The estimation of spatial distribution patterns of different socio-economic status (SES) groups by housing advertisement data and machine learning techniques: a case study in brooklyn, new york

Wei Yuan

Edinburgh School of Architecture and Landscape Architecture, The University of Edinburgh, Lauriston, United Kingdom

ivyweiyuan@hkust-gz.edu.cn

Abstract. Poverty eradication has long been a central issue for sustainable development goals (SDGs), which draws attention to the issue of urban inequalities that can hinder regional economic development and increase unemployment and crime rates. It is critical to understand the local socio-economic distribution pattern for better urban policies and planning strategies. Traditional SES measurements are mainly based on census data and surveys, which are slowly updated and often fail to apply in the latest analysis. The SES inference methods using other data (e.g., satellite maps, nighttime lighting data) lack a theoretical basis and are of coarse resolution. The study takes advantage of the latest data (i.e., online housing advertisement data) and point of interests (POIs) to infer fine-grained block-group-level SES in Brooklyn through machine learning techniques. In addition, natural language processing (NLP) methods are used to derive twelve housing-related SES predictors. The results show that the speculative models and predictors are feasible, and the Global decision tree (GBDT) algorithm is the most efficient of the seven algorithms. The SES distribution map demonstrates a clear socio-economic stratification in Brooklyn. The rich are mainly concentrated in the western and northern areas with a high density of facilities. Based on the analysis of the local SES, three policy recommendations are proposed. First, for the inequitable distribution of facilities, additional investment should be made in the central and eastern regions. Second, a high level of greenery should be given priority in urban planning. Third, in terms of housing, disadvantaged groups should be given attention.

Keywords: socio-economic status, machine learning, natural language processing, SES predictors, Brooklyn.

1. Introduction

1.1. The importance of understanding spatial distribution pattern of SES

Socio-economic status (SES) is a reflection of an individual or group's overall social and economic condition [1]. It is considered as an essential metric of the health of a social society, which is closely related to civil progress [2]. A growing body of literature recognizes the importance of SES since it plays a critical role in spatial analyses of physical and mental illnesses, the phenomenon of urban stratification, gentrification, and changes in social structure [3,4]. In 2019, New York released its

"OneNYC 2050" strategy, outlining eight goals and thirty initiatives aligned with the Sustainable Development Goals (SDGs). Among them, poverty eradication, economic equality, and social justice are of great concern. These issues are in large part due to the spatial segregation between the rich and the poor. It is of critical importance, therefore, to have a deep understanding of these issues through the analysis of local SES distribution pattern. This requires efforts on the advanced SES measurement techniques, using the latest and fine-grained spatial data.

1.2. Problems of previous SES measurement methods

The measurements of SES in previous studies rested on two primary data sources: surveys and census data. Despite the high reliability and public availability of census data, it is slowly updated so that the latest data may not be available for up-to-date research [5]. Also, the spatial resolution of its interpretation is limited to the scale of the census unit, which impedes more detailed studies [6]. With regards to survey data, it may run the risk of containing certain types of errors such as incorrect filling, non-filling, or rejection. To create more accurate, granular and timely statistics at lower costs, many countries are exploring new data and techniques. For example, the U.S. Census Bureau proposed the big data mission project, aiming to create high quality data and improve current statistical products using data from both outside and inside the U.S. Census Bureau. Under this general trend, alternative methods of measuring SES started to emerge with advances in technology. These innovative methods utilized open-source data, such as satellite images, nighttime light data, social media data, Wikipedia articles, and POIs data, to estimate area SES situations [7–11]. Although these approaches are quite innovative and promising, one of the weaknesses is the lack of theoretical basis [8]. In addition, some data, such as satellite images and nighttime light data, are not applicable to high-precision studies (e.g., at neighborhood level, block group level)

1.3. New method

To advance the SES measurement, this study harnesses the potential of online housing advertisement data and POIs to infer block-group-level SES. The advantages of this method are as follows. First, it is anchored in the theoretical linkage between housing and area SES. Bourdieu has mentioned that housing plays a vital role in urban social stratification, not only as a symbol of social wealth but also as an indicator of socio-economic status [12]. Residents of different classes have great differences in their housing tastes and location choices. Second, online housing advertising data are available in real-time, allowing for more accurate, fine-grained, and up-to-date results than census data. Using housing advertisement data, a methodological flow first proposed by Wang et al. has successfully extrapolated the neighborhood SES [5]. However, it is not yet universally applicable due to contextual limitations. Drawing upon NLP and machine learning techniques, this study develops an SES inference model and predictors in the context of Brooklyn. Visualizing the block-group level SES distribution patterns assists planners and policymakers in better understanding social stratification and urban inequalities. On the other hand, the housing-related predictors derived in the inference model can inform the choices of metrics for future SES measurements.

2. Background of the study area

Brooklyn is the largest and fastest-growing borough in New York City. Over the past decade, Brooklyn has grown to become one of New York's economic centers and a haven for academic and cultural prosperity. Yet Brooklyn struggles with severe socio-economic imbalances. It ranks third in the state for income inequality, with nearly a quarter of residents living in poverty and over half of the household incomes above the median. Statistics reveal that over half of the poorest census tracts of New York City are in Brooklyn. The situation in Brooklyn reflects the issues of many fast-growing American cities - economic segregation and polarization.

Block Groups (BGs) are small geographic units that are artificially divided to present census data, which generally contain between 600 and 3,000 people. They are selected as the high-precision study units to better understand the local phenomenon of social stratification and urban inequalities (Figure 1).

