully be identified within the trained city and across different cities (RQ4). In some cases, new, alternative or previously unidentified neighborhood names are proposed based on this approach (RQ3). Finally, the biases associated with these data are further exposed through this method (RQ5) and are discussed in further detail.

As mentioned when discussing the biases associated with this approach, these data really represent the property listers' views of the city. In most cases, the people listing these properties represent a small subset of the city's population, either property owners or real estate agents, both of which tend to exist within a narrow socio-economic group. The neighborhood names identified in the results are therefore heavily influenced by this group. While the methods presented are agnostic to the underlying source of the data, it is important to understand that the neighborhood results depicted in this work are reliant on data contributed to a single online platform.

Similarly, the three example cities used in this research are all within North America. Future work should examine how the results and accuracy values are affected by a change in location. European Cities such as Berlin, for example, could be vastly different given the unique historical context through which the city is understood. Additional work will focus on increasing the diversity of the data sources, languages of the rental property listings and inclusion of additional structured content (e.g., number of bedrooms, price, etc.). From a statistical perspective, further research will attempt to reduce the dimensionality of this approach by further investigating the correlations between the various spatial statistical measures. Furthermore, a deeper investigation into the role of spatial-autocorrelation, specifically the lack of significance in the results of Moran's I analysis, will be conducted, as this lack of significance is quite interesting and surprising to the researchers. Finally, this work presents the first step of identifying neighborhood names. Our next step is to identify the boundaries associated with these neighborhood names with the goal of developing local listing-based neighborhood datasets.

Author Contributions: All authors contributed substantially to this work. Conceptualization and methodology was done by all listed authors. Formal Analysis, Validation, and Data Curation was executed by G.M. and Z.L. Writing—Original Draft Preparation was done by G.M.; Writing—Review & Editing was done by all authors; Supervision and Project Administration was done by G.M.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Pearson's correlation matrix for all predictive spatial statistics measures. * indicates $p < 0.1$; *** indicates no significance; and all other values are significant to $p < 0.01$.

