RESEARCH ARTICLE

Using OpenStreetMap point-of-interest data to model urban change—A feasibility study

Liming Zhang[☯]*, **Dieter Pfoser**[☯]

Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA, United States of America

☯ These authors contributed equally to this work.
* lzhang22@gmu.edu

Abstract

User-generated content is a valuable resource for capturing all aspects of our environment and lives, and dedicated Volunteered Geographic Information (VGI) efforts such as OpenStreetMap (OSM) have revolutionized spatial data collection. While OSM data is widely used, considerably little attention has been paid to the quality of its Point-of-interest (POI) component. This work studies the accuracy, coverage, and trend worthiness of POI data. We assess the accuracy and coverage using another VGI source that utilizes editorial control. OSM data is compared to Foursquare data by using a combination of label similarity and positional proximity. Using the example of coffee shop POIs in Manhattan we also assess the trend worthiness of OSM data. A series of spatio-temporal statistical models are tested to compare change in the number of coffee shops to home prices in certain areas. This work overall shows that, although not perfect, OSM POI data and specifically its temporal aspect (changeset) can be used to drive urban science research and to study urban change.

OPEN ACCESS

Citation: Zhang L, Pfoser D (2019) Using OpenStreetMap point-of-interest data to model urban change—A feasibility study. PLoS ONE 14 (2): e0212606. https://doi.org/10.1371/journal.pone.0212606

Editor: Michael Szell, IT University of Copenhagen, DENMARK

Received: October 23, 2017

Accepted: February 6, 2019

Published: February 25, 2019

Data Availability Statement: Raw data of changelog can be downloaded from OpenStreetMap.org at http://wiki.openstreetmap.org/wiki/Planet.osm/full Home price data can be downloaded from Zillow.com at https://www.zillow.com/home-values/ Foursquare data can be crawled from API: https://developer.foursquare.com/docs/api/venues/search.

Funding: This work was supported by Department of Defense grant HM02101410004; the National Science Foundation, Grant No. 1637541; and George Mason University, Presidential graduate

1 Introduction

The inexorable trend towards urbanization worldwide presents a pressing challenge to our understanding of the scale and speed of urban processes. *Scale* refers to the quantity of data involved given the changing spatial and temporal resolution of the data. Traditionally, urban theories examine a coarse level such as a whole city observed over decades. Nowadays we try to understand the urban processes at the building and the citizen level. This trend creates a considerable data science research challenge. Complementing scale, *speed* refers to the analysis of data in real-time, which should lead to faster and more responsive decision making. These aspects give rise to the emerging data science field called *Urban Analytics*. As part of Urban Analytics we use novel types of data to evaluate contemporary and future cities through methods including GIS, Remote Sensing, Big Data and Geodemographics [1]. In studying *urban change*, we try to assess and verify urban theories using data science methods. Urban change relates to the physical environment [2], including land use [3], infrastructure, business locations [4] and other assets of a city. Such an effort necessitates basic data collection, which until

research scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

recently has been the responsibility of public authorities. This creates a bottleneck with respect to scale and speed, since such efforts rely on considerable man power and funding support. Updates to the data follow a 5+ years collection cycle. Studying urban change at the levels of granularity outlined above, requires data of a higher spatial and temporal resolution. Depending on existing collection infrastructure and regulations, such data might not be available and/ or trustworthy [5].

Our work tries to assess the suitability of user-generated content in the form of OpenStreetMap POI data as a means to infer urban change. Specifically, we will explore two aspects of the data: (i) accuracy and coverage and (ii) trend worthiness. The *accuracy and coverage* of OSM POI data is not always evident and has been the focus of a sizeable research community over the years (cf. Section 2). In this work, we compare OSM POI data to Foursquare data. Foursquare data although a crowdsourced resource, exhibits some editorial control. Our comparison will use the name and location of POIs in both sources to reason about the respective coverage. Besides positional accuracy, a question we would like to answer is whether VGI is updated in a timely manner and can be used to capture trends and, in our specific context, urban change. To satisfy this so-called *trend worthiness* criterion, we assess whether VGI can match some commonly recognized trends or theories. Often urban analytics study the interaction of an environment to social phenomena. Our approach here is that by using statistical modeling, we consider VGI as trend worthy if the *modeling error is acceptable while maintaining its statistical behavior*. Such a test is useful in a social science context, since it would conditionally permit the use of VGI and, more generally, user-generated content in an otherwise data-poor environment. To this effect, we create statistical models based on population-based power law relationships in urban science [6]. This underlying theory suggests that population growth is one of the fundamental parameters behind change to an urban environment. For example, some basic power-law functions can relate population size to urban phenomena such as electricity usage, salaries, road network length, intellectual output, and even the citizens' average speed of walking and heart rate. With evaluate a range of models in terms of their quality and how well they allow us to use OSM POI as an indicator of urban change.

The remainder of this work is organized as follows. Section 2 discusses related work such as Power Law Relationships and how they capture urban phenomena as well as quality aspects and fitness-for-use of user-generated content. Section 3 describes the data sources, related collection methods, and provides some data visualizations. Section 4 outlines our methodology to assess data quality. Section 5 presents the results of the quality assessment, including error distribution, the performance of statistical modeling, and interesting observations such as "inverse coffee shop effects". Section 6 concludes and gives directions for future research.

2 Related work

Our discussion of related work focuses on two aspects, (i) user-generated content and its quality and fitness-for-use issues, and (ii) urban science theory as related to our work.

2.1 User-generated content and quality

Given the advent of volunteered geographic information (VGI) [7] and geospatial crowdsourcing [8], more and more relevant user-generated content becomes available. The most well-known geo-crowdsourcing effort is OpenStreetMap (OSM) (http://www.openstreetmap.org), colloquially referred to as the Wikipedia of Maps. OSM is a "free" vector dataset covering the entire planet. It has started out as a road network dataset, but by now includes general spatial feature information (transportation networks, buildings, land use data) and, important for this effort, point-of-interest data. A concern with VGI and user-generated content is data quality.

Given the lack of quality control in data collection, the error in the data could be wild. OSM coverage and accuracy are for example examined in [9–12]. The work in [9] examines the accuracy and coverage of OSM by comparing it to British Ordnance Survey datasets. In this case, the error of OSM was found to be within 6*m* and the data had a 26% coverage. It should be noted that this study was conducted almost a decade ago. The authors of [11] assessed spatial coverage and ground-truth positional accuracy for five cities and towns in Ireland by comparing OSM data to Google and Bing maps. No method is proposed for POI data. A more systematic quality assessment is conducted in [12], which proposes metrics for geometric, attribute, semantic and temporal accuracy, as well as logical consistency, completeness, lineage, and usage. Most work however has focused on road networks and considerably little attention has been paid to POI data.

A recent OSM use cases survey [13] mentions an increasing number of data mining efforts in support of urban analytics. Urban form and function is but one area. [14] discusses different aspects of urban form and function and how they can be captured from user-generated content. Examples include map construction algorithms [15–17] and social media mining methods [18]. [5, 19] developed algorithms to better infer land use from OSM. Some other studies evaluated the overall potential for using OSM to assess land use and land cover [20, 21]. Another study in [22] presents an Urban-Rural Index derived from OSM to create an objective understanding of rural and urban classification. To quantify urban change, business directories and geocoding have been used in various efforts [23–26].

2.2 OSM POI quality assessment

Focusing on city scale, various authors have examined POI data quality. In [27], the authors propose a *spatial-semantic interaction* methodology to analyze the internal reference of different types of features of OSM POIs. The method developed so-called variograms and clusters of different feature types. The similarity of two POI features is defined as the spatiotemporal co-occurrence of different feature types of the same POI. A spatial process statistic is calculated to see if the processes are independent or not. Similar to our challenge, the authors mention that a weakness in their approach is the absence of a reference dataset. Recently, fitness-for-use of POIs [28] is defined at the levels of geo referencing, i.e., is the location accurate and allows for inferring an unambiguous reference to a real-world entity. Unfortunately, the authors do not propose a systematic metric for their measure. An overview paper [29] compares different aspects of the quality of OSM data, such as coverage, accuracy, and historical change using methods developed by different researchers. This work includes a reference dataset, the BD TOPO database produced by IGN, and it compares the data to a dataset derived from Flickr. The ambition of this quality assessment is limited to accuracy and coverage. To the best of our knowledge, our work is the first study that assesses the "fitness-for-use" of OSM POI data for studying urban phenomena and change. While we also consider accuracy and coverage, our most important contribution relates to the examination of *trend worthiness* of OSM POI data as a means to assess urban change. We consider a POI dataset trend worthy if the relative rate of change in the number of POIs accurately reflects a change in the real world, while conceding that the number of POIs at any given time deviates from the actual number of POIs existing in the real world, i.e., actual coffee shops vs. coffee shop POIs recorded in OSM.

2.3 Power law relationships

The fundamental dynamics of cities can be related to power law scaling relationships [6], i.e., urban phenomena scale with population size based on an universal power law function $P = \alpha N^{\beta}$, in which P is a specific kind of urban indicator, N is population size, α and β is scaling parameters. Based on the value of β, we can develop several growth models, which introduce different rates of growth for different cities. The authors here draw a comparison between the size of cities and the size of life forms and the energy demand each has to be sustainable. This work lead to countless results utilizing this theory, including the shape of cities [30], mobility patterns [31], and urban metabolism theory [32].

Other works have examined coffee shops and their deep connection to our social life. Coffee shops are social capital [33], in that a good coffee shop can provide a stronger sense of community for the residents of a neighborhood. One can consider places such as coffee shops, plazas, market places etc. as "the heart of a community's social vitality and the grassroots of democracy" [34]. They represent a "Third Place" and are considered an essential part of society besides the workplace and home. With this social view of urban spaces, [35] provides a detailed review of how computing and analytics are augmenting people's experiences of cities. The author argues that urban sensing and analytics will lead to considerable urbanization improvements. Some quantitative studies have empirically evaluated these visions. The authors of [36] investigate the effect of coffee shops on crime. The authors argue that coffee shops provide "an on-the-ground and visible manifestation of a particular form of gentrification. . . and lifestyle."

With this theoretical background, we are confident in the use of power law relationships when using the change of POI data to study urban phenomena. *We would like to point out that this work should not be considered the one and only approach to studying urban change, but rather proposes a method that leverages OSM POI data in this context.*

3 Data

Various datasets are used to provide for a sound quality assessment of OSM POI data. We assess (i) the accuracy and coverage of OSM as a POI data source by using/comparing it to respective Foursquare data, and we assess (ii) its trend worthiness by building statistical models that relate the change in coffee shops to a change in home prices. The following sections discuss each data source as well as the pre-processing steps needed to make the data actionable.

3.1 OSM coffee shop data

OSM allows anybody to freely edit a global map dataset. To keep track of the edits, OSM uses an independent data object called "changeset" to record changes from editing operations and users. Changesets record all changes such as tags, coordinates, and comments. Such a database then includes for all OSM objects (nodes, ways, relations) respective metadata information as shown in Fig 1).

For our work, we are interested in changeset data that reflects actual real-world change, i.e., the addition or deletion of a coffee shop node in close temporal proximity to the actual opening or closing of the coffee shop. Since we do not have ground truth data, i.e., municipal records, we use quantitative information such as the monthly coffee shop count plot in Fig 2, which shows the overall change in OSM data.

Assuming OSM has matured as a dataset, the plot shows a slow growth in numbers after OSM's inception, as not too many people were aware of it. Following 2010, the rate of growth in edits increases and after 2015 the growth slows again, suggesting that the POIs recorded in OSM start reflecting real world changes, i.e., coffee shops that opened were recorded in a timely manner. Based on these trends, we select coffee shop changesets for the period of 11/2014 to 11/2016 for our experimentation and study.

In a first step, we compute a seasonal average. Here, November, December, and January are defined as Winter, and so forth. Although not strictly correct, this definition has been used in

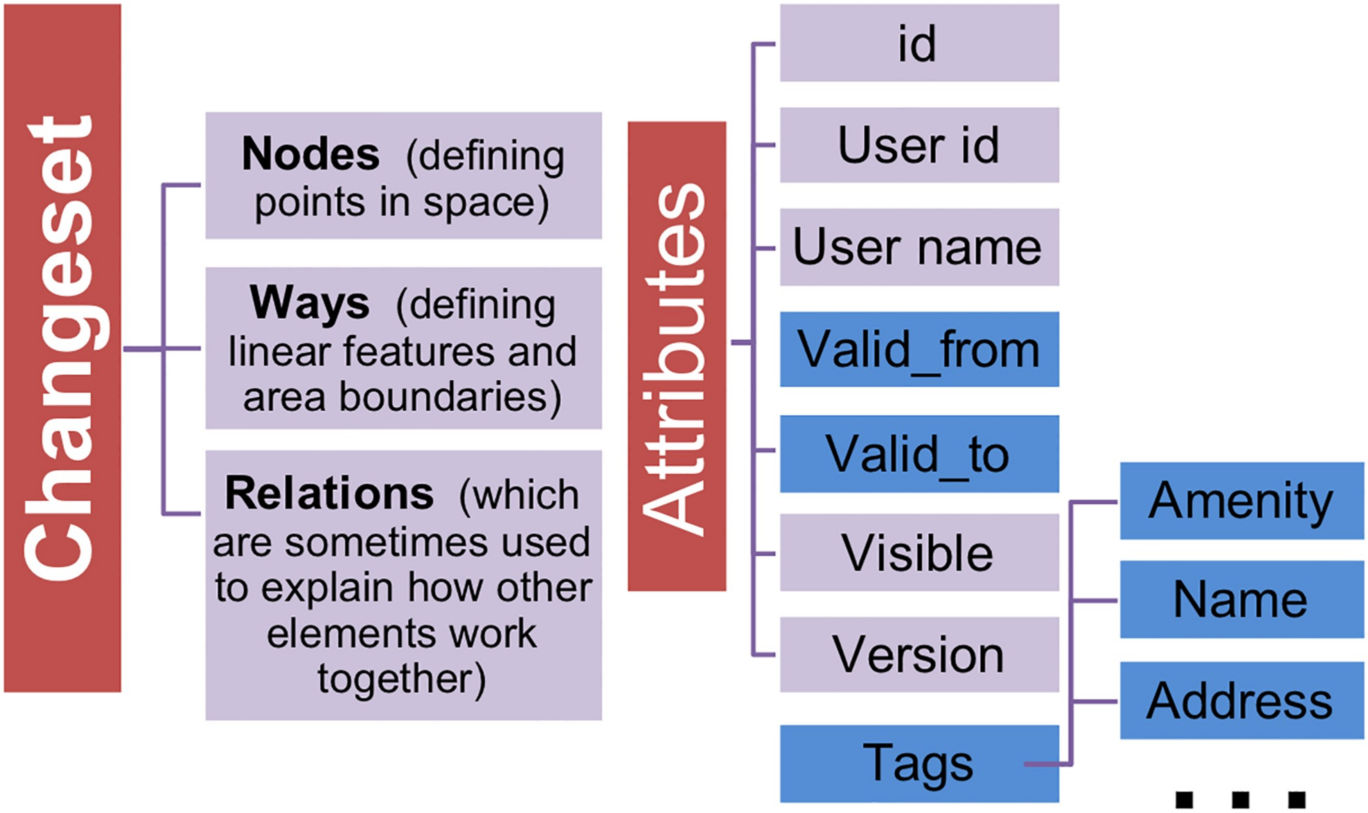
Changeset
Nodes (defining points in space)
Ways (defining linear features and area boundaries)
Relations (which are sometimes used to explain how other elements work together)
Attributes
id
User id
User name
Valid_from
Valid_to
Visible
Version
Tags
Amenity
Name
Address
. . .

Fig 1. OSM's changeset data object model.

https://doi.org/10.1371/journal.pone.0212606.g001

[37] reporting on real estate pricing trends. We can define the seasonal coffee shop density as the number of coffee shops in a neighborhood divided by the area of the neighborhood (km^2). We will use the term "Coffee Shop Density" in the remainder of this work. The changes to coffee shop density for each season are shown in Fig 3. We see that different neighborhoods seem to have different trends. While in most cases the coffee shop density is increasing, some neighborhoods, such as the Financial District and Soho, seem to suffer through periods of decreasing numbers. To better show different patterns between pairs of neighborhoods, a pairwise Pearson correlation (r score) [38] is shown in Fig 4. About half of the neighborhood pairs are strongly correlated (Pearson's $r > 0.5$). Only the Financial District is negatively correlated with most of the other neighborhoods. This mix of trends could also be an indication for the reliability of OSM data, i.e., the closing of shops is actually reflected in the crowdsourced OSM dataset.

3.2 Foursquare coffee shop data

We use Foursquare data [39] as a means to verify the coverage of OSM POI data. Foursquare, a location-based social-media app, utilizes POIs as the location where the user "socializes" with friends. Users are encouraged to review and visit different locations as often as possible to claim them. Since POIs are a core data aspect of this app, it also exerts editorial control, i.e., POIs are actively curated, As such, we can consider this dataset a good reference dataset when it comes to evaluating crowd-sourced OSM data.