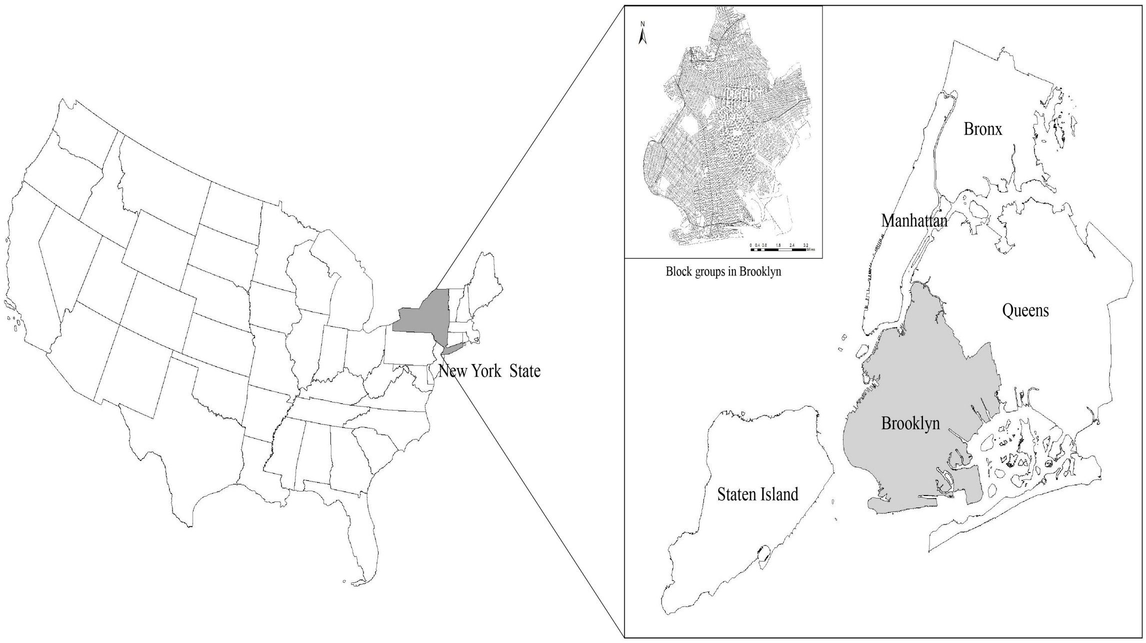


Figure 1. Location of Brooklyn and its block groups distribution.

3. Data and methods

The overall methodological flow can be divided into four sections: (1) Data collection and cleaning; (2) Features derivation through semantic analysis, sentimental analysis, and kernel density estimation; (3) SES inference through seven machine learning algorithms; (4) predictor validation and model validation (Figure 2).

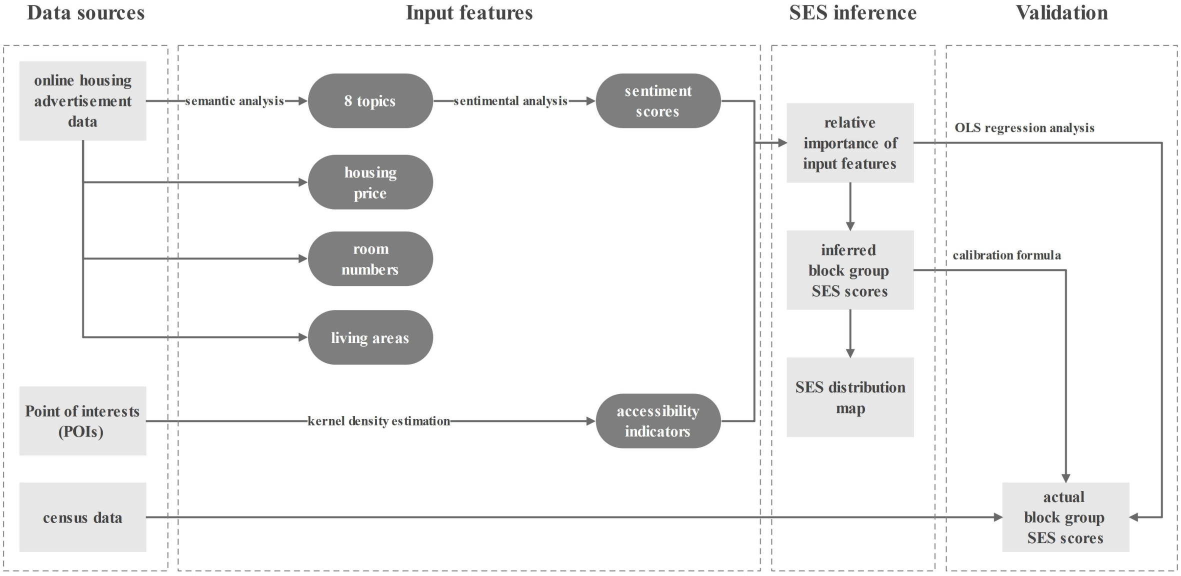


Figure 2. The overall methodological flow.

3.1. Data collection and cleaning

Two main data types used in this study are housing advertisement data and POIs data, both geospatial data that enable the follow-up spatial visualization. POIs data supplement information on the ability of residents to access service facilities. The Zillow website, the most frequently used rental website in the United States, serves as an online housing advertisement data source. The advertising data include all properties listed for rent and sale in Brooklyn from June 2022 to August 2022. Key information (including housing address, home type, longitude, latitude, living area, textual description, number of bathrooms, number of bedrooms, and housing price) is further selected and cleaned in R software. The processed advertising data contains a total of 3,297 entries, covering all block groups in Brooklyn. POIs data are obtained through Goole Map API and are divided into eight major categories: bank, school, entertainment facilities, catering facilities, green facilities, medical care facilities, shopping facilities, and sports facilities.