	N Match	SD	Count	Mean NN1	Mean NN2	Mean NN3	Med. NN1	Med. NN2	Med. NN3	Mean Pair	Med. Pair	Ripley 1	Ripley 2	Ripley 3	Ripley 4	Ripley 5	Ripley 6	Ripley 7	Ripley 8	Ripley 9	Ripley 10	Kurtosis	Skewness	Convex Area	Convex Density
N Match	1.000	−0.363	0.029 *	−0.125	−0.131	−0.136	−0.089	−0.108	−0.114	−0.371	−0.337	0.109	0.289	0.381	0.427	0.447	0.443	0.435	0.429	0.422	0.402	0.0212 ***	−0.121	−0.201	−0.001 ***
SD	−0.363	1.000	−0.041 *	0.179	0.201	0.210	0.137	0.213	0.237	0.981	0.932	−0.166	−0.362	−0.505	−0.611	−0.683	−0.738	−0.774	−0.802	−0.823	−0.844	0.056	0.017 ***	0.419	−0.309
Count	0.029 *	−0.041 *	1.000	−0.793	−0.820	−0.831	−0.428	−0.510	−0.570	0.010 ***	−0.039 *	−0.296	−0.219	−0.139	−0.086	−0.047	−0.010 ***	0.022 ***	0.044	0.053	0.065	−0.174	−0.029 *	0.676	0.857
Mean NN1	−0.125	0.179	−0.793	1.000	0.961	0.948	0.779	0.807	0.824	0.169	0.165	−0.138	−0.204	−0.237	−0.234	−0.233	−0.244	−0.259	−0.255	−0.234	−0.221	−0.083	0.101	−0.257	−0.921
Mean NN2	−0.131	0.201	−0.820	0.961	1.000	0.984	0.739	0.824	0.842	0.188	0.184	−0.117	−0.198	−0.241	−0.243	−0.245	−0.261	−0.280	−0.278	−0.259	−0.246	−0.095	0.126	−0.276	−0.943
Mean NN3	−0.136	0.210	−0.831	0.948	0.984	1.000	0.732	0.813	0.852	0.194	0.186	−0.104	−0.186	−0.234	−0.240	−0.245	−0.263	−0.281	−0.280	−0.263	−0.252	−0.081	0.123	−0.291	−0.944
Med. NN1	−0.089	0.137	−0.428	0.779	0.739	0.732	1.000	0.906	0.871	0.178	0.126	−0.533	−0.539	−0.495	−0.417	−0.360	−0.337	−0.321	−0.292	−0.249	−0.218	−0.335	0.143	0.011 ***	−0.607
Med. NN2	−0.108	0.213	−0.510	0.807	0.824	0.813	0.906	1.000	0.964	0.254	0.209	−0.446	−0.505	−0.505	−0.456	−0.417	−0.401	−0.393	−0.366	−0.323	−0.295	−0.317	0.171	−0.021 ***	−0.703
Med. NN3	−0.114	0.237	−0.570	0.824	0.842	0.852	0.871	0.964	1.000	0.273	0.235	−0.369	−0.445	−0.468	−0.437	−0.412	−0.406	−0.404	−0.382	−0.342	−0.316	−0.262	0.162	−0.077	−0.742
Mean Pair	−0.371	0.981	0.010 ***	0.169	0.188	0.194	0.178	0.254	0.273	1.000	0.955	−0.242	−0.445	−0.591	−0.694	−0.762	−0.810	−0.838	−0.857	−0.870	−0.881	−0.001	0.047	0.465	−0.276
Med. Pair	−0.337	0.932	−0.039 *	0.165	0.184	0.186	0.126	0.209	0.235	0.955	1.000	−0.147	−0.352	−0.510	−0.627	−0.710	−0.768	−0.811	−0.847	−0.877	−0.903	0.072 ***	0.004 ***	0.364	−0.285
Ripley 1	0.109	−0.166	−0.296	−0.138	−0.117	−0.104	−0.533	−0.446	−0.369	−0.242	−0.147	1.000	0.906	0.749	0.589	0.478	0.414	0.370	0.319	0.268	0.230	0.578	−0.189	−0.555	−0.042 *
Ripley 2	0.289	−0.362	−0.219	−0.204	−0.198	−0.186	−0.539	−0.505	−0.445	−0.445	−0.352	0.906	1.000	0.930	0.808	0.708	0.641	0.588	0.532	0.479	0.440	0.639	−0.337	−0.581	0.071
Ripley 3	0.381	−0.505	−0.139	−0.237	−0.241	−0.234	−0.495	−0.505	−0.468	−0.591	−0.510	0.749	0.930	1.000	0.948	0.875	0.809	0.754	0.698	0.646	0.607	0.569	−0.393	−0.544	0.156
Ripley 4	0.427	−0.611	−0.086	−0.234	−0.243	−0.240	−0.417	−0.456	−0.437	−0.694	−0.627	0.589	0.808	0.948	1.000	0.969	0.916	0.862	0.812	0.767	0.731	0.429	−0.371	−0.503	0.204
Ripley 5	0.447	−0.683	−0.047	−0.233	−0.245	−0.245	−0.360	−0.417	−0.412	−0.762	−0.710	0.478	0.708	0.875	0.969	1.000	0.974	0.930	0.888	0.850	0.817	0.298	−0.310	−0.469	0.238
Ripley 6	0.443	−0.738	−0.010 ***	−0.244	−0.261	−0.263	−0.337	−0.401	−0.406	−0.810	−0.768	0.414	0.641	0.809	0.916	0.974	1.000	0.979	0.946	0.912	0.878	0.192	−0.236	−0.445	0.271
Ripley 7	0.435	−0.774	0.022 ***	−0.259	−0.280	−0.281	−0.321	−0.393	−0.404	−0.838	−0.811	0.370	0.588	0.754	0.862	0.930	0.979	1.000	0.984	0.954	0.920	0.112	−0.160	−0.425	0.300
Ripley 8	0.429	−0.802	0.044	−0.255	−0.278	−0.280	−0.292	−0.366	−0.382	−0.857	−0.847	0.319	0.532	0.698	0.812	0.888	0.946	0.984	1.000	0.986	0.957	0.039 *	−0.090	−0.401	0.314
Ripley 9	0.422	−0.823	0.053	−0.234	−0.259	−0.263	−0.249	−0.323	−0.342	−0.870	−0.877	0.268	0.479	0.646	0.767	0.850	0.912	0.954	0.986	1.000	0.986	−0.012 ***	−0.050	−0.380	0.312
Ripley 10	0.402	−0.844	0.065	−0.221	−0.246	−0.252	−0.218	−0.295	−0.316	−0.881	−0.903	0.230	0.440	0.607	0.731	0.817	0.878	0.920	0.957	0.986	1.000	−0.037 *	−0.035 *	−0.360	0.312
Kurtosis	0.0212 ***	0.056	−0.174	−0.083	−0.095	−0.081	−0.335	−0.317	−0.262	−0.001 ***	0.072	0.578	0.639	0.569	0.429	0.298	0.192	0.112	0.039 *	−0.012 ***	−0.037 *	1.000	−0.764	−0.277	−0.042 *
Skewness	−0.121	0.017 ***	−0.029*	0.101	0.126	0.123	0.143	0.171	0.162	0.047	0.004 ***	−0.189	−0.337	−0.393	−0.371	−0.310	−0.236	−0.160	−0.090	−0.050	−0.035 *	−0.764	1.000	0.060	−0.077
Convex Area	−0.201	0.419	0.676	−0.257	−0.276	−0.291	0.011 ***	−0.021 ***	−0.077	0.465	0.364	−0.555	−0.581	−0.544	−0.503	−0.469	−0.445	−0.425	−0.401	−0.380	−0.360	−0.277	0.060	1.000	0.294
Convex Density	−0.001 ***	−0.309	0.857	−0.921	−0.943	−0.944	−0.607	−0.703	−0.742	−0.276	−0.285	−0.042 *	0.071	0.156	0.204	0.238	0.271	0.300	0.314	0.312	0.312	−0.042 *	−0.077	0.294	1.000