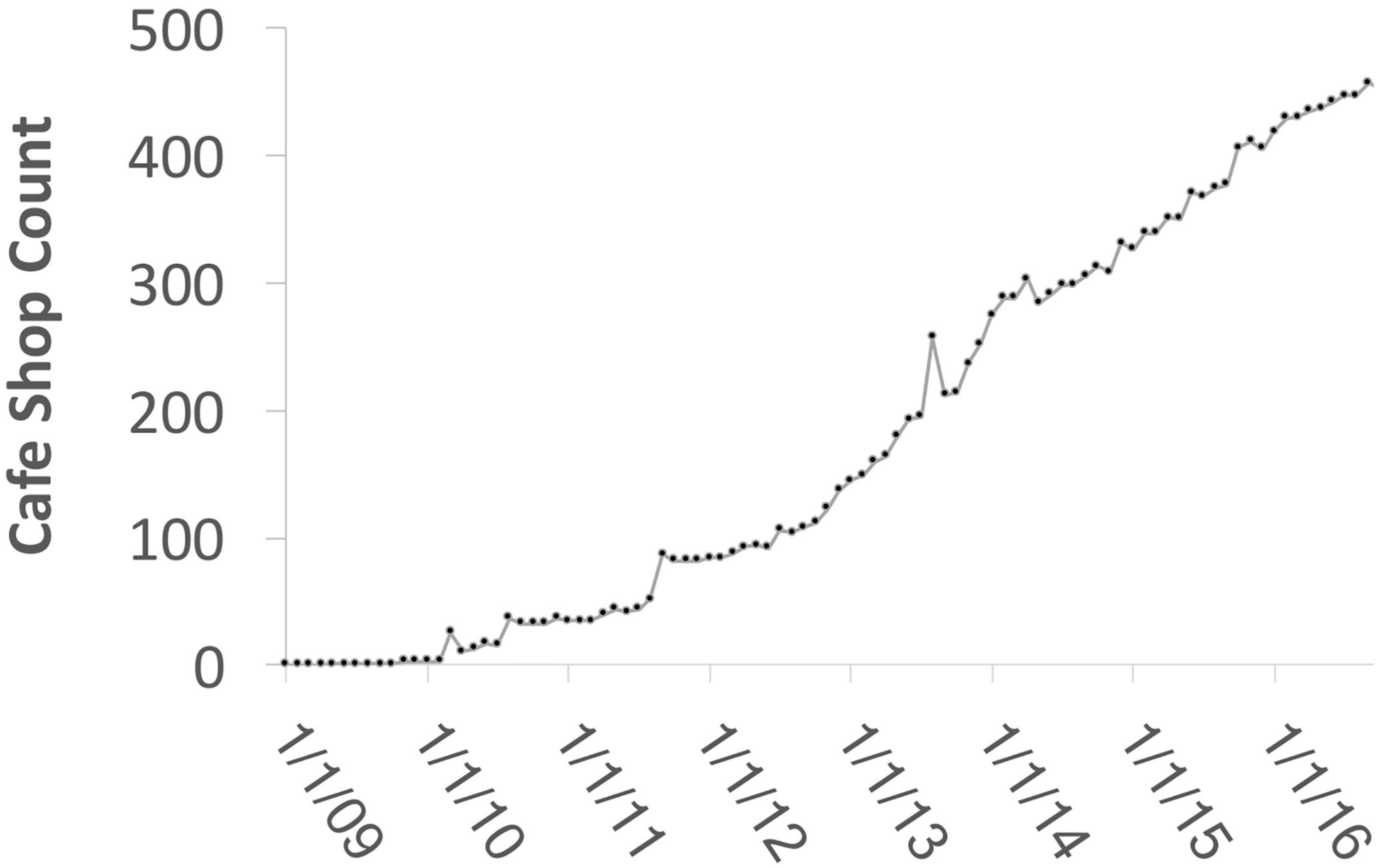

Fig 2. Coffee shop count in Manhattan from OSM—Monthly 1/1/2009–11/31/2016.

https://doi.org/10.1371/journal.pone.0212606.g002

Unfortunately, Foursquare does not provide access to their entire database in a fashion similar to OSM. An API with limited service rates allows one to interact with the service and in our case to retrieve POI information. To account for some API limitations, we use two types of queries as detailed in Tables A and Table B in S1 Appendix. Type I uses the collected OSM data to retrieve all Foursquare POIs that have a matching label within a 50*m* radius. Type II uses a regular spatial grid (200*m* spacing) to retrieve all coffee shop POIs in relation to the centroid of each cell. This strategy ensures that less than 50 POIs are retrieved per request (Foursquare limit), while it also covers POIs that are not in the immediate area of our OSM POIs. The mapping of categories of OSM data and Foursquare data is shown in Table C in S1 Appendix. The resulting Foursquare POI dataset is obtained by fusing these two datasets. The location of all OSM POIs for query Type I and grid center points for Type II are shown in Figs 5 and 6.

In total, 851 Foursquare POIs were retrieved on June 30, 2018. Section 4 will show how Foursquare and OSM POIs match up. An interesting observation is that Foursquare has a very different definition of POI categories when compared to OSM. For example, while Foursquare has a "donut" category, OSM considers it "cafe" (Table C in S1 Appendix).

Ideally our ground-truth data should have a historical dimension. Unfortunately, neither Foursquare nor any other data sources besides OSM captures this aspect. As such, we are only able to compare the accuracy of the current POIs between sources.

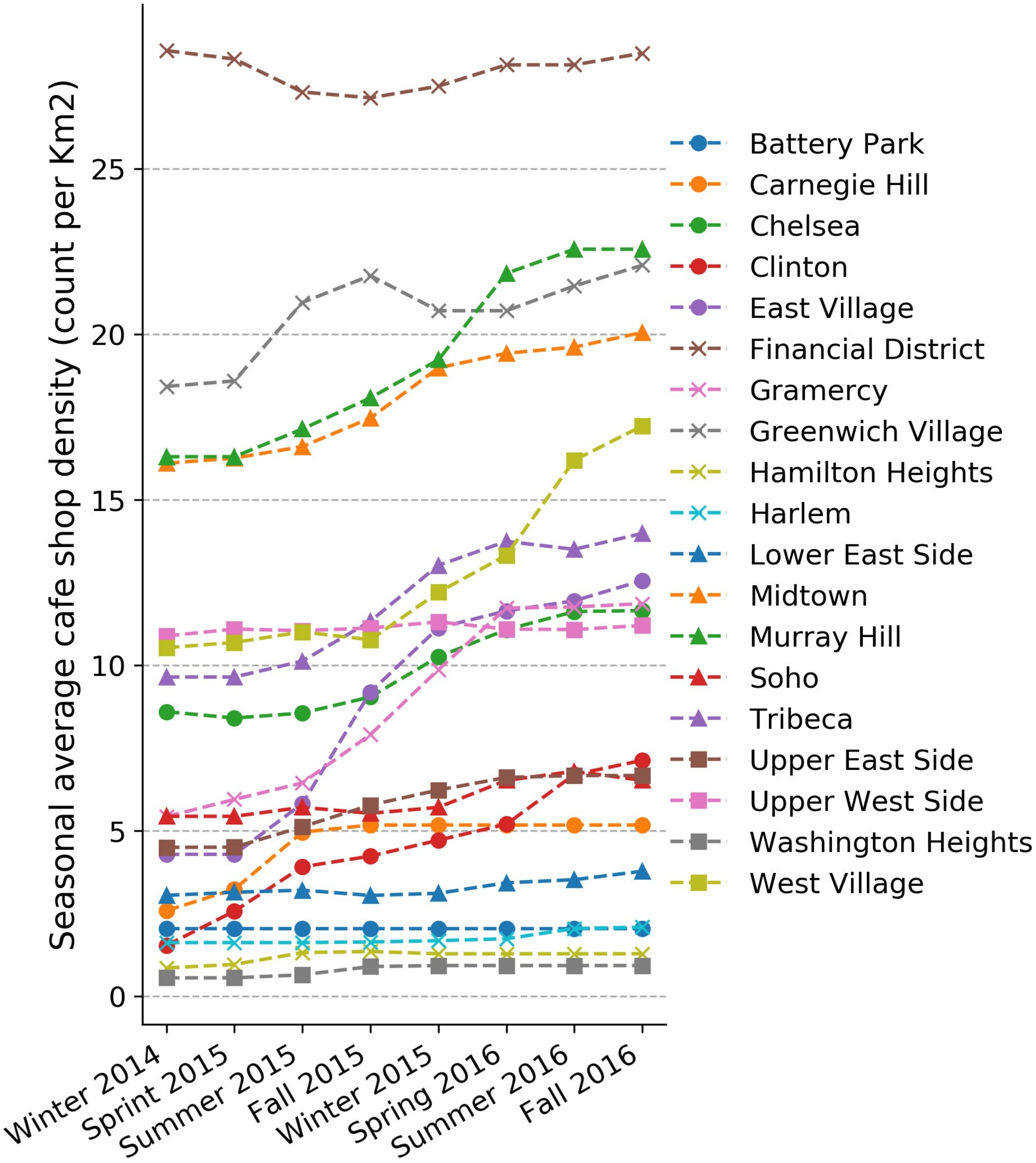

Fig 3. Coffee shop density [km^2]—Seasonal average 03/2015–11/2016.

https://doi.org/10.1371/journal.pone.0212606.g003

3.3 Home prices

To model the relationship between coffee shop densities and home prices, we obtain a dataset from Zillow, a real estate listing Web site. Zillow provides a data analysis product called "Home Value Index" (https://www.zillow.com/research/zhvi-methodology-6032/), which captures home prices at different spatial granularities ranging from city to neighborhood levels. This data is based on listing prices posted on the web site, and as such, can also be considered

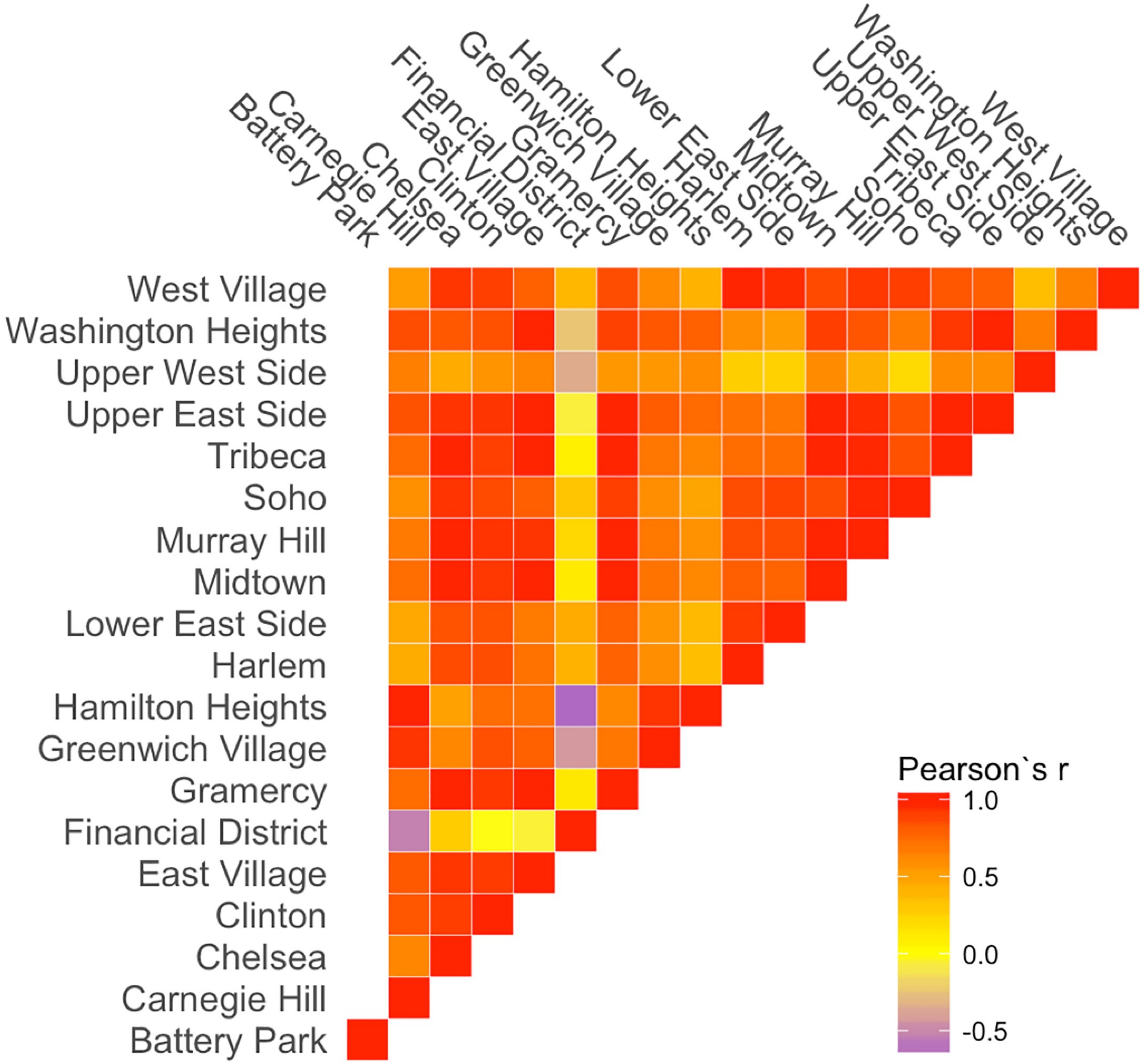

Fig 4. Pairwise Pearson correlation of neighborhoods' coffee shop densities.

https://doi.org/10.1371/journal.pone.0212606.g004

crowdsourced. For Manhattan, different datasets are provided at the neighborhood level. We selected the "Median List Price Per Sq Foot" from the Home Value Index. This data is more resilient with respect to outliers and removes a house size bias. We calculate the seasonal average of "Median List Price Per Sq Foot", to which we refer to as home prices in the remainder of the paper. The fluctuations of this measure are shown in Fig 7. Seeing this data spatially, we observe that adjacent neighborhoods tend to have similar home prices. Fig 8 visualizes this relationship. Neighborhoods for which no data is available from Zillow or if they have no real estate market (Central Park) are left blank (white).

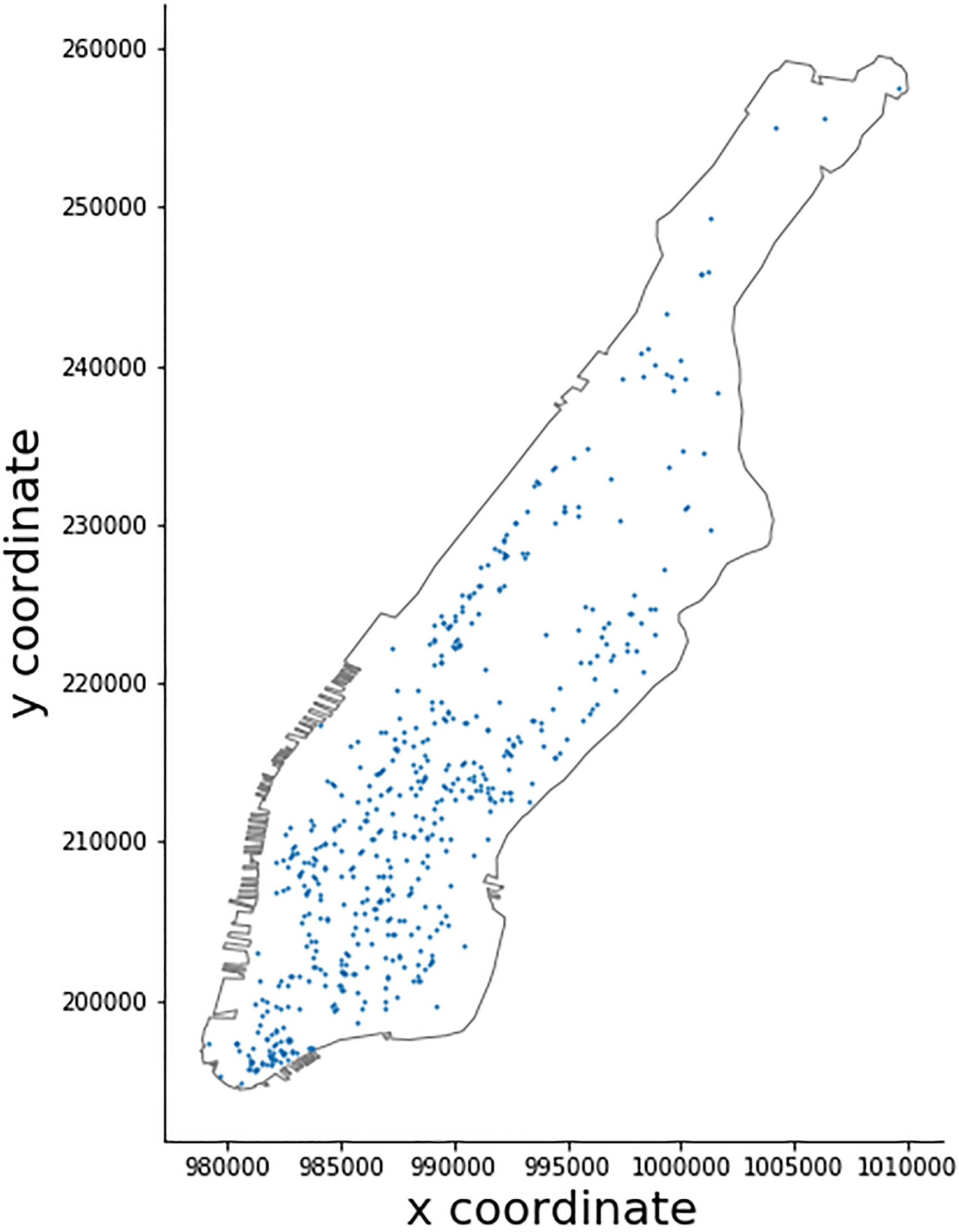

Fig 5. OSM coffee shop locations.

https://doi.org/10.1371/journal.pone.0212606.g005

4 Methodology

Our main objective is to assess the quality of user generated content and argue for its use in urban science research. The following sections provide a detailed discussion of the overall methodology that is employed. Fig 9 provides and overview.