3.2. Semantic analysis of advertisement descriptions

The publishers' textual descriptions of the residences and living environments can be converted into a set of critical attributes to infer SES. To begin with, textual descriptions need to be pre-processed in R - long paragraphs are divided into separate sentences without punctuation. The manipulation of sentence description is on the grounds of text classification, a necessary natural language processing (NLP) application, which aims to predict topics, sentiments, or other aspects of given textual materials. Naïve Bayes classifier is used as an effective method for identifying the topic of each description sentence through supervised-learning techniques. Specifically, 1500 entries are selected as training data, which are further manually labelled with topics. For example, "walking distance to bus stop" is marked as "transit accessibility"; "The entire house is in move-in ready condition" is defined as "decoration level". The remaining descriptive sentences are used as test data, for which the Naïve Bayes model infers themes based on the training data. Ultimately, eight topics are identified for all housing descriptions: transit accessibility (time or distance to reach public transportation), decoration level (interior decoration condition), natural light (lighting condition), security level (security of the residence), number of private gardens, number of yards, number of garages, and basement level (availability and finish of the basement).

3.3. Sentimental analysis

For the eight topics that have been defined, lexicon-based sentiment analysis is applied to quantify publishers' descriptions into scores. To be specific, all sentences are scored according to emotions contained in words, utilizing sentiment dictionaries and customized words [13]. Two dictionaries are used, i.e., English positive opinion lexicon and English negative opinion lexicon (http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html). To ensure accuracy, words and phrases used in the ads expressing degrees are added to the dictionary, such as "semi-finished", "well-maintained". Words in the ad descriptions are scored individually according to negative, positive, and neutral sentiments. Further, the sentiment scores for words in each sentence are summed and used as the score for the sentence topic.

3.4. Facility accessibility measurement

Facility accessibility reflects a community's ability to obtain the services and resources needed for living and working. It has been proven to be a significant indicator of the socio-economic status of residential areas [9]. In this study, the indicator of facility accessibility is subdivided into eight categories: the accessibility of bank, school, entertainment facilities, catering facilities, green facilities, medical care facilities, shopping facilities, and sports facilities. These metrics are derived using eight main types of POIs. Specifically, R software is used for this process. The buffer zone is set as the area within a 500-meter radius of a housing unit. The density of POIs within the buffer zone are quantified as the indicators of facility accessibility of each housing unit.

3.5. SES inference model

3.5.1. The rationale of the inference model. A weighing technique that is frequently used to assess value dispersion in decision-making is the entropy weight method (EWM). When the dispersion increases, more information can be obtained with greater degree of differentiation. The principle is established that entries with lower dispersion are given less weights. The EWM is an objective weighting method and avoids the bias brought by human factors. Based on the rationale, machine learning algorithms can be used to calculate the weight of each input feature and infer the composite SES score for each housing unit.

3.5.2. Normalize input features. Based on the previous steps, twenty-one housing-related features are derived to be examined for the inference model, including housing price, living area, bathroom numbers, bedroom numbers, quantity of private gardens, quantity of yards, number of garages, basement level, decoration level, natural light, security level, transit accessibility, bank accessibility, school accessibility, entertainment facility accessibility, catering facility accessibility, green facility accessibility, medical care facility accessibility, shopping facility accessibility, and sports facility accessibility. To determine whether these features are reliable predictors for local SES, there is a need to refer to the calculated weight of each feature in the next section. Since the measurements of the features are not consistent, it is necessary to normalize them for comprehensive indicators. To achieve this purpose, Equations (1-2) are used to convert the absolute values of these features into relative values. It is assumed that there are n features and m housing units. Then for the jth housing unit, the absolute value of the i-th indicator is expressed as X_{ij}, the relative values are expressed as Y_{ij}.

For positive features:

$$Y_{ij} = \frac{X_{ij} - \min\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right)}{\max\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right) - \min\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right)} \tag{1}$$

For negative features:

$$Y_{ij} = \frac{\max\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right) - X_{ij}}{\max\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right) - \min\left(X_{1j}, X_{2j}, \ldots, X_{nj}\right)} \tag{2}$$

3.5.3. Model implementation. To speculate SES for housing properties, the following algorithms are chosen for comparison: the multi-layer perceptron neural network (MLP-NN), k-nearest neighbor regression (k-NN), decision tree regression (DT), gradient boosting decision tree regression (GBDT), random forest regression (RF), extra-trees regression (ET) and support vector regression (SVR). Table 1 presents the list of parameters set for each algorithm.

For input features, Equations (3-5) are used to derive the weight of each indicator (W_j) for estimating local SES.

$$P_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_{ij}}, \mathrm{i} = 1, \ldots, \mathrm{n}, \mathrm{j} = 1, \ldots, \mathrm{m} \tag{3}$$

$$E_j = -\ln(n)^{-1} \sum_{i=1}^{n} P_{ij} \ln P_{ij} \tag{4}$$

$$W_j = \frac{1 - E_j}{m - \sum_{i=1}^{m} E_j}, \mathrm{j} = 1, \ldots, \mathrm{m} \tag{5}$$

Features with relatively high weights are used as reliable predictors (Y'_{ij}) for the subsequent operation. For each property unit, the composite SES score (S_i) can be derived from Equation (6):

$$S_i = \sum_{j=1}^{m} W_j Y_{ij}^{'}\,, \mathrm{i} = 1, \ldots, \mathrm{n}, \mathrm{j} = 1, \ldots, \mathrm{m} \tag{6}$$

Table 1. Parameters of each algorithm for SES inference model.