Appendix B

Neighborhood names (both true and false positives) as identified by the random forest ensemble learning model.

Washington, DC:
adams, adams morgan, alexandria va, american, and downtown, apartments in alexandria, arlington, arlington va, bloomingdale, branch, brookland, capitol, capitol hill, cathedral, chase, chevy, chevy chase, chinatown, circle, circle and, cleveland, cleveland park, columbia, columbia heights, crystal, crystal city, downtown, downtown bethesda, downtown silver, downtown silver spring, dupont, dupont circle, foggy, foggy bottom, forest, fort, friendship, friendship heights, from downtown, george, georgetown, georgetown and, georgetown university, georgia, glover, glover park, green, heights, howard, in alexandria, in arlington, kalorama, logan, logan circle, morgan, navy, navy yard, noma, of old town, old town, old town alexandria, petworth, pleasant, potomac, shaw, silver spring, silver spring md, spring, spring md, stadium, takoma, takoma park, to downtown, to dupont, to dupont circle, to georgetown, to silver, to union, to union station, town alexandria, triangle, u corridor, union, union station, university, vernon

Seattle, WA:
admiral, alki, alki beach, and redmond, anne, ballard, ballard and, beacon, beacon hill, belltown, bothell, bothell wa, broadway, by windermere, capitol hill, columbia, columbia city, corridor, eastlake, first hill, fremont, green lake, greenlake, greenwood, heart of capitol, heart of downtown, interbay, international district, junction, lake city, lake union and, lincoln, lower queen, lower queen anne, madison, magnolia, mercer, ne seattle, ne seattle wa, north seattle, northgate, northgate mall, of ballard, of capitol, of capitol hill, of lake union, of queen, of queen anne, phinney, phinney ridge, pike, pike pine, pike place, pike place market, pine, pine corridor, pioneer, pioneer square, queen anne, ravenna, roosevelt, seattle center, seattle central, seattle downtown, seattle university, shoreline, south lake, south lake union, stevens, the junction, the university district, to green, to green lake, u district, union and, university district, university village, uptown, uw campus, wallingford, west seattle, westlake, windermere, woodland, woodland park