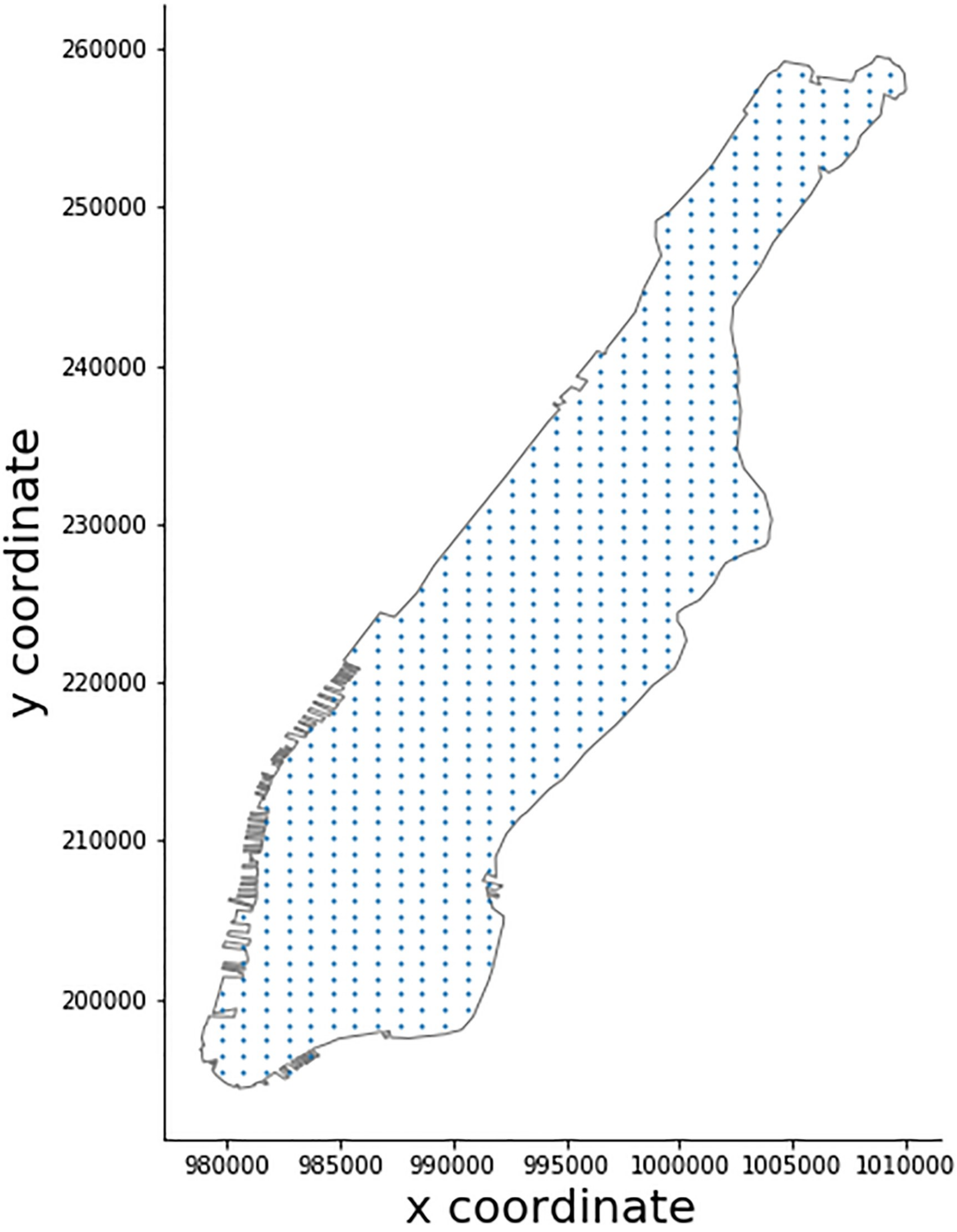

Fig 6. API search grid points.

https://doi.org/10.1371/journal.pone.0212606.g006

4.1 POI accuracy and coverage assessment

To assess the accuracy and coverage of OSM POI data, we compare it to a reference dataset. However, since there are no authoritative datasets available, we chose Foursquare POI data, which is crowdsourcing data with some editorial control. Unfortunately, no historical versions

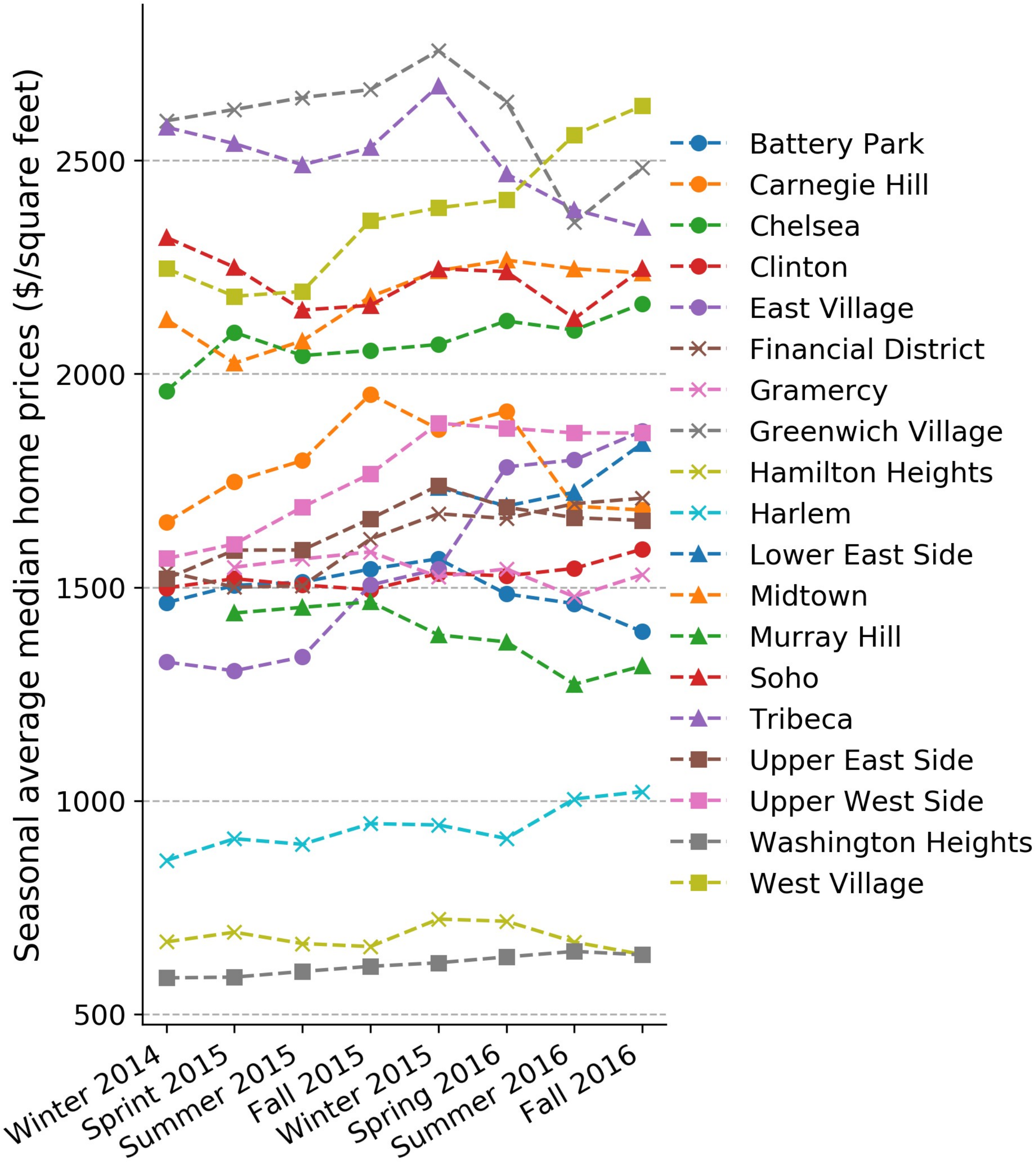

Fig 7. Unit home price (UHP) ($/$ft^2$) changes—Seasonal average 12/2014–11/2016.

https://doi.org/10.1371/journal.pone.0212606.g007

of the data are available. Our dataset was retrieved on June 30, 2018. Comparing two POI sources that cover the same geographic area (Manhattan) and context (coffee shops) is not trivial, given a potential label mismatch ("Ben's Cafe and Eatery" vs. "Ben's") and location uncertainty (a daunting problem for any user-generated content). In the following, we try to match POIs from both data sets using string similarity measures and location proximity.

4.1.1 Label similarity. To compare POI labels between data sources, we selected the Longest Common Sub-sequence (LCS) method and the Levenshtein distance (cf. [40]). The

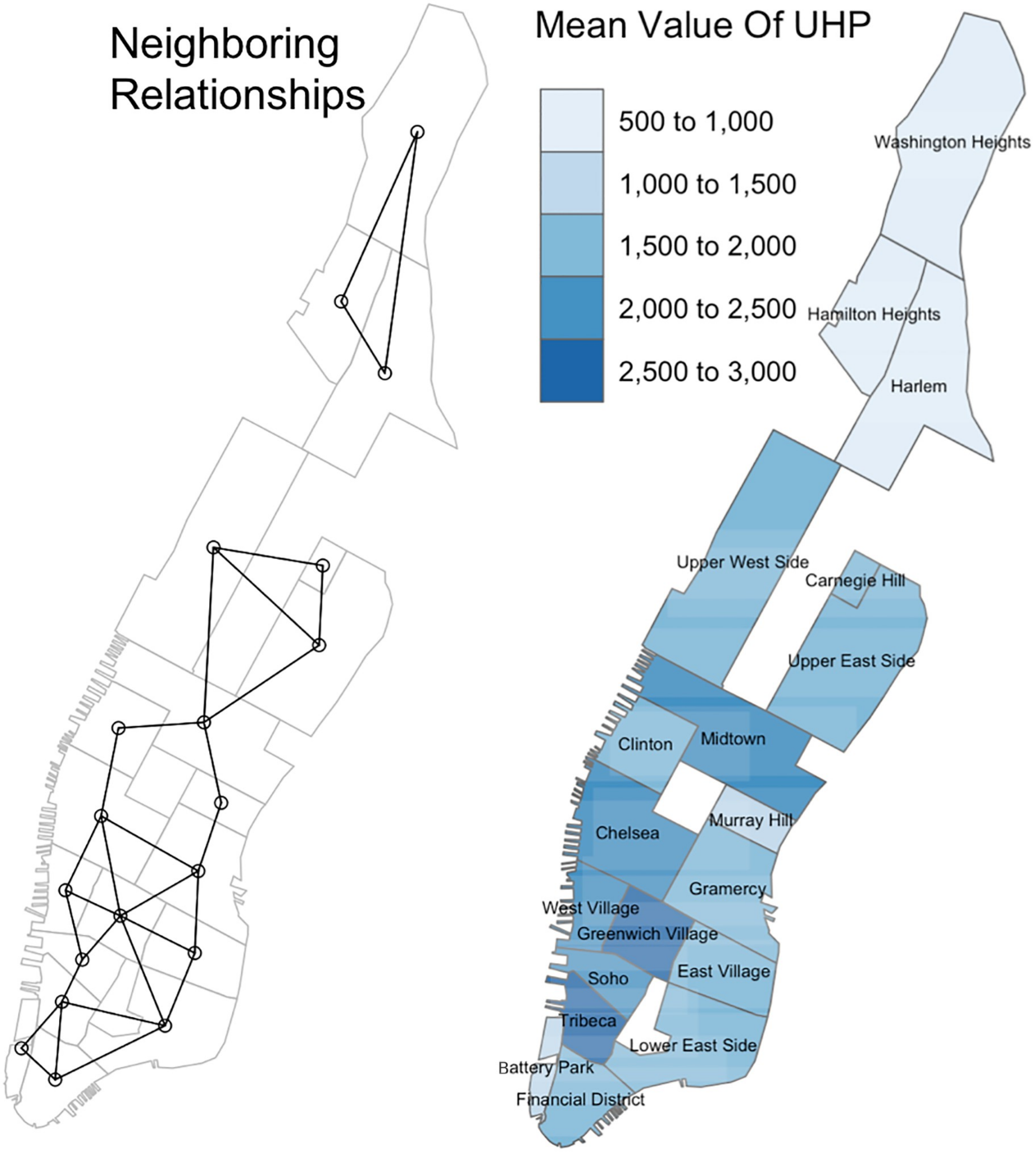

Fig 8. Relationships and mean values: (left) adjacency relationship map of neighborhoods (dots represent neighborhoods, edges indicate adjacent neighborhoods), (right) mean values of average unit house price (UHP).

https://doi.org/10.1371/journal.pone.0212606.g008

Levenshtein distance has been used in similar contexts, e.g., [29]. LCS gets its advantage over Levenshtein distance since it measures the difference of two strings that are compared to a common substring, while the Levenshtein distance uses one string as its reference. Given two crowdsourced data sources, it is unlikely that two labels for the same POI match up exactly. Both methods can provide an approximate match and can account for labels of varying length and spelling differences. Both measures are distance measures, which need to be converted to a similarity score. Similarity scores can be mapped to the interval [0, 1]. To formally define the

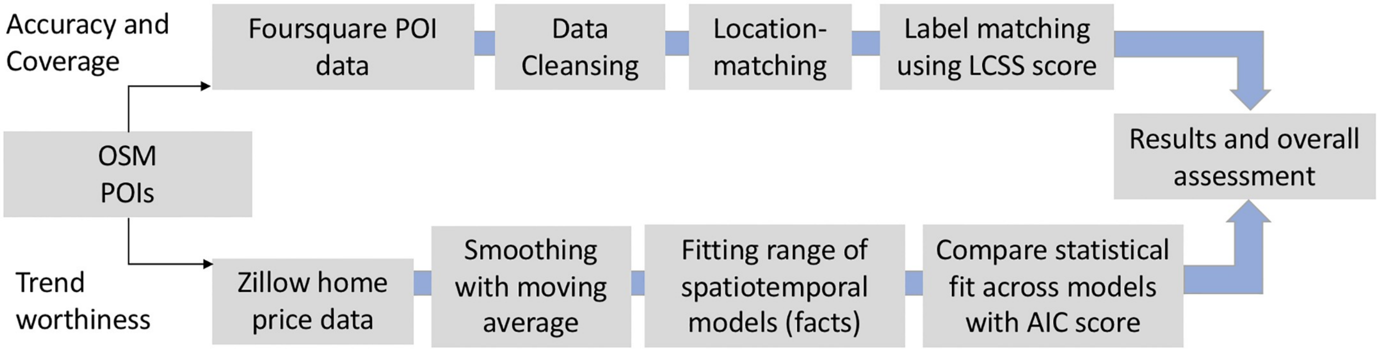
Accuracy and Coverage
Foursquare POI data
Data Cleansing
Location-matching
Label matching using LCSS score
OSM POIs
Results and overall assessment
Trend worthiness
Zillow home price data
Smoothing with moving average
Fitting range of spatiotemporal models (facts)
Compare statistical fit across models with AIC score

Fig 9. Quality assessment workflow overview.

https://doi.org/10.1371/journal.pone.0212606.g009

similarity of two labels, the two *Label Similarity Scores* are defined as follows.

$$S_{LCS} = \frac{L_{LCS} * 2}{L_{FSQ} + L_{OSM}} \tag{1}$$

$$S_{LD} = \frac{L_{OSM} - L_{LD}}{L_{OSM}} \tag{2}$$

Here, L_{LCS} is the length of longest sub-sequence and L_{LD} is the number of changes needed to transform one label to another (reference). L_{FSQ} and L_{OSM} capture the lengths of the Foursquare and OSM labels, respectively. Identical labels generate a score of 1 in both cases. Both functions are monotonically increasing as the length of a common sub-sequence increases (LCS), or the difference between strings decreases (LD), which makes both a valid and good similarity measure.