Algorithm	Parameters
MLP-NN	solver: 'adam', alpha: 0.0002, batch_size: 'auto', max_iter: 1000,
k-NN	n_neighbors: 5, weights: ' uniform'

DT	criterion: 'mse',
	splitter = 'best',
	max_features = n_features

GBDT	n_estimators: 500, learning_rate: .5, loss: 'ls', max_depth: 3, criterion: ' friedman_mse', max_features: n_features
RF	n_estimators: 500, criterion: "mse", max_features: n_features
ET	n_estimators: 500, criterion: "mse", max_features = n_features
SVR	kernel: 'sigmoid', gamma: 1/n_features, epsilon: 0.2, max_iter: 1000,

To follow up, the SES scores of housing units are grouped and averaged, converting to the SES scores of block groups. The block-group level SES scores are further imported into GIS for spatial visualization, to plot the SES distribution map in Brooklyn.

3.5.4. Validation of the inference model and relative predictors. The validation process rests on the actual SES measurement, which uses census data of the same year as advertising data. The SES score of each block group is obtained according to eight indicators (population percentages of aged under 14, aged 65 or more, aged 25 years or more with less than high school education, aged 25 years or more with at least some college education, the labor force unemployed, residents who were non-Hispanic blacks, renter-occupied households paying 30% or more of income to rent, and median household income) and the corresponding assessment criteria.

The verification of the inference model involves calibration formulas based on estimated and actual SES scores. The cross-validation method is used to verify the skill of each algorithm. Specifically, the comprehensive scores of block groups are divided into eight groups of approximately equal size. The first group is used for validation, and the remaining seven groups are used for training. The validation metrics include the accuracy, the percentage of root mean squared error (%RMSE), and the percentage of mean absolute error (%MAE), which can be calculated using Equations (7-9).

$$Accuracy = \frac{\sum (1 - |\frac{y_{i,0} - y_{i,p}}{y_{i,0}}|)}{n} \tag{7}$$

$$\%RMSE = \frac{\sqrt{\frac{1}{n}\sum (y_{i,0} - y_{i,p})^2}}{\bar{y}_0} \tag{8}$$

$$\%MAE = \frac{1}{n}\sum \frac{|y_{i,0} - y_{i,p}|}{y_{i,0}} \tag{9}$$

where n is the number of block groups; $y_{i,0}$ and $y_{i,p}$ are actual and estimated SES scores of the i-th block group; y_0 is the mean value of actual block-group-level SES score. The verification results of the seven algorithms are presented in Table 2.

Table 2. The performance of seven machine learning algorithms.

	Input of all features			Input of reliable features		
Algorithm	Accuracy	%RMSE	%MAE	Accuracy	%RMSE	%MAE
MLP-NN	0.41	25.17	20.07	0.59	17.62	15.47
K-NN	0.36	27.42	24.49	0.48	19.84	17.06
DT	0.56	18.71	14.43	0.81	15.20	16.82
GBDT	0.62	12.76	14.32	0.85	11.05	10.02
RF	0.54	13.74	11.68	0.83	11.56	10.24
ET	0.53	15.23	16.11	0.67	12.07	14.83
SVR	0.43	24.78	23.29	0.51	20.14	19.08

After removing all unreliable features, the performances of all seven algorithms are improved. In terms of accuracy, GBDT (0.85), RF (0.83) and DT (0.81) are ranked first, second and third. The accuracies of the other algorithms do not exceed 0.8, and the final results of the calculations will not be considered. The GBDT algorithm has the best performance in predicting block-group level SES, and its outputs are used to map the SES distribution. For the validation of reliable predictors, multivariate regression analysis is used to examine the correlation of the reliable metrics with actual block-group level SES scores (Table 3). In the regression analysis, the actual SES scores serve as the dependent variables, and reliable indicators are the independent variables. If the results demonstrate a clear correlation, reliable indicators can be regarded as accurate predictors.

4. Results and discussions

4.1. Interpretation of reliable local predictors

Figure 3 shows the weights of the input features calculated by each algorithm. Taking the results of the seven algorithms under consideration, twelve features with relatively high weights are identified as reliable predictors for block-group SES in Brooklyn. The reliable metrics are related to three aspects of the housing: the locations (entertainment facility accessibility, shopping facility accessibility, catering facility accessibility, sports facility accessibility, green facility accessibility, bank accessibility, and school accessibility), housing price, and physical attributes (living area, number of bedrooms, number of bathrooms, decoration level, and basement level). These indicators vary in importance and together serve as the criteria for estimating the socio-economic status in Brooklyn. The results of the regression analysis further justify these indicators.

From the result of the t-test, it can be found that none of the t-values for independent variables is zero, which means that all of them should be included as predictors in the model. Except for bank accessibility and shopping facility accessibility, the p-values for the predictors are far below 0.05, indicating that they altogether have impact on the actual SES. The regression model is proved valid by F-test and Shapiro-wilk normality test (with p-values less than 0.05).