Montréal, QC:
and downtown, canal lachine, cote des, cote des neiges, dame, dame de, des neiges, downtown, downtown and, downtown montreal, du mont royal, du plateau, from downtown, griffintown, heart of downtown, henri, in downtown, in ndg, lachine, lasalle, laurent, le plateau, loyola, marie, mile end, minutes to downtown, monk, monkland, monkland village, mont royal, mont royal et, mount, mount royal, neiges, nord, notre, notre dame de, of downtown, of downtown montreal, of the plateau, old montreal, old port, outremont, plateau, plateau mont, plateau mont royal, rosemont, royal, saint, saint laurent, snowdon, st henri, te des neiges, the lachine, the plateau, to downtown, tro mont, tro mont royal, verdun, villa maria, village, ville, ville marie, villeray, westmount

References

1. Riesz, M. Borders Disputed! Brooklynites Take Issue with Google's Neighborhood Maps, 2014. Available online: https://www.brooklynpaper.com/stories/37/18/all-google-maps-neighborhoods-2014-04-25-bk_37_18.html (accessed on 1 July 2018).
2. Folven, E. Residents Voice Anger of Redistricting Maps, 2012. Available online: http://beverlypress.com/2012/02/residents-voice-anger-of-redistricting-maps/ (accessed on 1 July 2018).
3. Usborne, S. Disputed Territories: Where Google Maps Draws the Line, 2018. Available online: https://www.theguardian.com/technology/shortcuts/2016/aug/10/google-maps-disputed-territories-palestineishere (accessed on 1 July 2018).
4. Sutter, J. Google Maps Border Becomes Part of International Dispute, 2010. Available online: http://edition.cnn.com/2010/TECH/web/11/05/nicaragua.raid.google.maps/index.html (accessed on 1 July 2018).