When calculating similarity, we have to consider some particularities of the data. In both datasets, common terms are "coffee", "cafe", and "cafeteria", all referring to the same concept. Given the same POI, it might be labeled as "Starbucks" in one dataset and "Starbucks Coffee" in the other. In this case, its similarity scores would only be $S_{LCS} = \frac{9*2}{9+25} = 0.529, S_{LD} = \frac{25-9}{25} = 0.64$. As such we removed all commonly used terms from the labels to provide for a fair assessment (stop words). In addition, we also removed all white spaces and punctuation. The experiments use a rather strict similarity threshold of 0.9 for both methods (cf. Section 5).

4.1.2 Location similarity. POIs are typically captured as point locations and the expectation is that the recorded coordinates do not match up exactly across data sources. To examine whether two coordinates capture the same POI (and without considering the label) one can use a buffer region to see whether one location is close to the other. The choice of a proper threshold is critical, as various coffee shops with the same name (chain) might be close by. Using a projected coordinate system, Euclidean distance can be used. Fig 9 gives an overview of the overall processing pipeline.

4.2 Trend worthiness of OSM POI data

While the expectations are that OSM data is timely and has considerable coverage, it might not always perfectly match the real-world situation, i.e., accuracy and coverage might not always reflect reality. As such we want to assess whether such data captures overall urban trends and change using a statistical modeling approach.

We first introduce the concept of scaling relationships to model urban phenomena based on population size. This allows us then to relate change in coffee shop numbers and home prices to population and to eventually model the direct relationship between them. We use a range of spatial and temporal analysis methods and adjustments to try and improve the overall model fit. Section 5 will finally tell us the adjustments that work best and consequently the model that has the best fit.

4.2.1 Scaling relationship based on population. The fundamental driving force behind urban change is human activity and population size. In [6], it is shown that larger metropolitan areas produce comparatively more wealth, innovation and activity following a power law function based on population size.

This power law scaling relationship is shown in Eq 3. The variance of each urban indicator is the exponential scaling factor β, which can be used to derive three types of urban phenomena. With (i) $\beta \approx 1$ (linear growth) we can describe individual human needs, like jobs and water consumption, (ii) $\beta < 1$ (sublinear growth) characterizes infrastructure, and (iii) $\beta > 1$ (superlinear growth) signifies quantities related to social currencies, like innovation. The coefficient estimation is done using linear regression after a log transformation of the scaling relationship model (Eq 4) [6]. In the following equations, N_t is the population at a certain time, Y_0 is the initial state of an indicator, Y_t is the current state of an indicator, and ϵ_t is the error.

$$Y_t = Y_0 N_t^{\beta} \tag{3}$$

$$\log(Y_t) = \log(Y_0) + \beta \log(N_t) + \epsilon_t \tag{4}$$

Coming back to our problem of coffee shops and home prices, in [36] the authors found that coffee shops are a good indicator for gentrification. Gentrification is a common and controversial topic in politics and urban planning and refers to improving deteriorated urban neighborhoods by an influx of a wealthier demographic. Intuitively, the establishment of a coffee shop relies on certain population numbers (customers) to support it. To open a coffee shop, a much smaller investment is needed than, for example, a super market. As such, *the number of coffee shops is sensitive to relatively small changes in population numbers (volatility)*. It is easier to close a coffee shop than a supermarket when customers stay away. Another argument here is that coffee is a cheap commodity that is consumed frequently. Hence, it is not as susceptible to personal spending cuts as more expensive products such as entire meals, clothing or jewelry.

Power low relationship between coffee shops and home prices. Coffee shop numbers correlate with population numbers over time and place and can be treated as an indicator of human activity. Thus, they should follow a power law function of population. On the other hand, the real estate market is an economic phenomenon also related to human activity. With Eqs 3 and 4, we have two relations between coffee shops c and population, and between home prices p and population: $\log(c_t) = log(c_0) + \beta_c \log(N_t)$ and $\log(p_t) = log(p_0) + \beta_p \log(N_t)$. In combining them, we infer a function between coffee shops and home prices. Eq 5 also represents a power low scaling relationship and can be fitted using regression techniques. As a side benefit, since both datasets are user generated, it would also establish the usefulness of such data for the investigation of urban phenomena.

$$\begin{aligned} \log(P_t) &= \tau + \beta \log(C_t) \\ \tau &= log(P_0) - \ \beta_P/\beta_C \ log(C_0) \\ \beta &= \beta_P/\beta_C \end{aligned} \tag{5}$$

In Eq 5, $\log(P_t)$ is the log transformed home price and $\log(C_t)$ is the log transformed coffee

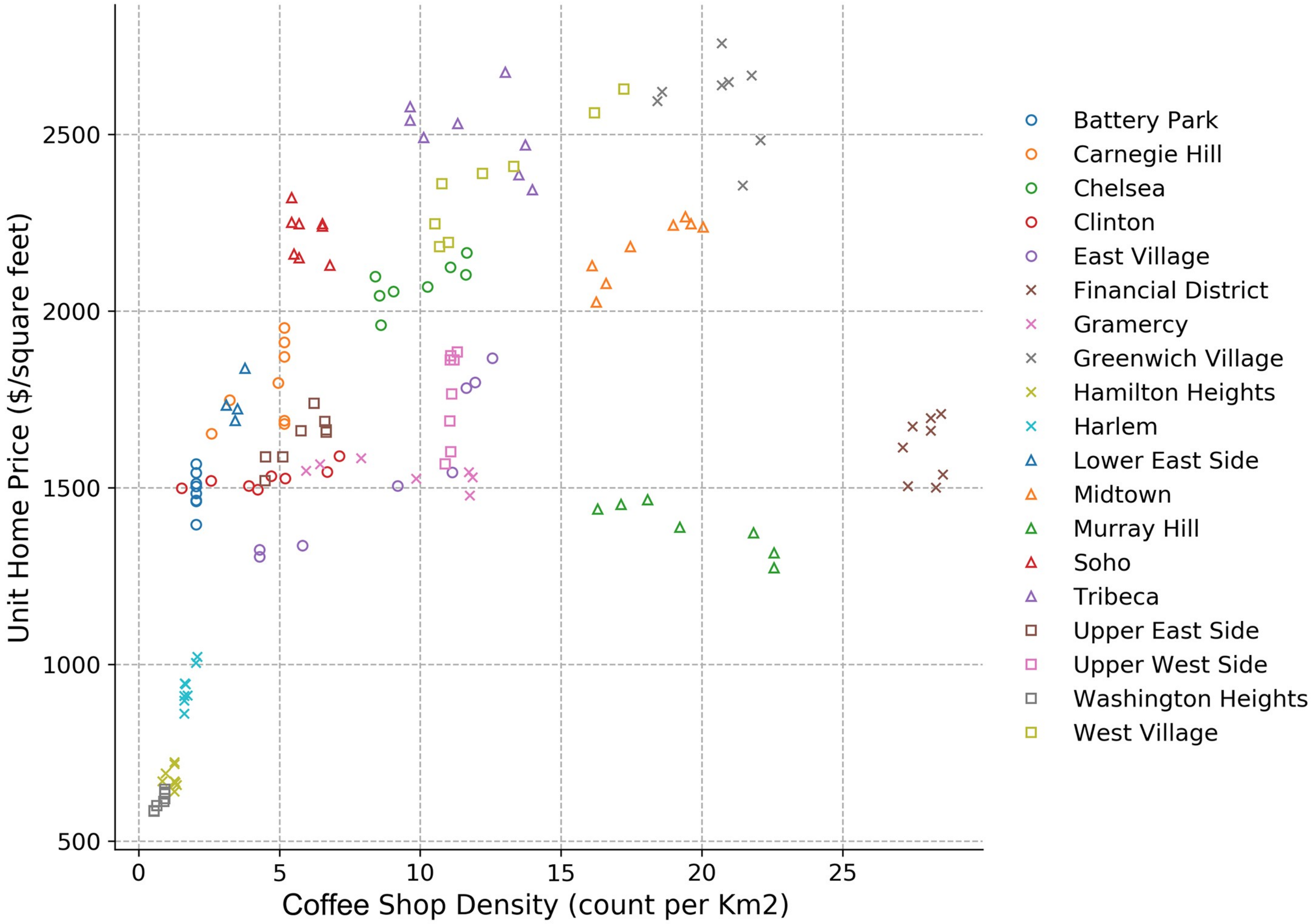

Fig 10. Log(coffee shop density) vs. log(unit home price) grouped by neighborhoods.

https://doi.org/10.1371/journal.pone.0212606.g010

shop density. In the remainder of this paper, these log transforms are still referred to as “coffee shop density” and “home price”.

Fig 10 shows that different neighborhoods over time (shown using different colors and symbols) exhibit different patterns. The home prices of each neighborhood do not always increase as the coffee shop density increases. Fig 11 shows seasonal patterns across all neighborhoods and they seem to be more consistent. Different sub-figures represent different seasons in different years. Home prices seem to increase with coffee shop density for each season in general. In the lower-left corner, both, coffee shop density and home prices are low. In the upper-right corner, home price and coffee shop density have diverging trends.

These different patterns in both figures indicate that there might be other spatiotemporal effects at work beyond the basic scaling relationship model. We will use this model as our basic model and introduce a series of adjustments that utilize some commonly recognized temporal and spatial patterns in urban science.

4.2.2 Temporal trend and lag. **Global trend.** Every city’s real estate market is affected by global market forces and the economy. The economic cycle is a commonly recognized trend [41] and refers to market fluctuations over certain periods, e.g., ten years. While there are many modeling techniques to estimate economic cycles, our data does only cover a two-year

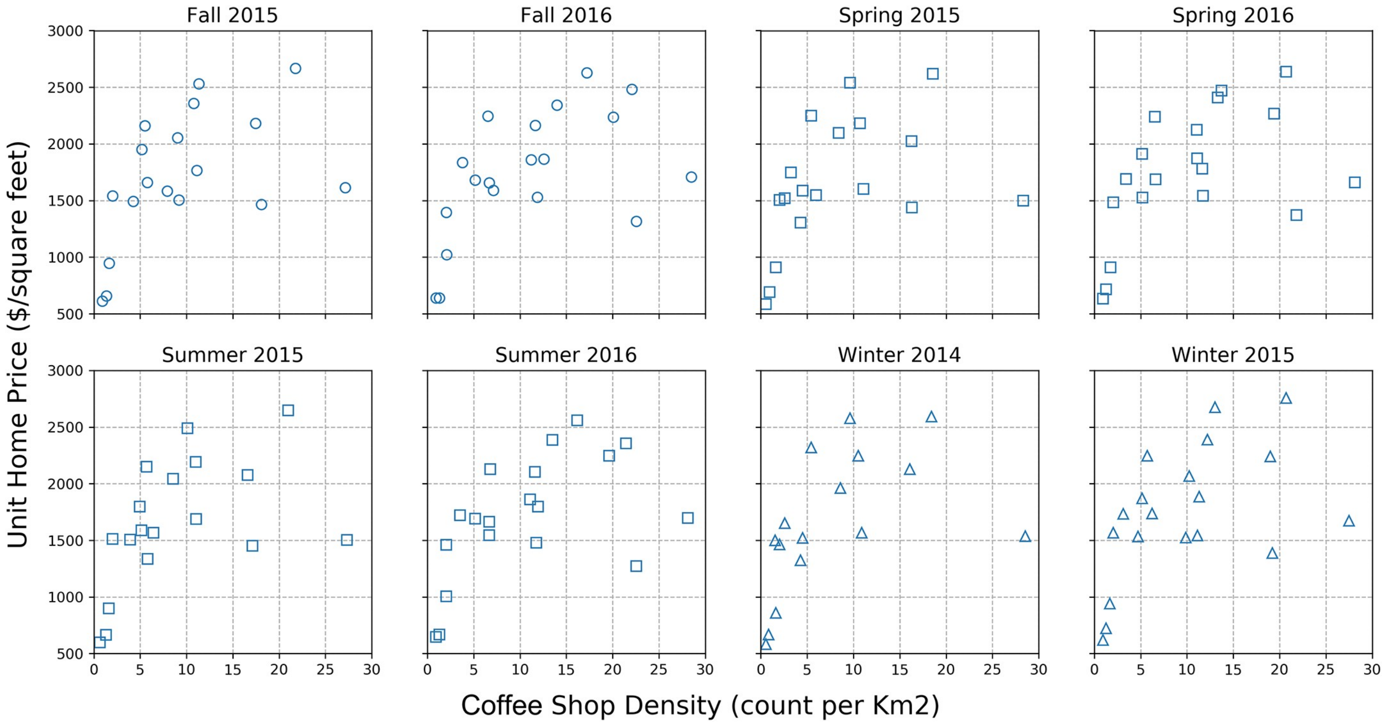

Fig 11. Log(coffee shop density) vs. log(unit home price) grouped by seasons.

https://doi.org/10.1371/journal.pone.0212606.g011

period of a typical ten-year cycle and as such it could cover any portion of the cycle (peak, bottom, etc.). We use the following general polynomial model for this trend: $\tau + \alpha_1 t + \alpha_2 t^2$. Here t is an index capturing sequential seasons ([1, 8] of our two year period), τ is the intercept term, α_1 and α_2 are coefficients. This function can generalize the situation mentioned before. If α_2 is zero (or not statistically significant) this model is reduced to a linear trend model. If α_1 and α_2 are both zero (or not statistically significant), it becomes a stationary process.

Creating one model that captures all neighborhoods, we need to address the differences in magnitudes of prices existing in those neighborhoods. We observed that during a recession, the absolute devaluation of properties of higher value would be much more than the absolute devaluation of properties with lower value. The reverse can be said during boom periods. We use the mean value of the unit home price for the entire two year period $\overline{P_n}$, with n representing different neighborhoods to normalize home prices. The scaled polynomial model becomes $\tau + \alpha_1 \overline{P_n} t + \alpha_2 \overline{P_n} t^2$. The modified scaling relationship model is then as follows.

$$\log(P_t) = \tau + \beta \log(C_t) + \alpha_1 \overline{P_n} t + \alpha_2 \overline{P_n} t^2 + \epsilon_t \tag{6}$$

Seasonality. Many articles investigating the real estate market discuss the "seasonality" of home prices (cf. [42–44]). Following a typical approach, seasonality is modeled using dummy variables ([45, 46]). Our model can be restated as follows.

$$\log(P_t) = \beta \log(C_t) + \alpha_1 \overline{P_n} t + \alpha_2 \overline{P_n} t^2 + \sum_{i}^{4} I_{(i)} w_i + \epsilon_t \tag{7}$$

Here, $I_{(i)}$ and w_i are seasonal dummy variables and their coefficients, respectively. With seasonal dummy variables present, the intercept term τ of the scaling relationship model is dropped from the linear model, since a collinearity issue would exist between τ and seasonality. Both seasonality and τ are implicitly estimated by w_i. A larger w_i means a larger seasonality factor, and vice versa.

Temporal lag. The relationship between coffee shops and home prices could be that an increase of coffee shops either (i) leads to higher home prices, (ii) coincides, or (iii) follows home prices. As such, we can add a lag variable for coffee shop density to our model.

$$\log(P_t) = \beta \log(C_t) + \alpha_1 \overline{P_n} t + \alpha_2 \overline{P_n} t^2 + \sum_i^4 I_{(i)} w_i + \sum_j^3 \eta_j \log(C_{t-j}) + \epsilon_t \tag{8}$$

In this new model, $\log(C_{t-j})$ is the last j season's difference to the current season's coffee shop density and η_j is its coefficient.

4.2.3 Spatial trend and autoregression. **Spatial trend**. Many physical or social phenomena, such as the earth's gravity, snow thickness, and population are correlated with their location. Using this basic spatial analysis technique, the spatial trend in observational data is described by means of a two-dimensional polynomial equation [47]. Despite some claims that it does not perform well for the case of real estate [48, 49], we still want to investigate its performance given that it is a basic geospatial method.

For Manhattan, (cf. Fig 8 right) it shows that home prices are higher in the south of the city than in the north. Two-dimensional coordinates themselves do not influence home prices. However, since many spatial phenomena in a city follow a certain spatial configuration, e.g., subway lines, tourism, etc., home prices might be affected by them and as such are implicitly correlated with location.