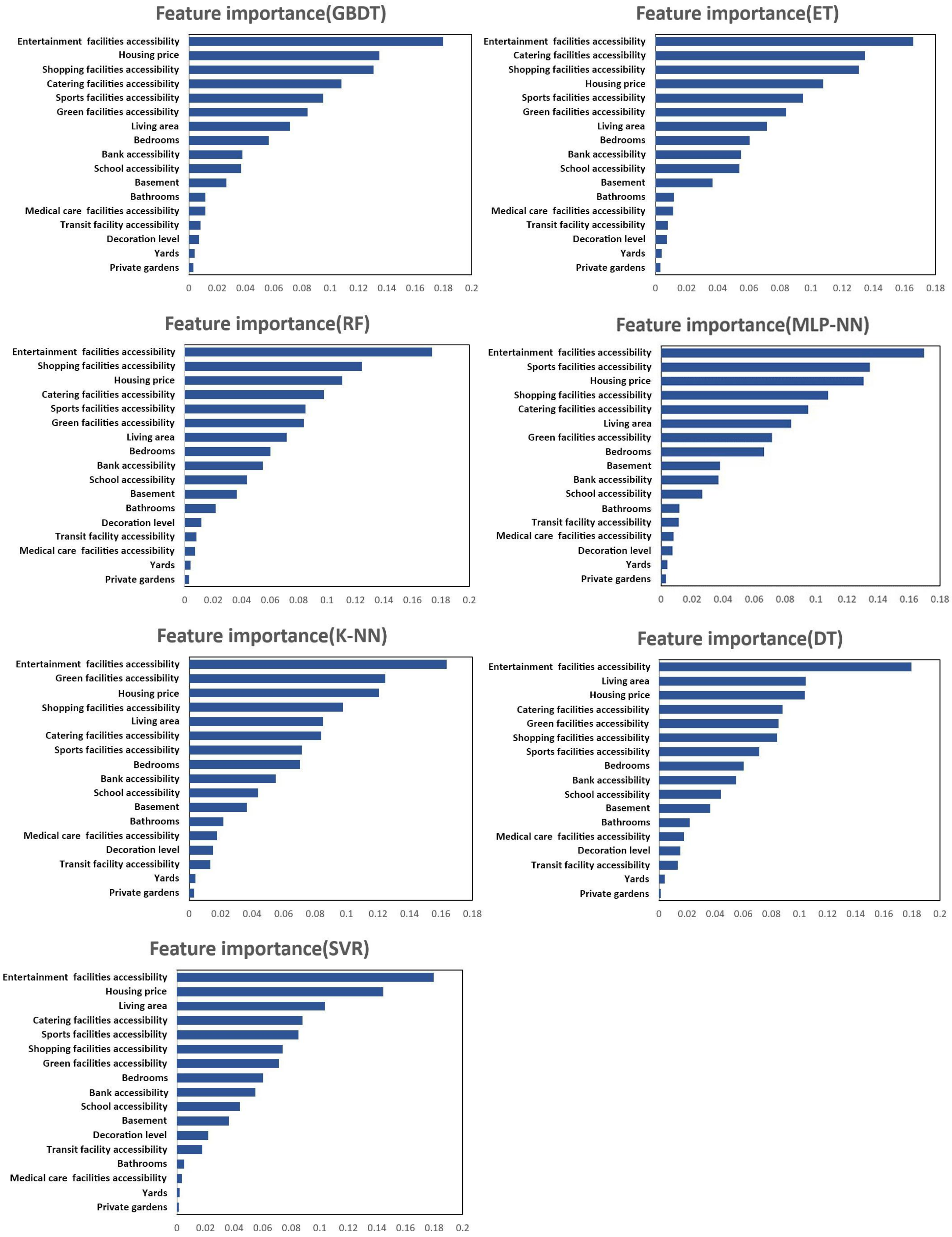


Figure 3. Relative importance of the input features by each algorithm.

Table 3. Regression analysis of reliable indicators and actual block-group level SES scores.

	Estimate	Std. Error	T value	Pr(>|t|)	Signif.
Bank accessibility	1.710e-02	1.678e-02	1.019	0.308	
School accessibility	-1.176e-01	1.209e-02	-9.722	< 2e-16	***
Entertainment facility accessibility	1.705e-01	1.663e-02	10.254	< 2e-16	***
Catering facility accessibility	2.173e-01	2.835e-02	7.665	2.01e-14	***
Green facility accessibility	3.619e-02	9.533e-03	3.796	0.0001	***
Medical care facility accessibility	-5.006e-02	1.825e-02	-2.743	0.006	**
Shopping facility accessibility	-4.602e-02	2.987e-02	-1.541	0.123	
Sports facility accessibility	2.200e-01	1.854e-02	11.867	< 2e-16	***
Bathrooms	1.057e-01	1.383e-02	7.639	2.45e-14	***
Bedrooms	-2.632e-01	1.344e-02	-19.576	< 2e-16	***
Living area	-3.514e-02	9.522e-03	-3.690	0.0002	***
Housing price	2.218e-01	1.193e-02	18.591	< 2e-16	***
Basement level	-2.872e-02	1.112e-02	-2.583	0.0098	**
Decoration level	-4.395e-02	1.090e-02	-4.031	5.60e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1
Residual standard error: 0.7792 on 1206 degrees of freedom
Multiple R-squared: 0.3998, Adjusted R-squared: 0.3928
F-statistic: 57.37 on 14 and 1206 DF, p-value: < 2.2e-16
Shapiro-Wilk normality test
W = 0.98431, p-value < 2.2e-16

Traditionally, the area SES measurements are based on ethnicity, age, and educational attainment indicators. With the development of big data and technology, more factors have been taken into account in the estimations of SES, such as crime rates, quality of the living environment, and access to facilities. Yet housing is rarely mentioned as a factor strongly associated with SES. A few studies have validated the feasibility of several housing-related factors (e.g., housing type, number of bedrooms, and furnishing materials) as indicators of SES, most of which are context-specific. This study demonstrates that the physical attributes of residences and housing prices are related to the identification of different socio-economic groups in Brooklyn, which further corroborates the idea of using housing features as metrics of area SES. The housing price is also an effective predictor confirmed by this study. The possible explanations are two-fold. On the one hand, the value of housing can change residents' subjective perceptions of SES and identity construction logic. On the other hand, housing is an important component of household expenditure, a proxy for income levels [14].