5. Nicas, J. As Google Maps Renames Neighborhoods, Residents Fume, 2018. Available online: https://www.nytimes.com/2018/08/02/technology/google-maps-neighborhood-names.html (accessed on 1 July 2018).
6. Taylor, R.B.; Gottfredson, S.D.; Brower, S. Neighborhood naming as an index of attachment to place. *Popul. Environ.* **1984**, *7*, 103–125. [CrossRef]
7. Mitrany, M.; Mazumdar, S. Neighborhood design and religion: Modern Orthodox Jews. *J. Archit. Plan. Res.* **2009**, *26*, 44–69.
8. Knopp, L. Gentrification and gay neighborhood formation in New Orleans. In *Homo Economics: Capitalism, Community, and Lesbian and Gay Life*; Psychology Press: Hove, UK, 1997; pp. 45–59.
9. Alderman, D.H. A street fit for a King: Naming places and commemoration in the American South. *Prof. Geogr.* **2000**, *52*, 672–684. [CrossRef]
10. Hernandez, J. Redlining revisited: Mortgage lending patterns in Sacramento 1930–2004. *Int. J. Urban Reg. Res.* **2009**, *33*, 291–313. [CrossRef]
11. Northcraft, G.B.; Neale, M.A. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organ. Behav. Hum. Decis. Process.* **1987**, *39*, 84–97. [CrossRef]
12. Altschuler, A.; Somkin, C.P.; Adler, N.E. Local services and amenities, neighborhood social capital, and health. *Soc. Sci. Med.* **2004**, *59*, 1219–1229. [CrossRef] [PubMed]
13. Calem, P.S.; Gillen, K.; Wachter, S. The neighborhood distribution of subprime mortgage lending. *J. Real Estate Financ. Econ.* **2004**, *29*, 393–410. [CrossRef]
14. Romero, M. How Real Estate Websites Define Fishtown's Boundaries, 2016. Available online: https://philly.curbed.com/2016/10/31/13458206/fishtown-neighborhood-boundaries-map (accessed on 2 June 2018).
15. Grether, D.M.; Mieszkowski, P. Determinants of real estate values. *J. Urban Econ.* **1974**, *1*, 127–145. [CrossRef]
16. Mumford, L. The neighborhood and the neighborhood unit. *Town Plan. Rev.* **1954**, *24*, 256–270. [CrossRef]
17. Talen, E. Constructing neighborhoods from the bottom up: The case for resident-generated GIS. *Environ. Plan. B Plan. Des.* **1999**, *26*, 533–554. [CrossRef]
18. Sieber, R. Public participation geographic information systems: A literature review and framework. *Ann. Assoc. Am. Geogr.* **2006**, *96*, 491–507. [CrossRef]
19. United States Department of Housing and Urban Development; Office of Policy Development and Research. *The Behavioral Foundations of Neighborhood Change*; University of Michigan Library: Ann Arbor, MI, USA, 1979.
20. Keller, S.I. *The Urban Neighborhood: A Sociological Perspective*; Random House: New York, NY, USA, 1968; Volume 33.
21. Hoyt, H. *The Structure and Growth of Residential Neighborhoods in American Cities*; Washington, U.S. Govt.: Washington, DC, USA, 1939.
22. Coulton, C.J.; Korbin, J.; Chan, T.; Su, M. Mapping residents' perceptions of neighborhood boundaries: A methodological note. *Am. J. Community Psychol.* **2001**, *29*, 371–383. [CrossRef] [PubMed]
23. Lee, B.A.; Reardon, S.F.; Firebaugh, G.; Farrell, C.R.; Matthews, S.A.; O'Sullivan, D. Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *Am. Sociol. Rev.* **2008**, *73*, 766–791. [CrossRef] [PubMed]
24. Sampson, R.J.; Morenoff, J.D.; Gannon-Rowley, T. Assessing "neighborhood effects": Social processes and new directions in research. *Ann. Rev. Sociol.* **2002**, *28*, 443–478. [CrossRef]
25. Schockaert, S.; De Cock, M. Neighborhood restrictions in geographic IR. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, 23–27 July 2007; pp. 167–174.
26. Hollenstein, L.; Purves, R. Exploring place through user-generated content: Using Flickr tags to describe city cores. *J. Spat. Inf. Sci.* **2010**, *2010*, 21–48.
27. Hu, Y.; Gao, S.; Janowicz, K.; Yu, B.; Li, W.; Prasad, S. Extracting and understanding urban areas of interest using geotagged photos. *Comput. Environ. Urban Syst.* **2015**, *54*, 240–254. [CrossRef]
28. Gao, S.; Janowicz, K.; Montello, D.R.; Hu, Y.; Yang, J.A.; McKenzie, G.; Ju, Y.; Gong, L.; Adams, B.; Yan, B. A data-synthesis-driven method for detecting and extracting vague cognitive regions. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1245–1271. [CrossRef]
29. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]