In our case of incomplete knowledge and variables, space might be a proxy for the missing information and spatial trend analysis might still be valuable for our model. Since coffee shop density could follow a spatial trend, a collinearity issue might exist between location and coffee shop density. The updated model using a polynomial for the x and y coordinates is as follows.

$$\begin{aligned}\log(P_t) &= \beta \log(C_t) + \alpha_1 \overline{P_n} t + \alpha_2 \overline{P_n} t^2 + \sum_i^4 I_{(i)} w_i \\ &\quad + \theta_1 x + \theta_2 y + \theta_3 x^2 + \theta_4 y^2 + \theta_5 xy + \sum_j^3 \eta_j \log(C_{t-j}) + \epsilon_t\end{aligned} \tag{9}$$

Spatial autoregression model. In this model we consider spatial autocorrelation, which is the interdependence between different locations. It is widely used in economy, sociology and biology when one needs to analyze the correlation of neighboring phenomena. It helps in establishing that neighborhoods sometimes exhibit similar characteristics at a certain spatial scale [50]. Recently, the authors of [51] argue that latent factors could be universal behind all of those autoregressive models. In a similar argument, spatial autocorrelation is linked to missing values estimation and interpolation [52]. For example, in the case of Manhattan there are some externalities, such as transportation hubs, parks, and venues that affect more than one neighborhood. From the perspective of latent factors, autoregression models estimate additional hidden parameters that can increase the goodness-of-fit of our model. This is especially the case if its residuals do not show the expected random distribution and they are spatially clustered creating spatial interdependence. There are two basic spatial autoregression models. Eq 10 is spatial lag based and includes lagged dependent variables. Eq 11 is spatial error based and includes lagged error terms (cf. [50]). In both equations, Y_t is the vector of observed home prices for each neighborhood at time t with dimension $n \times 1$. W is a weighted matrix with dimension $n \times n$. X_t is a matrix of regressors with dimension $n \times m$, with m being the number of regressors in this model, including coffee shop density and the variables discussed in the adjustment models. β is the coefficients' vector with dimension $m \times 1$. ϵ_t is the error term. ρ is the coefficient of the spatial lag term. λ is the coefficient of the spatial error term.

$$Y_t = \rho W Y_t + \beta X_t + \epsilon_t \tag{10}$$

$$\begin{aligned} Y_t &= \beta X_t + \epsilon_t \\ \epsilon_t &= \lambda W \epsilon_t + \mu_t \end{aligned} \tag{11}$$

Modeling approach. In our study, we build models starting with a basic model up to involved models that include additional regressors. We do so in order to observe whether subsequent adjustments add power to the model. The simplest model uses a mean value of coffee shop density and home prices across all seasons. One model is built for each season. We use this simple approach to show the applicability of a scaling relationship model.

Next, a comprehensive model with data for all neighborhoods and seasons is introduced. It includes coffee shop density as an independent variable. Based on this model, different independent variables and spatial autoregression methods are added or deleted based on their respective p-value and modeling power.

To assess the performance of our various models, we choose the Akaike Information Criterion (AIC) [53] and the Bayesian Information Criterion (BIC) [54] instead of the typical R-squared metric since AIC and BIC are more resilient to overfitting [55]. In our modeling approach, there are two sources of complexity. The first is an increasing number of different independent variables. The second source comes from the fact that polynomial models have the potential risk of being overly complex for our modeling case. Both AIC and BIC are information-based criteria that assess model fit as metrics for selecting a finite set of models. They both maximize likelihood and penalize an increasing number of parameters and complexity. As such, both are resilient to overfitting and they are widely used for model comparisons in modern statistics. In general, the criteria for model comparison is that lower AIC or BIC values indicate a better model fit. Normally, BIC will give a higher score as it penalizes model complexity more than AIC. One strategy to add or eliminate a variable is that if a variable (i) is not considered statistically significant (p-value), or (ii) a model using it has a similar or worse AIC or BIC value than a simpler model, then this variable would not be included in the next (improved) model. Another strategy is that if coefficients of coffee shop density change significantly when new variables are added, the model is considered a bad one, since they eliminate the contribution of coffee shop density in our model. This reasoning relates to the discussion of spatial trends and spatial auto-regression and that these models might implicitly (location) estimate coffee shop density.

5 Experimental results

The ambition of this section is simple. We want to evaluate the POI Accuracy & Coverage and trend worthiness methods that we defined in Section 4.

5.1 Accuracy and coverage

The OSM dataset extract contains 529 coffee shops, while we were able to retrieve 851 locations from Foursquare. While the former strictly contains coffee shops, so does the latter also include other types of venues (e.g., sandwich and donut shops) due to the API characteristics. Using a strict label similarity score threshold of 0.9 and a 50*m* proximity threshold for location-matching, we were able to match 310 (LCS method) and 316 (Levenshtein method) OSM

POIs to Foursquare data. Out of the 561 unmatched Foursquare POIs, we were able to match 210 (LCS) and 135 (Levenshtein) of them to other POIs in OSM. Overall 351 (LCS) or 426 (Levenshtein) Foursquare POIs and 219 (LCS) or 394 (Levenshtein) OSM POIs could not be matched to a respective equivalent in the other data source. The matched OSM locations are shown in the Fig 12. To see the distribution of the label similarity score S_{LCS}, S_{LD} and location proximity, Fig 13 shows the CDF of the label matching score for pairs of names with distance threshold $< 50m$. We can see that about 35% of them have $S_{LCS} = 1$, which is much better than expected. It means that about 35% of the data can be matched exactly within $50m$. Only 30% of the data has $S_{LD} = 1$. Fig 14 shows the CDF for label pairs with a proximity threshold $< 50m$. Close to 80% of all matching labels are within $30m$ of each other, and about 95% are within $40m$ for both methods. It tells us that the matched labels for both methods have a spatial accuracy of $30m - 40m$.

5.2 Fitting scaling relationships

For our first model case, we try to fit one scaling relationship model per season. Additionally, we use one baseline model for the mean value of coffee shop densities and home prices across seasons. The two variables are shown in Fig 15. The blue line is the fitted regression line. The grey area is the confidence interval of the predicted values. The coefficients and model fitting results of Table 1 suggest that across different seasons the scaling factor β is stable at around 0.30. All the models have very small p-values (less than 0.01), which indicates a good fit. *These results are a strong indication towards the existence of a scaling relationship between coffee shop densities and home prices*. Fig 16 shows the normal probability plot of residuals, which follows a normal distribution with a high goodness-of-fit (R-squared = 0.9389).

Choosing $\beta \approx 0.3$ is a reasonable value, since this scaling relationship model is based on the coffee shop density vs. population and home price vs. population models (cf. Eq 5). The addition of one coffee shop would imply the addition of a sizeable population. Thus, our β is considerably smaller than the values identified in [6], which are directly related to population. Several weaknesses of our basic model are also evident. About half of the points are outside of the confidence interval, and the changing temporal pattern inside one neighborhood in Fig 10 is in stark contrast to the stable spatial pattern of one season in Fig 11. We will address this issue in the models when considering adjustments.

5.3 Model results with adjustments

Section 4 presented a series of models, which added spatiotemporal variables or spatiotemporal methods to the basic model. We use the labels M1, M2, etc. to distinguish the various models (cf. Table 2). The coefficients of parameters, p-values, AIC, and BIC of each model are shown in Table 3.

As mentioned in Section 4.2.1, **M1** is the basic model considering only coffee shop density as independent variable. β is estimated to be 0.3032, which is almost the same as in the case of the simple model using mean values. We will use this model as a baseline in our comparison to assess the various adjustments.

Temporal trend modeling (**M2**) has coefficients that are all significant and with p-values close to 0. Here, α_1 and α_2 represent global trends. To understand what that means, we can assign an arbitrary coffee shop density value (2) and seasonal dummies (winter) to this model. The estimated global trend is shown in Fig 17. During those two years, real estate markets reached a peak, exhibiting a growth pattern and only shrinking somewhat towards the end of the period. The coefficient of coffee shop density, β, drops from 0.3032 in the basic model to now 0.1824. It shows that probably this global market change is correlated with the coffee shop

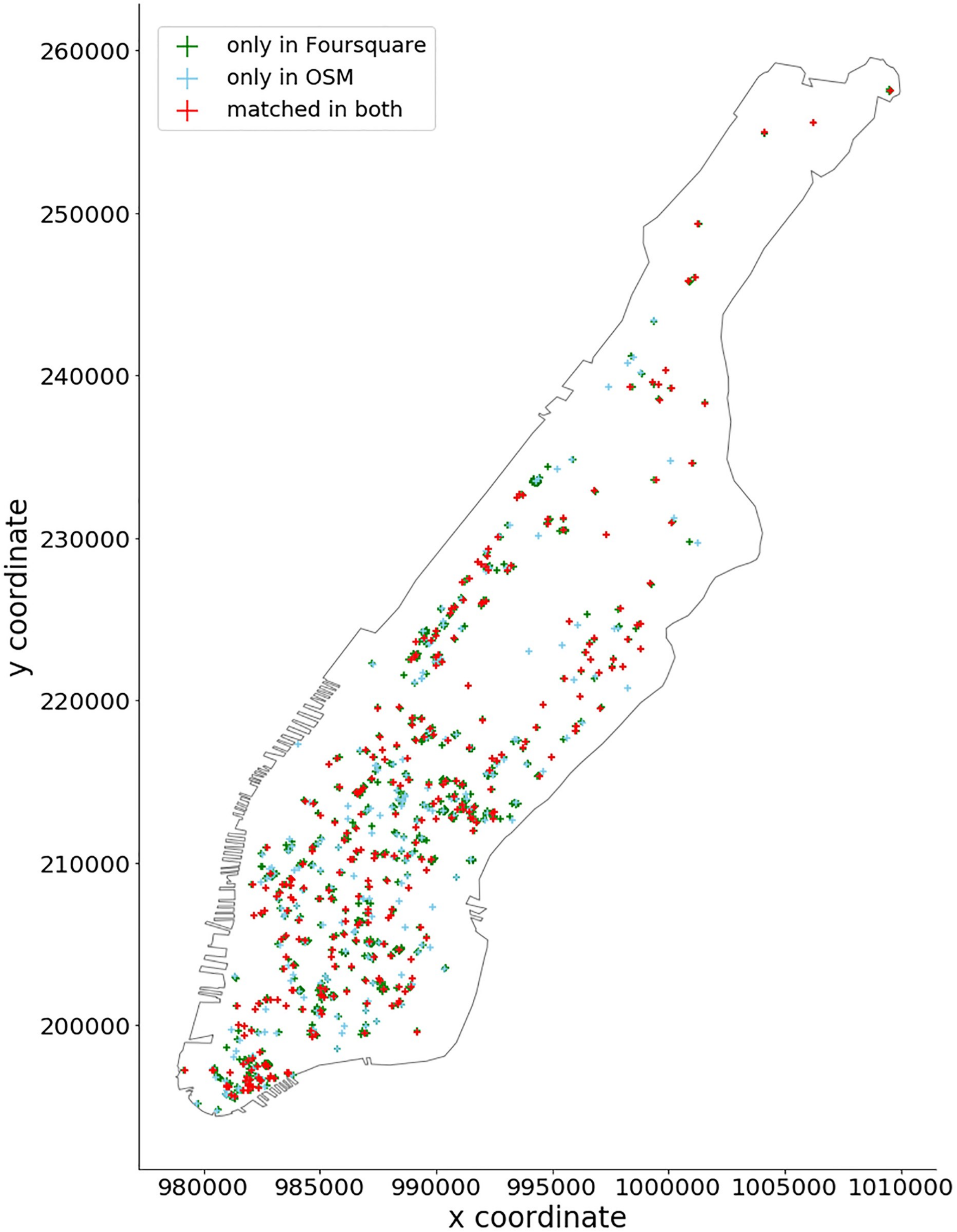

Fig 12. Result of POI accuracy and coverage assessment.

https://doi.org/10.1371/journal.pone.0212606.g012

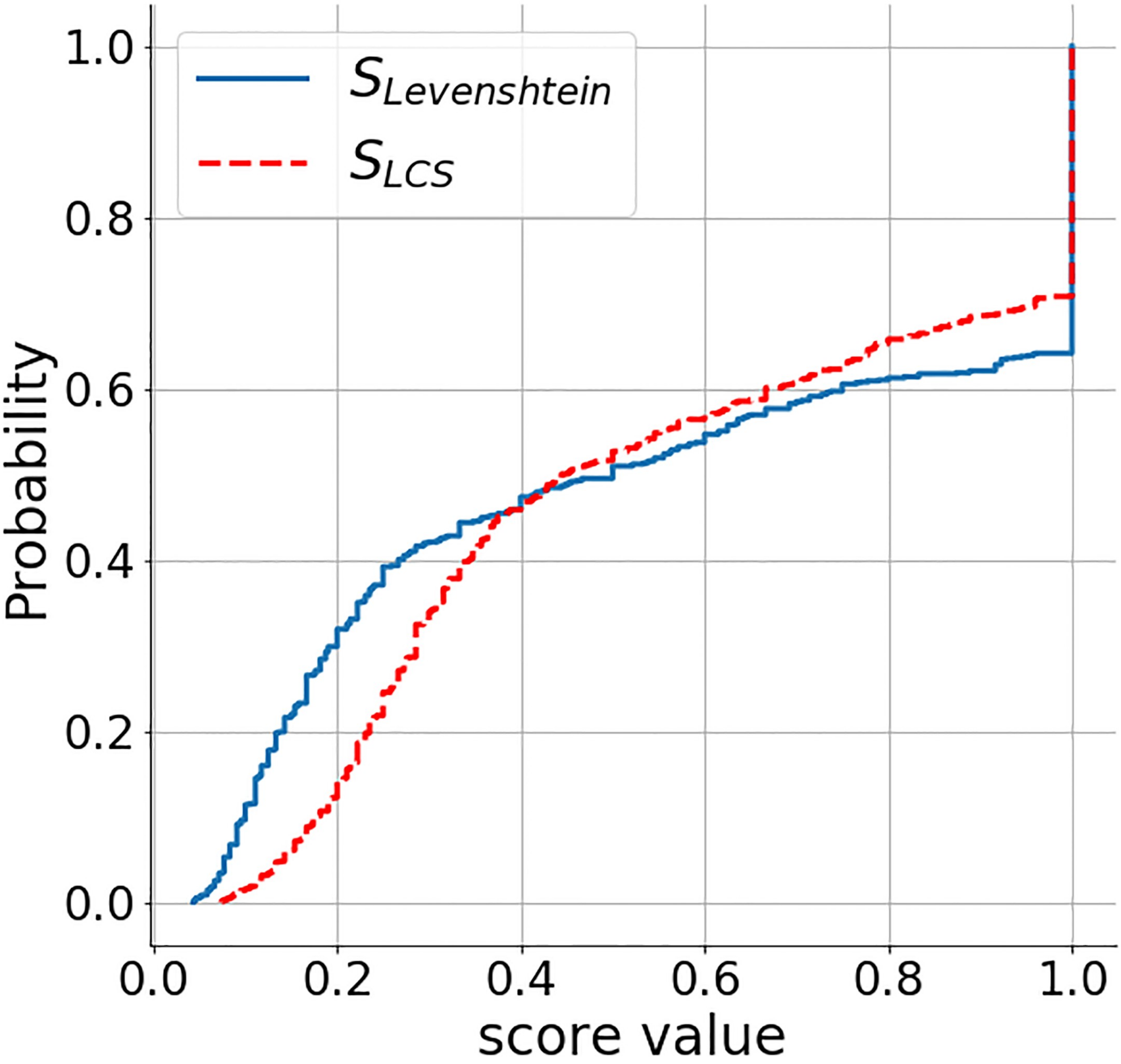

Fig 13. CDF of S_{LCS}, S_{LD} (proximity threshold $\leq 50m$).

https://doi.org/10.1371/journal.pone.0212606.g013

density's overall change during the observation period. The model also captures seasonality. With w_1 representing winter, this means that home prices in winter would have a higher value than during other seasons (cf. Fig 18). For our different models, this seasonal pattern is the same across all models. However, it is in contrast to some other reports, e.g., [56], which claim that prices are lowest during winter. Limited by our available sparse data (only two years) and methods, we cannot conclusively explain this difference, only that it is statistically significant.