The findings of this study recognize the conclusions of prior studies, which suggest that accessibility to schools, shopping facilities, entertainment facilities, green facilities, and medical care facilities play a significant role in identifying high SES clusters and determining poverty rates [4,9,15]. It is worth noting that green facility accessibility may not be a relevant factor in shaping class stratification when there are abundant landscape resources [5]. Therefore, it is possible that local green resources in Brooklyn are insufficient to meet residents' needs. While contrary to previous studies that regarded transport infrastructure as a trigger of social differentiation, a significant relationship is not found between transit accessibility and block-group level SES in Brooklyn. This inconsistency may be due to New York's well-developed subway, bus and ferry systems, ensuring that public transportation is equally accessible to all classes of citizens regardless of their location [16].

4.2. Brooklyn SES distribution pattern

On the basis of SES scores, all block groups are classified into five socio-economic classes. Figure 4 provides basic information on the overall socio-economic situation in Brooklyn. It can be noticed that the majority of block groups are classified as disadvantaged and less disadvantaged, and the phenomenon of social stratification is apparent.

Further, Global moran's I is conducted to quantify the spatial autocorrelation pattern of 5 SES groups. The value of moran's I statistic is 0.76 with a p-value less than 0.05, indicating a significant spatial autocorrelation of the socio-economic pattern in Brooklyn. The local indicators of spatial associations (LISA) map show the relationship of the various socio-economic groups (Figure 5). High-high (HH) group represents the high SES group adjacent to the high SES group, and the high-low (HL) represents the high SES group adjacent to the low SES group. As illustrated in the map, HH clusters are mainly located in the north and northwest, where the social stratification is most severe. This phenomenon is likely related to the concentration of high-density facilities in northern Brooklyn (Figure 6). The absence of high SES groups in East Brooklyn may be explained by the domination of Black and Latino Americans and high crime rates in the area.

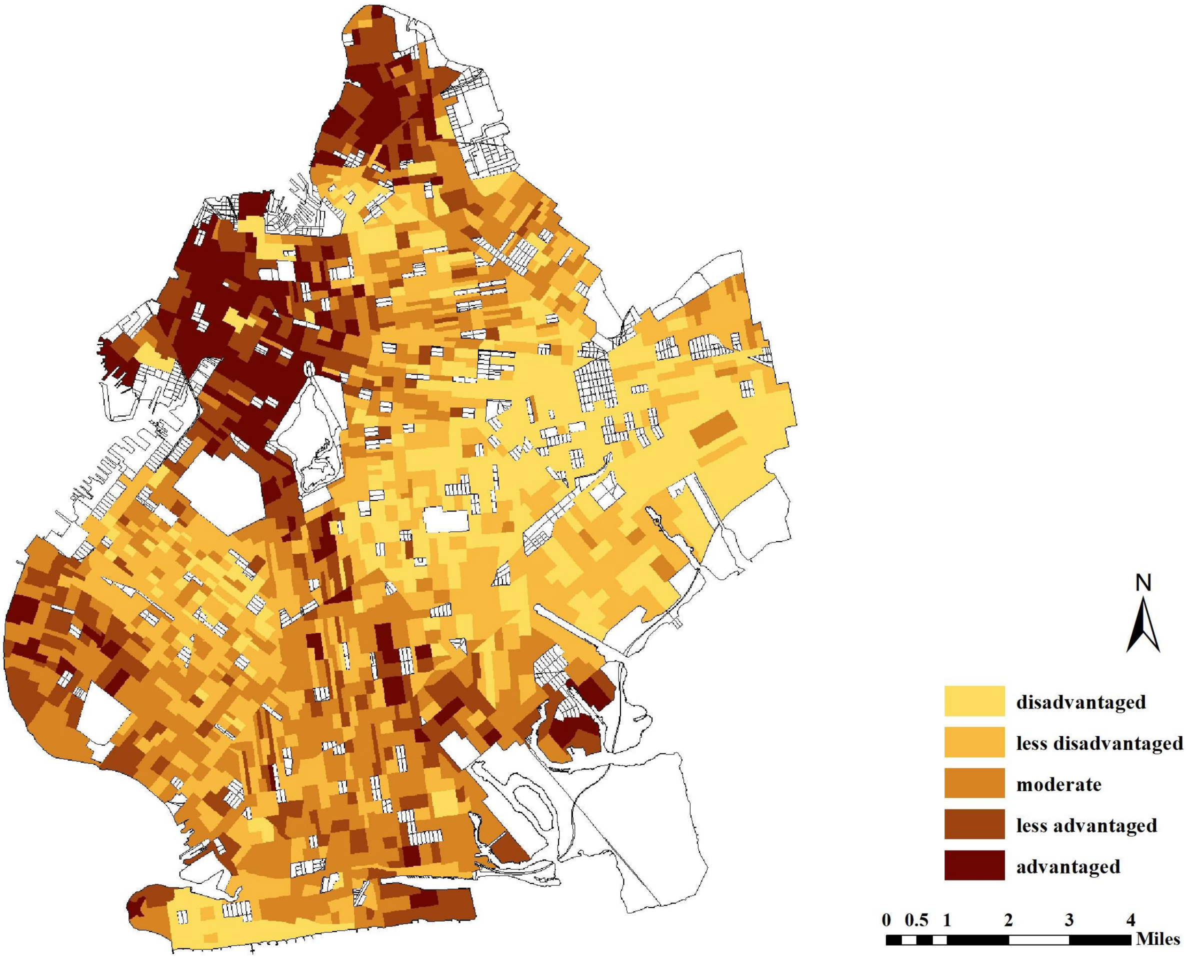


Figure 4. Inferred SES distribution pattern in Brooklyn, New York.