30. McKenzie, G.; Adams, B. Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content. In *Leibniz International Proceedings in Informatics (LIPIcs), Proceedings of the 13th International Conference on Spatial Information Theory (COSIT 2017), L'Aquila, Italy, 4–8 September 2017*; Clementini, E., Donnelly, M., Yuan, M., Kray, C., Fogliaroni, P., Ballatore, A., Eds.; Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik: Dagstuhl, Germany, 2017; Volume 86, pp. 1–14, doi:10.4230/LIPIcs.COSIT.2017.20.
31. Cranshaw, J.; Schwartz, R.; Hong, J.I.; Sadeh, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 4–7 June 2012.
32. Wahl, B.; Wilde, E. Mapping the World. . . One Neighborhood at a Time. *Directions Magazine*, 4 December 2008.
33. McKenzie, G.; Hu, Y. The "Nearby" Exaggeration in Real Estate. In Proceedings of the Cognitive Scales of Spatial Information Workshop (CoSSI 2017), L'Aquila, Italy, 4–8 September 2017.
34. Chisholm, M.; Cohen, R. The Neighborhood Project, 2005. Available online: https://hood.theory.org/ (accessed on 2 June 2018).
35. Hu, Y.; Mao, H.; McKenzie, G. A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements. *Int. J. Geogr. Inf. Sci.* **2018**. [CrossRef]
36. Zhu, R.; Hu, Y.; Janowicz, K.; McKenzie, G. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Trans. GIS* **2016**, *20*, 333–355. [CrossRef]
37. Zhu, R.; Janowicz, K.; Yan, B.; Hu, Y. Which kobani? a case study on the role of spatial statistics and semantics for coreference resolution across gazetteers. In Proceedings of the International Conference on Geographic Information Science, Montreal, QC, Canada, 27–30 September 2016.
38. Brindley, P.; Goulding, J.; Wilson, M.L. A data driven approach to mapping urban neighbourhoods. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Fort Worth, TX, USA, 4–7 November 2014; pp. 437–440.
39. Brindley, P.; Goulding, J.; Wilson, M.L. Generating vague neighbourhoods through data mining of passive web data. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 498–523. [CrossRef]
40. Jones, C.B.; Purves, R.S.; Clough, P.D.; Joho, H. Modelling vague places with knowledge from the Web. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1045–1065. [CrossRef]
41. Derungs, C.; Purves, R.S. Mining nearness relations from an n-grams web corpus in geographical space. *Spat. Cogn. Comput.* **2016**, *16*, 301–322. [CrossRef]
42. Vasardani, M.; Winter, S.; Richter, K.F. Locating place names from place descriptions. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 2509–2532. [CrossRef]
43. Buscaldi, D.; Rosso, P. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 301–313. [CrossRef]
44. Gelernter, J.; Mushegian, N. Geo-parsing messages from microtext. *Trans. GIS* **2011**, *15*, 753–773. [CrossRef]
45. Inkpen, D.; Liu, J.; Farzindar, A.; Kazemi, F.; Ghazi, D. Location detection and disambiguation from Twitter messages. *J. Intell. Inf. Syst.* **2017**, *49*, 237–253. [CrossRef]
46. Liu, Y.; Wang, F.; Kang, C.; Gao, Y.; Lu, Y. Analyzing Relatedness by Toponym Co-O ccurrences on Web Pages. *Trans. GIS* **2014**, *18*, 89–107. [CrossRef]
47. Santos, J.; Anastácio, I.; Martins, B. Using machine learning methods for disambiguating place references in textual documents. *GeoJournal* **2015**, *80*, 375–392. [CrossRef]
48. Melo, F.; Martins, B. Automated geocoding of textual documents: A survey of current approaches. *Trans. GIS* **2017**, *21*, 3–38. [CrossRef]
49. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Newton, MA, USA, 2009.
50. Baddeley, A.; Rubak, E.; Turner, R. *Spatial Point Patterns: Methodology and Applications with R*; Chapman and Hall/CRC Press: London, UK, 2015.
51. Ripley, B.D. The second-order analysis of stationary point processes. *J. Appl. Probab.* **1976**, *13*, 255–266. [CrossRef]
52. Besag, J.E. Comment on 'Modelling spatial patterns' by BD Ripley. *JR Stat. Soc. B* **1977**, *39*, 193–195.
53. McKenzie, G.; Janowicz, K.; Gao, S.; Yang, J.A.; Hu, Y. POI pulse: A multi-granular, semantic signature–based information observatory for the interactive visualization of big geosocial data. *Cartographica* **2015**, *50*, 71–85. [CrossRef]

54. Graham, R.L. An efficient algorithm for determining the convex hull of a finite planar set. *Inf. Process. Lett.* **1972**, *1*, 132–133. [CrossRef]
55. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
56. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
57. Chesnokova, O.; Nowak, M.; Purves, R.S. A crowdsourced model of landscape preference. In *LIPIcs-Leibniz International Proceedings in Informatics*; Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik: Wadern, Germany, 2017; Volume 86.
58. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manag.* **2012**, *275*, 117–129. [CrossRef]
59. Hayes, M.M.; Miller, S.N.; Murphy, M.A. High-resolution landcover classification using Random Forest. *Remote Sens. Lett.* **2014**, *5*, 112–121. [CrossRef]
60. George-Cosh, D. July 1 Is Day for Mass, Messy Moves in Montreal., 2013 Available online: https://www.wsj.com/articles/SB10001424127887323300004578559722182821246 (accessed on 2 June 2018).
61. Boeing, G.; Waddell, P. New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings. *J. Plan. Educ. Res.* **2017**, *37*, 457–476. [CrossRef]