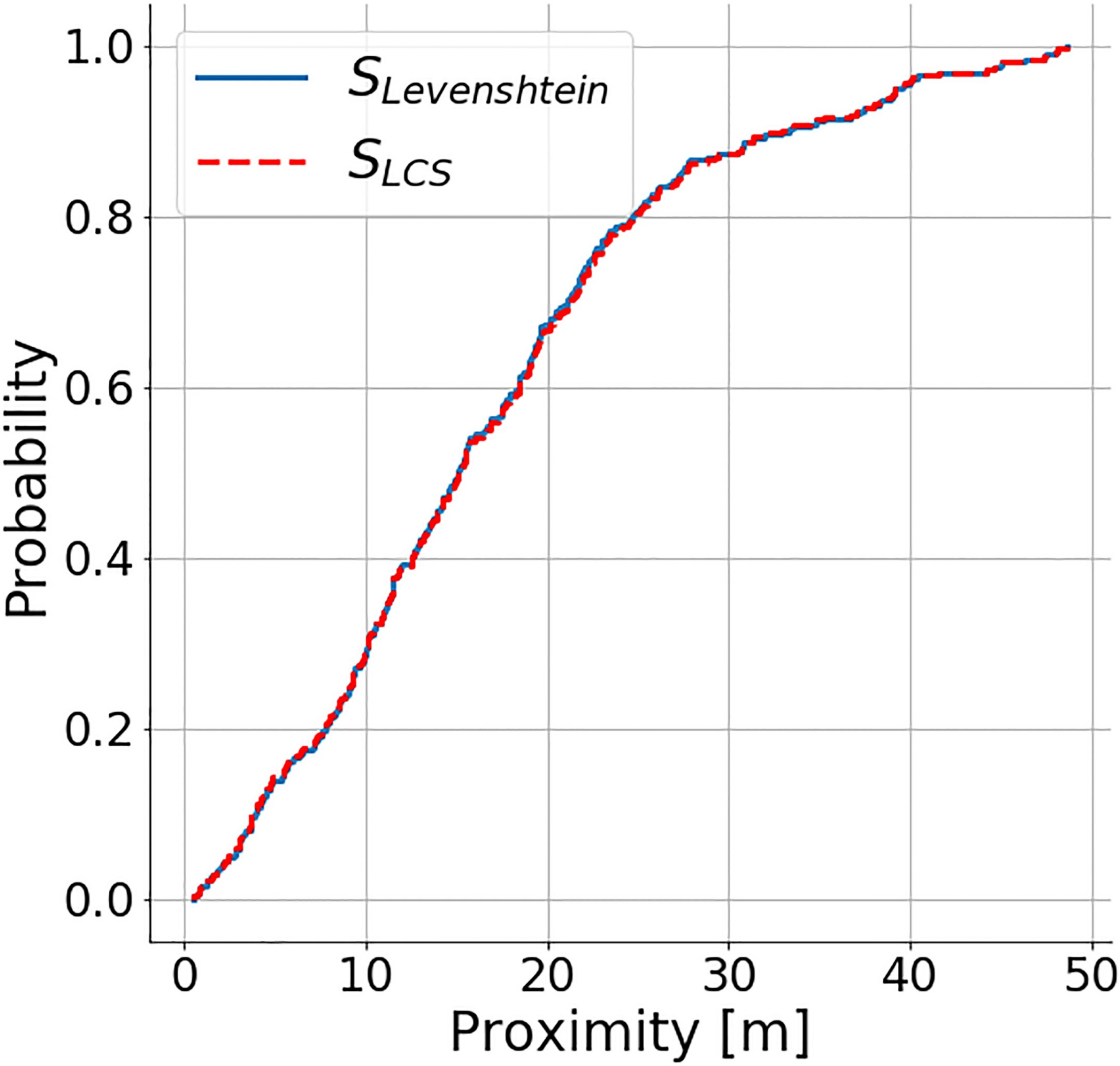

Fig 14. CDF of proximity (S_{LCS}, $S_{LD} \geq 0.9$, proximity threshold $\leq 50m$).

https://doi.org/10.1371/journal.pone.0212606.g014

Moreover, this model has an AIC value of -40.9 and a BIC value of -16.8, which is a significant decrease from M1's AIC and BIC values of 19.8 and 28.7 respectively. As such, we are quite confident that global trends and seasonality impact our model.

Temporal lag modeling (**M3**) did not perform well, since all of the lag parameters have significant p-values, even though AIC slightly dropped to 18.0 (from 19.8 in M1), BIC is 33.3 (increase from 28.7 in M1). This also shows that **M3** does not improve over **M1**. Changes to coffee shop density have no effect beyond a single season (three months period).

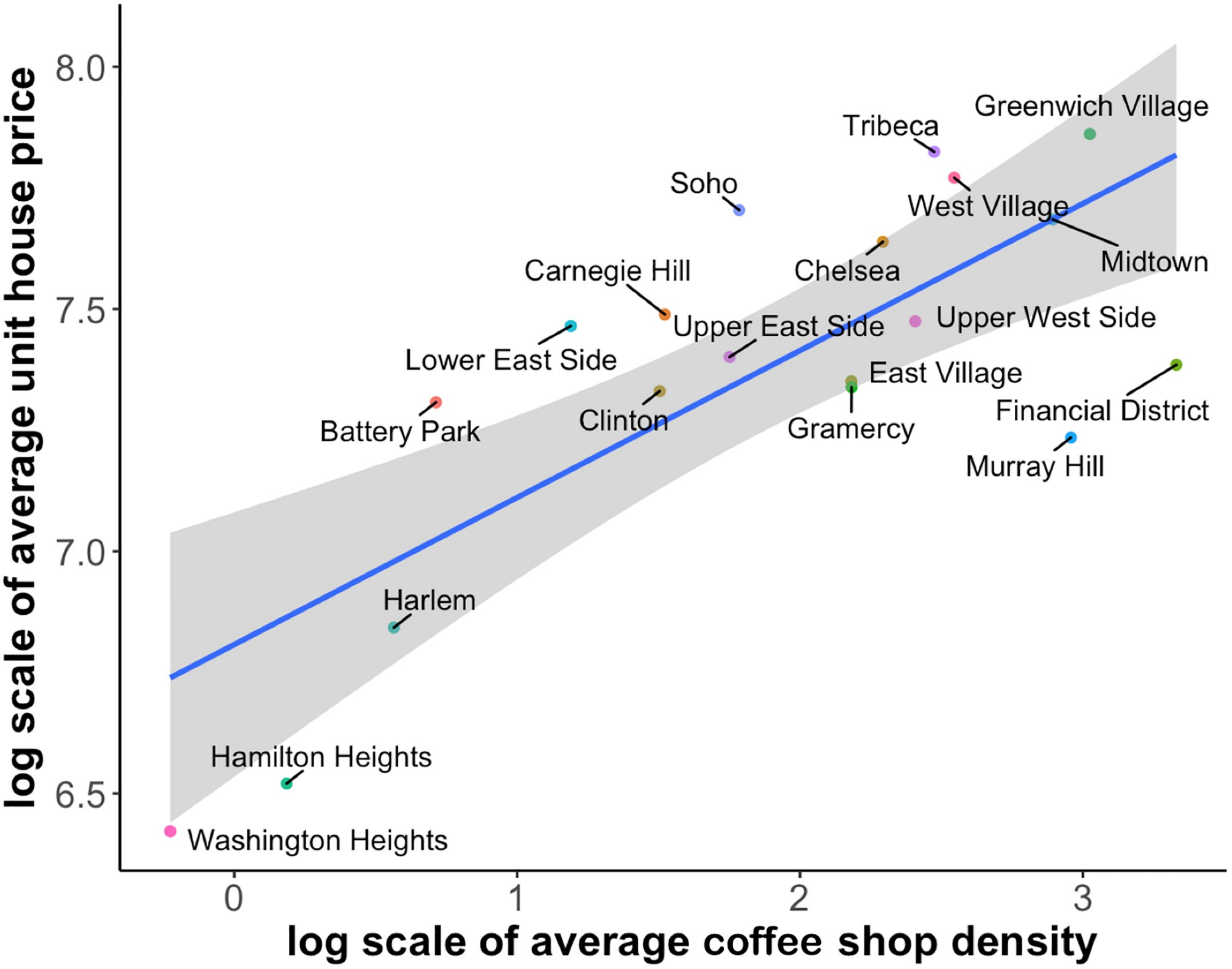

Fig 15. Coffee shop density means vs. home price means for the various neighborhoods.

https://doi.org/10.1371/journal.pone.0212606.g015

Table 1. Coefficients and p-values of basic scaling relationship model.

Models at different seasons	β	$log(Y_0)$	p-value
Winter 2014	0.31	6.85	0.00058
Spring 2015	0.29	6.85	0.00101
Summer 2015	0.32	6.77	0.00053
Fall 2015	0.33	6.76	0.00057
Winter 2015	0.30	6.82	0.00077
Sprint 2016	0.30	6.80	0.00038
Summer 2016	0.30	6.79	0.00038
Fall 2016	0.31	6.77	0.00030
Mean	0.30	6.80	0.00047

https://doi.org/10.1371/journal.pone.0212606.t001

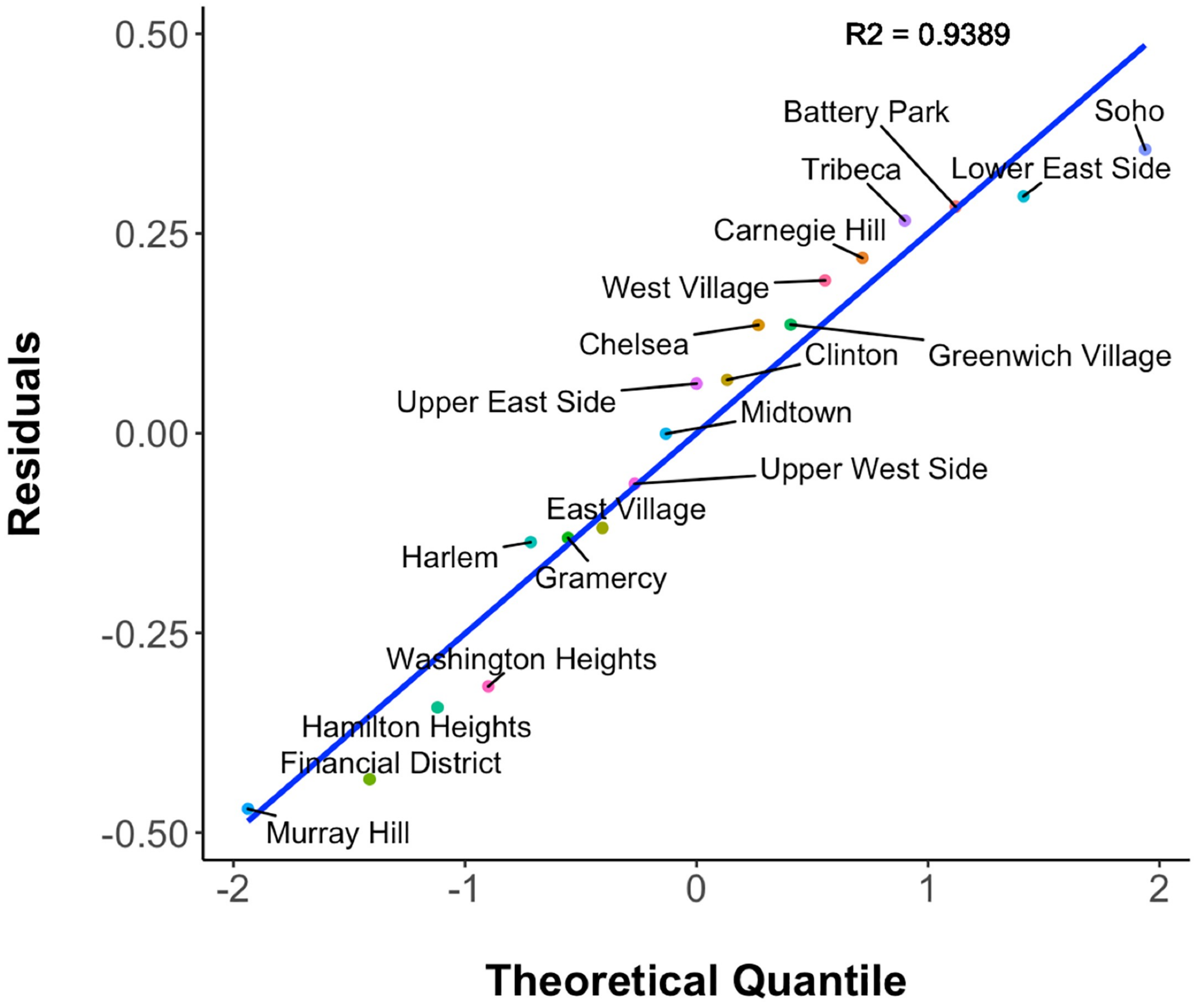

Fig 16. Q-Q plot of residuals of scaling relationship model.

https://doi.org/10.1371/journal.pone.0212606.g016

Table 2. Abbreviation of different models.

abbreviation	modeling components
M1	Basic power law scaling relationship model
M2	Model with temporal trend
M3	Model with temporal lag
M4	Model with temporal and spatial trend
M5	Model with temporal and spatial trend, and spatial lag
M6	Model with temporal and spatial trend, and spatial error

https://doi.org/10.1371/journal.pone.0212606.t002

Table 3. Estimated model parameters with adjustments.

Models with adjustments	Coefficients (p-value)	AIC/BIC	Models with adjustments	Coefficients (p-value)	AIC/BIC
M1	$y_0 = 6.8067(\approx 0)$ $\beta = 0.3032(\approx 0)$	19.8/28.7	M5	$\beta = 0.1895(\approx 0)$ $\alpha_1 = 1.1639e-04(\approx 0)$ $\alpha_2 = -1.0865e-05(\approx 0)$ $\omega_1 = 6.703(\approx 0)$ $\omega_2 = 6.652(\approx 0)$ $\omega_3 = 6.602(\approx 0)$ $\omega_4 = 6.627(\approx 0)$ $\theta_1 = -6.9627e-05(\approx 0)$ $\theta_2 = 2.2269e-05(0.018)$ $\theta_3 = 9.0755e-08(0.004)$ $\theta_4 = 1.5726e-08(0.001)$ $\theta_5 = -7.7248e-08(\approx 0)$ $\rho = -0.0055(0.0163)$	-65.3/-30.8
M2	$\beta = 0.1824(\approx 0)$ $\alpha_1 = 1.326e-04(\approx 0)$ $\alpha_2 = -1.234e-05(\approx 0)$ $\omega_1 = 6.622(\approx 0)$ $\omega_2 = 6.548(\approx 0)$ $\omega_3 = 6.495(\approx 0)$ $\omega_4 = 6.520(\approx 0)$	-40.9/-16.8			
M3	$y_0 = 6.8088(\approx 0)$ $\beta = 0.3057(\approx 0)$ $\eta_1 = -0.0810(0.834)$ $\eta_2 = -0.1623(0.640)$ $\eta_3 = -0.0401(0.888)$	18.0/33.3			
M4	$\beta = 0.2042(\approx 0)$ $\alpha_1 = 1.136e-04(\approx 0)$ $\alpha_2 = -1.071e-05(\approx 0)$ $\omega_1 = 6.551(\approx 0)$ $\omega_2 = 6.490(\approx 0)$ $\omega_3 = 6.442(\approx 0)$ $\omega_4 = 6.466(\approx 0)$ $\theta_1 = -0.1488(\approx 0)$ $\theta_2 = -0.1808(0.002)$ $\theta_3 = 1.27e-07(\approx 0)$ $\theta_4 = 2.004e-08(\approx 0)$ $\theta_5 = -1.015e-07(\approx 0)$	-61.5/-35.5	M6	$\beta = 0.1508(\approx 0)$ $\alpha_1 = 1.839e-04(\approx 0)$ $\alpha_2 = -1.671e-05(\approx 0)$ $\omega_1 = 6.484(\approx 0)$ $\omega_2 = 6.334(\approx 0)$ $\omega_3 = 6.263(\approx 0)$ $\omega_4 = 6.293(\approx 0)$ $\theta_1 = -6.4873e-05(0.001)$ $\theta_2 = 2.2995e-05(\approx 0)$ $\theta_3 = 8.7868e-08(\approx 0)$ $\theta_4 = 1.4554e-08(\approx 0)$ $\theta_5 = -7.0493e-08(\approx 0)$ $\lambda = 0.17997(\approx 0)$	-77.3/-33.6

https://doi.org/10.1371/journal.pone.0212606.t003

Another interesting aspect is a potential spatial trend, i.e., does Manhattan's home price follow a two-dimensional distribution? **M4** looks at this question. As we can see in Table 3, all θs are significant. Assigning arbitrary values to all the other parameters, we can generate a trend surface as shown in Fig 19. It shows a slope pattern with higher values in the Northwest area and lower values towards Southeast. The result does not seem to be intuitive at first. Assuming it is not a fundamental issue with the modeling methods themselves, then a possible explanation is that only three neighborhoods in this study are located to the north of Central Park. Thus, the model is geared towards the trend in the southern area. Neighborhoods around Central Park typically also have high home prices. The overall power of this model improves considerably when compared to the next best one, M2, with AIC/BIC dropping to -61.5/-35.5 respectively. The coffee shop coefficient $\beta = 0.2042$ is only slightly higher than the 0.1824 for the case of M2. Still, this model seems to be quite a good fit.