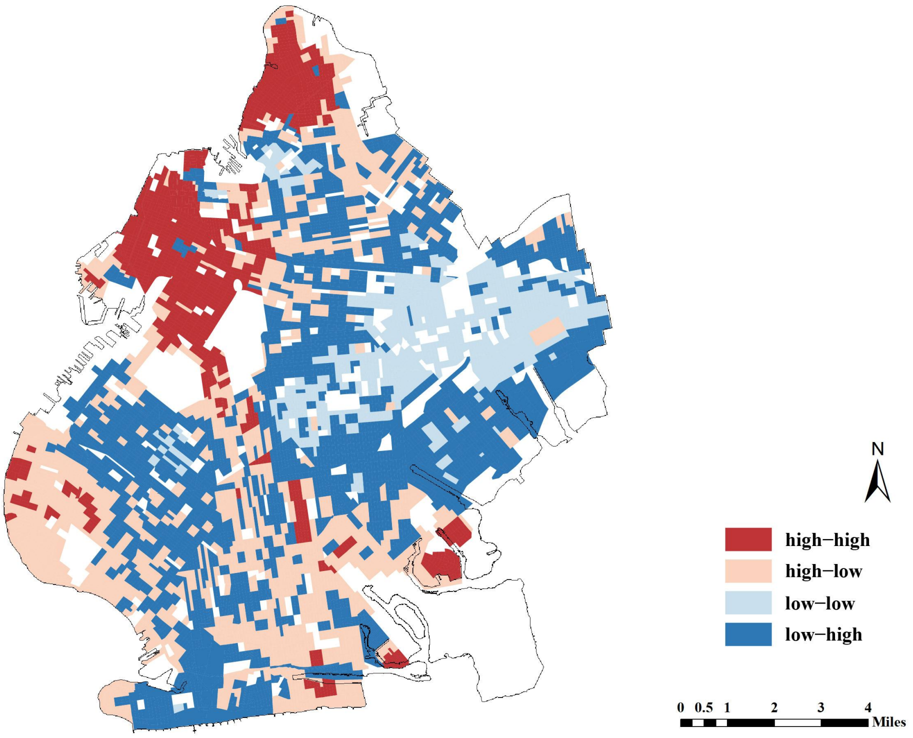
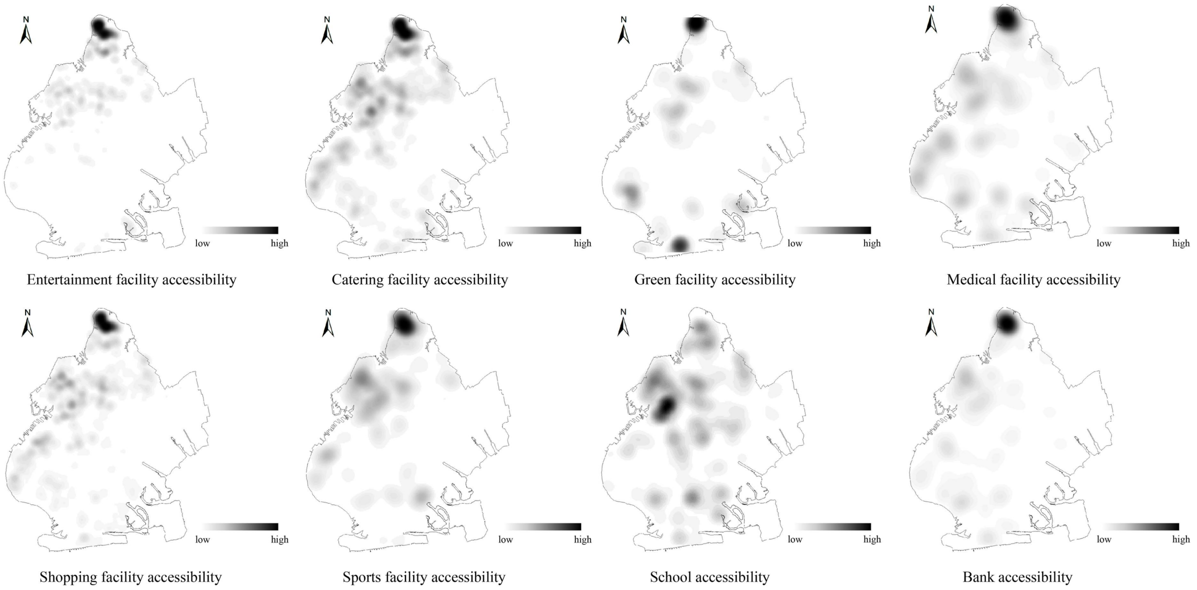


Figure 5. Spatial autocorrelation patterns of different SES groups.



Entertainment facility accessibility

Catering facility accessibility

Green facility accessibility

Medical facility accessibility

Shopping facility accessibility

Sports facility accessibility

School accessibility

Bank accessibility

Figure 6. Kernel density maps of facility accessibility.