M5 and **M6** consider a *spatial autocorrelation effect*. Spatial autoregression is widely used in home price analysis. However, it might implicitly estimate latent factors, which are also

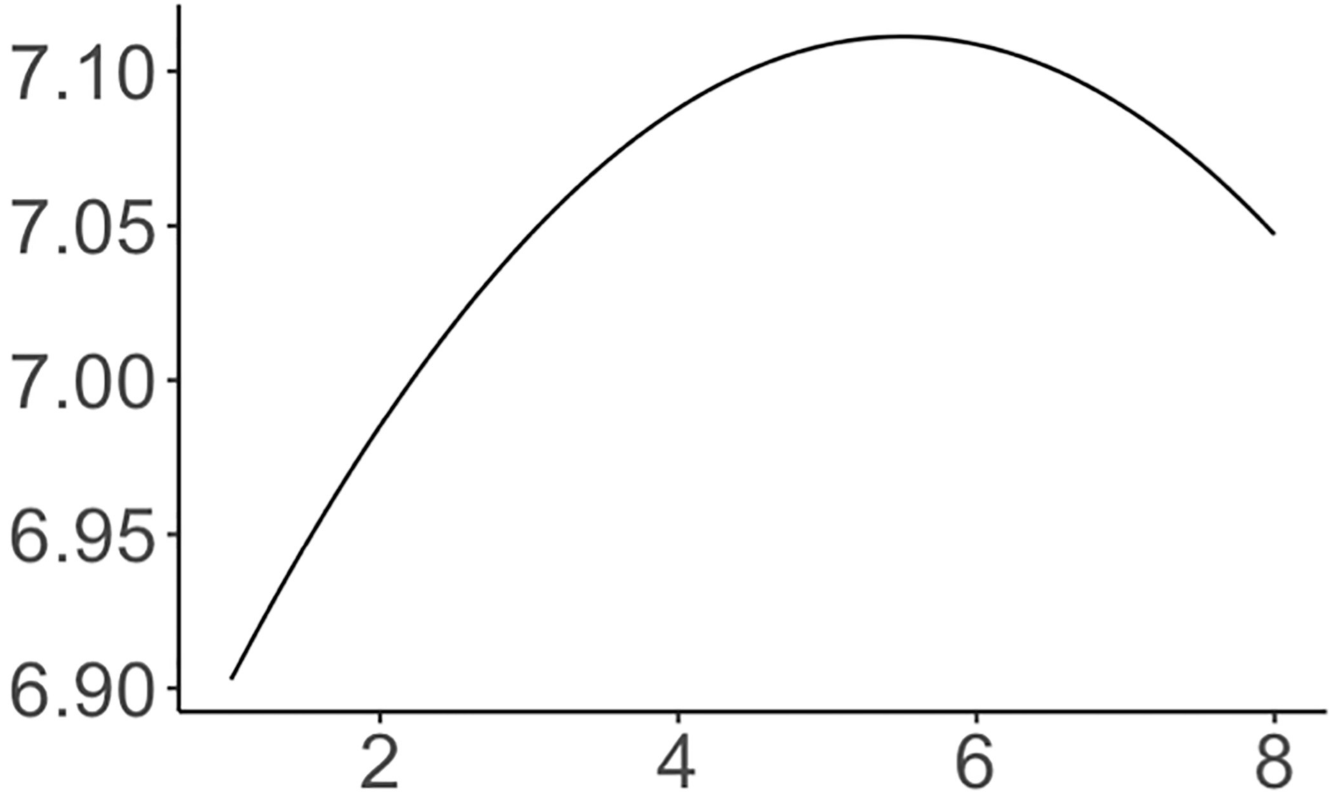

Fig 17. Estimated function of global temporal trend.

https://doi.org/10.1371/journal.pone.0212606.g017

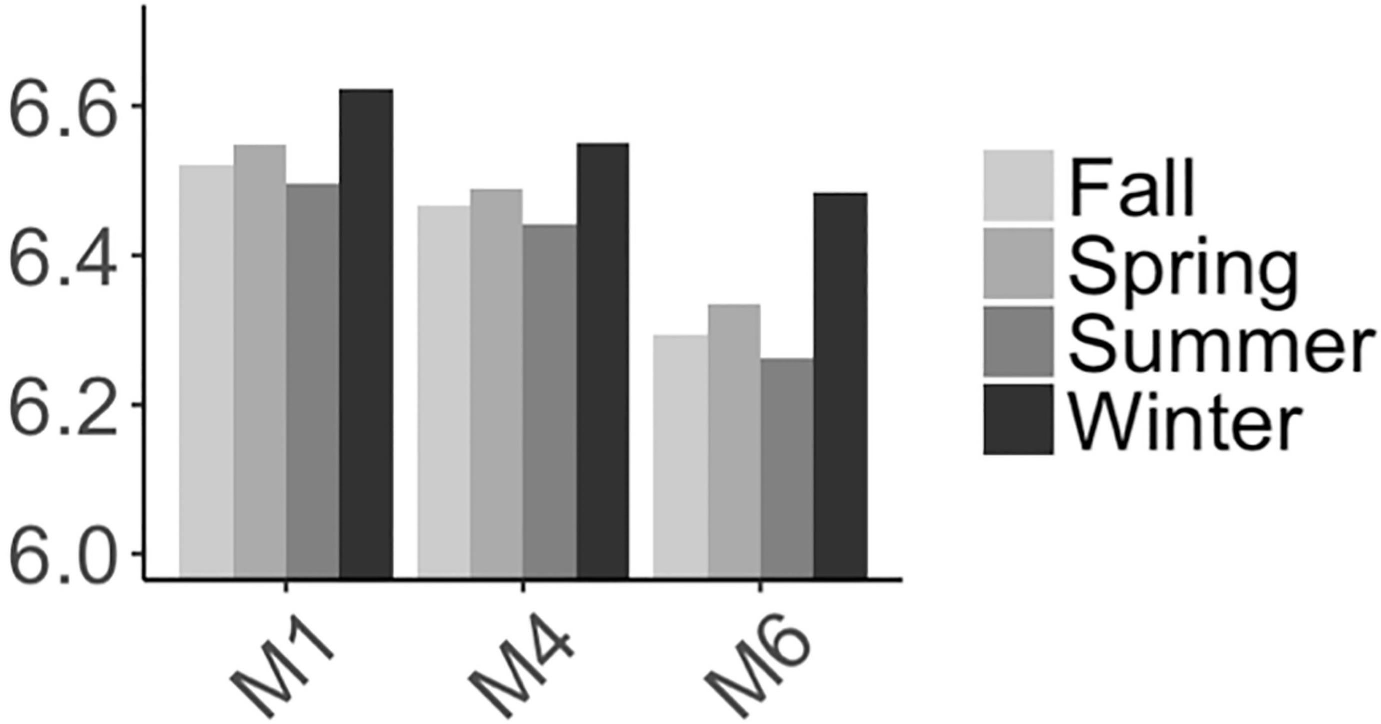

Fig 18. Estimated coefficients of seasonal trend of Manhattan.

https://doi.org/10.1371/journal.pone.0212606.g018

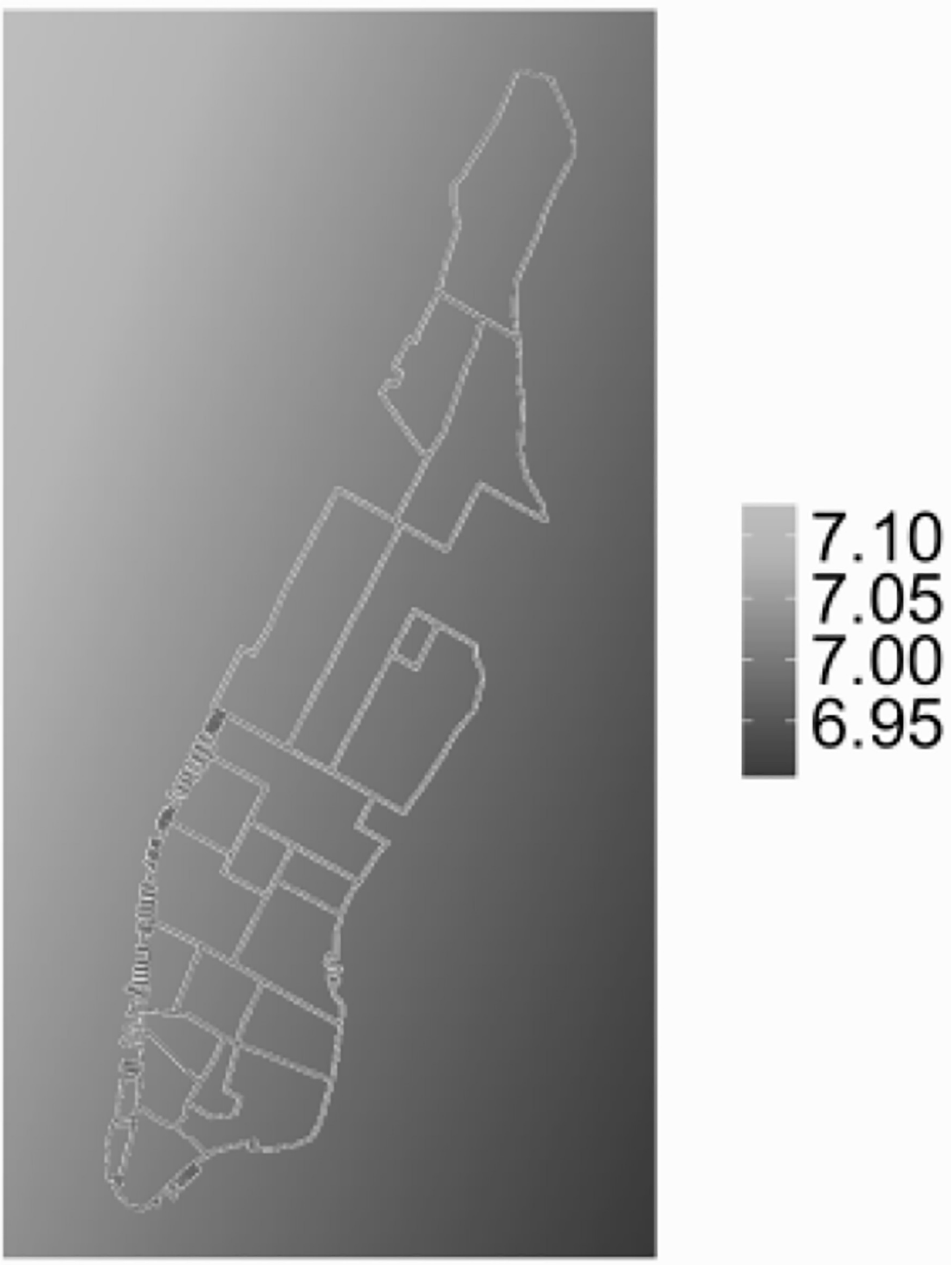

Fig 19. Estimated function of spatial trend of Manhattan.

https://doi.org/10.1371/journal.pone.0212606.g019

estimated by coffee shop densities and spatial trends. The modeling results strongly support this effect. To see whether M5 and M6 capture similar spatial effects, we observe that both models have lower AIC/BIC, -65.3/-30.8 for M5, and -77.3/-33.6 for M6. This means that they are both statistically good models. The variables in both models have very low p-values, which indicates that they explain their effects very well. The θ coefficient values capturing spatial trends in both M5 and M6 are now considerably smaller ($10^4\times$) when compared to M4. This means that spatial autoregression has taken over the interpolation power from spatial trends. The coffee shop density coefficient β is also smaller. However, **M5** and **M6** do have some differences. The autoregression coefficient, ρ, of M5's dependent variable is negative. This indicates a negative autocorrelation between neighbors. However, the error term's coefficient, λ, is positive, which suggests a positive effect between neighbors that is unexplained by coffee shop density or other trends. The AIC/BIC value of M6 is -77.3/-33.6. Both are lower and indicate an improvement over M5. It is hard to say which model, M5 or M6, is right or wrong, since both autoregression models work on latent factors. The model itself did not explicitly give us information about the factors they estimate. Of course, there are other advanced modeling techniques that could be applied in future research and which might have more explanatory power.

5.4 Best model?

After comparing the p-values of coefficients, AIC and BIC, we can identify the two "best model" candidates. **M4** is the best model without an autoregressive process given its low BIC score. **M6** uses spatial autoregression and is the best in terms of overall AIC. **M6** has a worse BIC score, since it is penalized for its complexity by using autoregression terms.

To visually compare the models in terms of prediction accuracy, we can use a Q–Q (quantile-quantile) plot, which is a probability plot that compares two probability distributions by plotting their quantiles against each other. Fig 20 plots the residuals of **M1**, **M4**, and **M6** against an expected normal distribution. M1 shows some deviation for both tails and also in the middle. The residuals for M6 are worse for the right tail portion. M4 seems to be the best of the three models. Its residuals fit to a normal distribution very well. However, M6's total range of residuals (0.8341) is smallest, followed by M4 (0.8641) and M1 (1.032). Fig 20 shows for each neighborhood (points with the same color) that residuals fluctuate more for M6 than for M4 and M1. This is also the reason behind M6's lowest AIC value. This pattern means a better fit to the normal distribution inside a neighborhood.

In general, based on our modeling approach, both M4 and M6 would be good candidates for modeling coffee shop densities in relation to home prices. With M6, even though this model has a better AIC score, there is a concern that the autoregressive process implicitly estimates the coffee shop density. This complexity is penalized as indicated by a larger BIC score than M4. The coffee shop density coefficient $\beta = 0.1508$, which is almost half of the β of the baseline model. For M4, $\beta = 0.2043$. Since it more convincingly considers coffee shop density, we consider *M4 to be the best overall model.*

5.5 Inverse coffee shop effects

The following power law based scaling relationship examines an intriguing, albeit intuitive observation about cities derived from our data. Using a simple model for mean home prices $P_t = \exp^{6.8} \times C_t^{0.3}$, a growth function can be obtained by calculating the derivative of the scaling relationship (cf. Eq 12). This growth function can help us in understanding the coffee shops' contribution to local communities, i.e., how does adding one coffee shop affect home

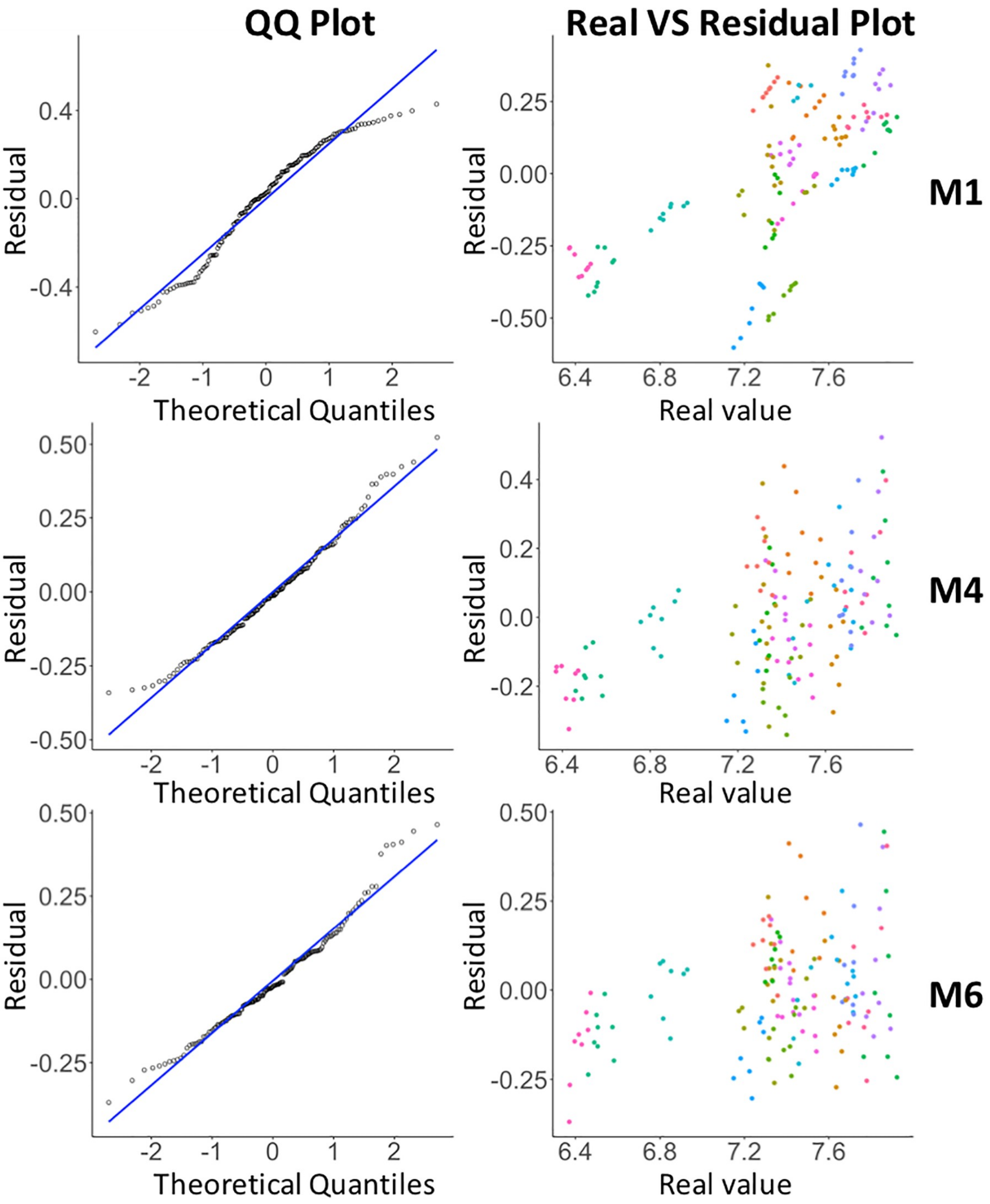

Fig 20. Residuals of fitting models: Comparison using Q-Q plots (left) and real vs. residual (right) for different models.

https://doi.org/10.1371/journal.pone.0212606.g020

prices?

$$\frac{dP_t}{dC_t} = \exp(5.596) \times C_t^{-0.7} \tag{12}$$

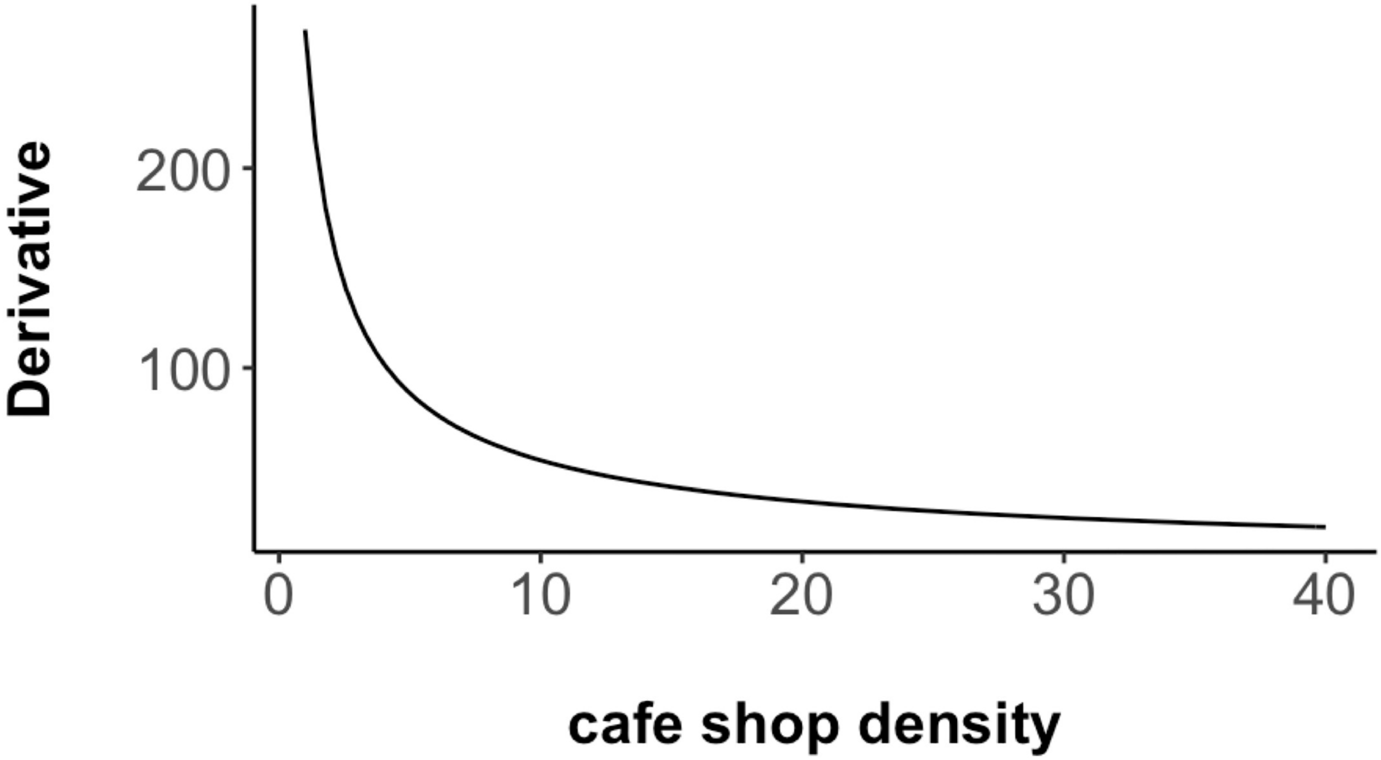

Fig 21. Derivative function of home price and coffee shop change.

https://doi.org/10.1371/journal.pone.0212606.g021

Eq 12's generic structure captures a fundamental feature of coffee shops and how they are related to neighborhood changes. The function shown in Fig 21 indicates that a change in coffee shop density has an inverse sublinear effect $C_t^{-0.7}$ on home prices. In underdeveloped neighborhoods with small coffee shop numbers (zero is not included in this discussion), home prices increase more rapidly when coffee shops are added. This effect corresponds to the left tail of the curve shown in Fig 21. In a highly developed neighborhood with many coffee shops, the impact of a new coffee shop is rather small. This also explains that the different *neighborhoods have different temporal correlations between coffee shop densities and home prices*. The lower left corner of Fig 10 captures neighborhoods in which home prices grow faster with the addition of coffee shops. The upper right corner in the same figure shows neighborhoods that have already a high concentration and the growth dynamic backed by cafe shops is less evident. This growth function now provides an explanation for this phenomenon. It also confirms the intuition that the first, for example, Starbucks coffee shop has the biggest impact on neighborhood development (Harlem vs. the Upper East Side). As such, our work supports that argument that one could really use coffee shop densities as an indicator for gentrification as suggested in [36].

6 Conclusions

The study of urban change has always been of critical importance and has only become widely feasible in recent years due to the availability of open data and user-generated content. However, the quality of such data for this particular use case still needs to be systematically examined as addressed in this work. By examining the accuracy and coverage of OSM POI data we found that it compares favorably to other more authoritative data sources. Compared to

Foursquare, 60% of the OSM POIS could be matched with high accuracy. A more important aspect we looked at is the trend worthiness of the data, i.e., even if it is not as accurate, does it still capture temporal trends such as the relative increase of coffee shops over time? Using statistical models that exploit the power law relationship of various factors in relation to population, we were able to relate coffee shop POI data to urban housing prices. The models that we derived allow us to show that even though OSM POI data (coffee shops in our case) might be incomplete, i.e., not all coffee shops are recorded in a timely matter, such data can still be used in urban analytics research. It is also interesting to observe that the estimated growth function decodes a generic process of urban change and shows that coffee shop data can be an important geo-social indicator.

We can give the following directions for future research. We can apply our models to evaluate further user-generated content datasets, such as restaurants, supermarkets, parking garages, and bike sharing systems. A goal could be to define a respective data matrix that shows which aspects are best suited to predict urban change. On the other hand, we want to extend this model to include all kinds of different urban phenomena and datasets including, but not limited to crime, street vitality, and local business climate. Further, we want to improve our spatiotemporal modelling techniques. Various approaches exist that rely on more advanced statistical models, such as spatial-temporal autoregression [57], and spatial filtering [52]. Overall, although we focused on a very specific model in this work, we argue that the proposed approach (modeling plus user generated content) has considerable potential for social science research.

Supporting information

S1 Appendix. Categories mapping and query details.
(PDF)

Author Contributions

Conceptualization: Liming Zhang, Dieter Pfoser.

Data curation: Liming Zhang.

Formal analysis: Liming Zhang.

Funding acquisition: Dieter Pfoser.

Investigation: Liming Zhang.

Methodology: Liming Zhang, Dieter Pfoser.

Resources: Liming Zhang, Dieter Pfoser.

Software: Liming Zhang.

Validation: Liming Zhang.

Visualization: Liming Zhang.

Writing – original draft: Liming Zhang.

Writing – review & editing: Liming Zhang.

References

1. Singleton AD, Spielman S, Folch D. Urban Analytics. SAGE; 2017.

2. Singh A. Review article digital change detection techniques using remotely-sensed data. International journal of remote sensing. 1989; 10(6):989–1003. https://doi.org/10.1080/01431168908903939

3. Verburg PH, Schot PP, Dijst MJ, Veldkamp A. Land use change modelling: current practice and research priorities. GeoJournal. 2004; 61(4):309–324. https://doi.org/10.1007/s10708-004-4946-y

4. Grant J, Perrott K. Where is the café? The challenge of making retail uses viable in mixed-use suburban developments. Urban Studies. 2011; 48(1):177–195. https://doi.org/10.1177/0042098009360232 PMID: 21174898

5. Liu X, Long Y. Automated identification and characterization of parcels with OpenStreetMap and points of interest. Environment and Planning B: Planning and Design. 2016; 43(2):341–360. https://doi.org/10.1177/0265813515604767

6. Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB. Growth, innovation, scaling, and the pace of life in cities. Proc of the National Academy of Sciences. 2007; 104(17):7301–7306. https://doi.org/10.1073/pnas.0610172104

7. Goodchild MF. Citizens as sensors: the world of volunteered geography. GeoJournal. 2007; 69(4):211–221. https://doi.org/10.1007/s10708-007-9111-y

8. Pfoser D. On user-generated geocontent. In: Proc. International Symposium on Spatial and Temporal Databases (SSTD); 2011. p. 458–461.

9. Haklay M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: Planning and design. 2010; 37(4):682–703. https://doi.org/10.1068/b35097

10. Zielstra D, Zipf A. Quantitative studies on the data quality of OpenStreetMap in Germany. In: Proc. of GIScience; 2010.

11. Ciepłuch B, Jacob R, Mooney P, Winstanley AC. Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In: Proc. 9th International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences; 2010. p. 337.

12. Girres JF, Touya G. Quality assessment of the French OpenStreetMap dataset. Transactions in GIS. 2010; 14(4):435–459. https://doi.org/10.1111/j.1467-9671.2010.01203.x

13. Arsanjani JJ, Zipf A, Mooney P, Helbich M. An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications. In: OpenStreetMap in GIScience. Springer; 2015. p. 1–15.

14. Crooks A, Pfoser D, Jenkins A, Croitoru A, Stefanidis A, Smith D, et al. Crowdsourcing urban form and function. Int'l Journal of Geographical Information Science. 2015; 29(5):720–741. https://doi.org/10.1080/13658816.2014.977905

15. Karagiorgou S, Pfoser D. On Vehicle Tracking Data-Based Road Network Generation. In: Proc. 20th ACM SIGSPATIAL GIS Conference; 2012. p. 89–98.

16. Ahmad M, Karagiorgou S, Pfoser D, Wenk C. Map Construction Algorithms. Springer-Verlag; 2015.

17. Ahmad M, Karagiorgou S, Pfoser D, Wenk C. A Comparison and Evaluation of Map Construction Algorithms using Vehicle Tracking Data. GeoInformatica Journal. 2015; 19(3):601–632. https://doi.org/10.1007/s10707-014-0222-6

18. Karagiorgou S, Pfoser D, Skoutas D. Geosemantic Network-of-Interest Construction Using Social Media Data. In: Proc. GISCIENCE conf.; 2014. p. 109–125.

19. Touya G, Reimer A. Inferring the scale of OpenStreetMap features. In: OpenStreetMap in GIScience; 2015. p. 81–99.

20. Estima J, Painho M. Investigating the potential of OpenStreetMap for land use/land cover production: A case study for continental Portugal. In: OpenStreetMap in GIScience; 2015. p. 273–293.

21. Estima J, Painho M. Exploratory analysis of OpenStreetMap for land use classification. In: Proceedings of the second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information; 2013. p. 39–46.

22. Schlesinger J. Using crowd-sourced data to quantify the complex urban fabric—OpenStreetMap and the urban–rural index. In: OpenStreetMap in GIScience; 2015. p. 295–315.

23. Angel A, Lontou C, Pfoser D, Efentakis A. Qualitative geocoding of persistent web pages. In: Proc. 16th ACM SIGSPATIAL GIS Conference; 2008. p. 1–10.

24. Bader MD, Ailshire JA, Morenoff JD, House JS. Measurement of the local food environment: a comparison of existing data sources. American Journal of Epidemiology. 2010; 171(5):609–617. https://doi.org/10.1093/aje/kwp419 PMID: 20123688

25. Carroll GR, Torfason MT. Restaurant Organizational Forms and Community in the US in 2005. City & Community. 2011; 10(1):1–24. https://doi.org/10.1111/j.1540-6040.2010.01350.x

26. Kubrin CE, Squires GD, Graves SM, Ousey GC. Does fringe banking exacerbate neighborhood crime rates? Criminology & Public Policy. 2011; 10(2):437–466. https://doi.org/10.1111/j.1745-9133.2011.00719.x

27. Mülligann C, Janowicz K, Ye M, Lee WC. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information. In: Proc. International Conference on Spatial Information Theory (COSIT); 2011. p. 350–370.

28. Jonietz D, Zipf A. Defining fitness-for-use for crowdsourced points of interest (POI). ISPRS International Journal of Geo-Information. 2016; 5(9):149. https://doi.org/10.3390/ijgi5090149

29. Touya G, Antoniou V, Olteanu-Raimond AM, Van Damme MD. Assessing crowdsourced POI quality: Combining methods based on reference data, history, and spatial relations. ISPRS International Journal of Geo-Information. 2017; 6(3):80. https://doi.org/10.3390/ijgi6030080

30. Batty M. The size, scale, and shape of cities. Science. 2008; 319(5864):769–771. https://doi.org/10.1126/science.1151419 PMID: 18258906

31. Simini F, González MC, Maritan A, Barabási AL. A universal model for mobility and migration patterns. Nature. 2012; 484(7392):96. https://doi.org/10.1038/nature10856 PMID: 22367540

32. Kennedy C, Pincetl S, Bunje P. The study of urban metabolism and its applications to urban planning and design. Environmental pollution. 2011; 159(8):1965–1973. https://doi.org/10.1016/j.envpol.2010.10.022 PMID: 21084139

33. Waxman L. The coffee shop: Social and physical factors influencing place attachment. Journal of Interior Design. 2006; 31(3):35–53. https://doi.org/10.1111/j.1939-1668.2006.tb00530.x

34. Oldenburg R. The great good place: Cafes, coffee shops, bookstores, bars, hair salons, and other hangouts at the heart of a community. Da Capo Press; 1999.

35. McKenna HP. Urbanizing the Ambient: Why People Matter So Much in Smart Cities. Enriching Urban Spaces with Ambient Computing, the Internet of Things, and Smart City Design. 2016; p. 209.

36. Papachristos AV, Smith CM, Scherer ML, Fugiero MA. More coffee, less crime? The relationship between gentrification and neighborhood crime rates in Chicago, 1991 to 2005. City & Community. 2011; 10(3):215–240. https://doi.org/10.1111/j.1540-6040.2011.01371.x

37. Renthop. Rental Seasonality 2016; 2017. Available from: https://www.renthop.com/study/national/seasonality-2016.html [cited 2019-02-10].

38. Pearson K. Note on regression and inheritance in the case of two parents. In: Proc. of the Royal Society of London. 1895; 58:240–242. https://doi.org/10.1098/rspl.1895.0041

39. Foursquare. Foursquare API documentation; 2019. Available from: https://developer.foursquare.com/docs [cited 2019-02-10].

40. Cormen TH. Introduction to algorithms. MIT press; 2009.

41. Kenton W. Economic Cycle; 2019. Available from: http://www.investopedia.com/terms/e/economic-cycle.asp [cited 2019-02-10].

42. Thwaites G, Wood R. The measurement of house prices. Bank of England Quarterly Bulletin. 2003.

43. McAvinchey ID, Maclennan D. A regional comparison of house price inflation rates in Britain, 1967-76. Urban Studies. 1982; 19(1):43–57. https://doi.org/10.1080/00420988220080041

44. Diewert WE, et al. Alternative approaches to measuring house price inflation. Discussion Paper 10-10, Department of Economics, The University of British . . .; 2010.

45. Kuo CL. Serial correlation and seasonality in the real estate market. The Journal of Real Estate Finance and Economics. 1996; 12(2):139–162. https://doi.org/10.1007/BF00132264

46. Case KE, Shiller RJ. Prices of single family homes since 1970: New indexes for four cities; 1987.

47. Agterberg FP. Trend surface analysis. In: Spatial statistics and models; 1984. p. 147–171.

48. Case B, Clapp J, Dubin R, Rodriguez M. Modeling spatial and temporal house price patterns: A comparison of four models. The Journal of Real Estate Finance and Economics. 2004; 29(2):167–191. https://doi.org/10.1023/B:REAL.0000035309.60607.53

49. Bourassa S, Cantoni E, Hoesli M. Predicting house prices with spatial dependence: a comparison of alternative methods. Journal of Real Estate Research. 2010.

50. Anselin L. Spatial econometrics: methods and models. vol. 4. Springer Science & Business Media; 2013.

51. Stewart B. Latent factor regressions for the social sciences. Harvard University: Department of Government Job Market Paper. 2014.

52. Griffith DA. Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization; Springer Science & Business Media; 2013.

53. Akaike H. Factor analysis and AIC. Psychometrika. 1987; 52(3):317–332. https://doi.org/10.1007/BF02294359

54. Schwarz G, et al. Estimating the dimension of a model. The annals of statistics. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136

55. Gayawan E, Ipinyomi RA. A comparison of Akaike, Schwarz and R square criteria for model selection using some fertility models. Australian Journal of Basic and Applied Sciences. 2009; 3(4):3524–3530.

56. Zhou WX, Sornette D. Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indices. Physica A: Statistical Mechanics and its Applications. 2008; 387(1):243–260. https://doi.org/10.1016/j.physa.2007.08.059

57. Pace RK, Barry R, Gilley OW, Sirmans C. A method for spatial–temporal forecasting with an application to real estate prices. International Journal of Forecasting. 2000; 16(2):229–246. https://doi.org/10.1016/S0169-2070(99)00047-3