4.3. Implications for planning and policy-making in Brooklyn

The spatial distribution pattern of socio-economic classes can greatly affect the economic development and social stability of the region. Researchers have demonstrated that the spatial segregation between the rich and poor can significantly contribute to increased unemployment, crime rates, and poverty rates. In addition, urban stratification is thought to induce disparities in the quality of life and social well-being [17,18]. Based on the findings of this paper, three aspects of recommendations are proposed for planning and policy-making. First, for the inequitable distribution of facilities, additional investment should be made in the central and eastern regions. In particular, Adequate recreational facilities need to be equipped throughout Brooklyn to ensure equal access to all residents. There is also a need to improve the diversity of facilities and the quality of public services to enhance the vitality of each block group. Second, a high level of greenery should be given priority in urban planning, which is related to the quality of life, and physical and mental health of the inhabitants. Third, in terms of housing, disadvantaged groups should be given attention. It is recommended that the government develop mixed-income communities by providing affordable housing in areas where the wealthy congregate, promoting social integration of various SES groups and reducing spatial segregation. Funding and grants are necessary to make affordable housing economically viable for low-income residents.

5. Conclusions

To assist decision-makers in comprehending the patterns of SES distribution in Brooklyn, this study applies an effective method to infer the local block-group level SES. There are three aspects of findings in this study. First of all, the GBDT algorithm is the best among seven algorithms in terms of the performance of SES inference model. This informs the choice of algorithms for future research. Second, twelve housing-related features derived from the semantic and sentimental analysis are identified as reliable indicators for SES inference in Brooklyn, which are in line with the results of previous studies in other areas. In particular, physical characteristics of residences and housing prices are shown to be linked to the identification of various socioeconomic groups in Brooklyn. This further supports the notion of employing housing features as measurements of local SES. However, this study does not consider transit accessibility as a predictor of local SES, which differs from the previous research. Third, the inference results show that the stratification between rich and poor in Brooklyn is obvious, and the high SES groups reside mainly in the north and northwest. A possible explanation for this might be that the density of infrastructures in the north and northwest is much higher than in other parts of the borough. Based on the analysis of SES-related factors, it is suggested that policies for infrastructure, greenery, and housing can be improved to reduce social inequalities in Brooklyn. This study provides researchers and policy makers with a new way of thinking about measuring area SES. By using big data and machine learning techniques, up-to-date regional SES information can be obtained. In addition, the results could further supplement the big data mission project with the latest data. Since the predictors inferred in this study are context-specific, further research can be undertaken to validate the effectiveness of these predictors in other regions. Additionally, more aspects of online housing advertisement data can be exploited, such as photos and virtual reality videos.

References

[1] Baker EH. Socio-economic Status, Definition. In: The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society. John Wiley & Sons, Ltd; 2014, 2210–4.

[2] Arrow K, Bowles S, Durlauf SN. Meritocracy and Economic Inequality. Princeton University Press; 2018. 367 p.

[3] Oakes JM, Kaufman JS. Methods in Social Epidemiology. 2017;603.

[4] Singh GK, Ghandour RM. Impact of Neighborhood Social Conditions and Household Socioeconomic Status on Behavioral Problems Among US Children. Matern Child Health J. 2012, 16(1):158–69.

[5] Wang L, He S, Su S, et al. Urban neighborhood socio-economic status (SES) inference: A machine learning approach based on semantic and sentimental analysis of online housing advertisements. Habitat International. 2022, 124:102572.

[6] Ilic L, Sawada M, Zarzelli A. Deep mapping gentrification in a large Canadian city using deep learning and Google Street View. Ribeiro HV, editor. PLoS ONE. 2019, 14(3):e0212814.

[7] Zhang G, Guo X, Li D, Jiang B. Evaluating the Potential of LJ1-01 Nighttime Light Data for

Modeling Socio-Economic Parameters. Sensors. 2019, 19(6):1465.

[8] Abitbol JL. Interpretable socio-economic status inference from aerial imagery through urban patterns. 2020;2:12.

[9] Niu T, Chen Y, Yuan Y. Measuring urban poverty using multi-source data and a random forest algorithm: A case study in Guangzhou. Sustainable Cities and Society. 2020, 54:102014.

[10] Sheehan E, Meng C, Tan M, et al. Predicting Economic Development using Geolocated Wikipedia Articles. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA: ACM; 2019. 2698–706.

[11] Suel E, Bhatt S, Brauer M, Flaxman S, Ezzati M. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. Remote Sensing of Environment. 2021, 257:112339.

[12] Bourdieu P. Distinction a Social Critique of the Judgement of Taste. In: Inequality Classic Readings in Race, Class, and Gender. Routledge; 2006.

[13] Hu M, Liu B. Mining and Summarizing Customer Reviews. 10.

[14] Chen W, Wu X, Miao J. Housing and Subjective Class Identification in Urban China. Chinese Sociological Review. 2019, 51(3):221–50.

[15] Leslie E, Cerin E, Kremer P. Perceived Neighborhood Environment and Park Use as Mediators of the Effect of Area Socio-Economic Status on Waiking Behaviors. 10.

[16] Dodson J, Gleeson B, Sipe N. Transport Disadvantage and Social Status: A review of literature and methods. 63.

[17] Mouw T. Job Relocation and the Racial Gap in Unemployment in Detroit and Chicago, 1980 to 1990. American Sociological Review. 2000;65(5):730–53.

[18] As Long as There are Neighborhood - John R. Logan, 2016.