Exploring Commercial Gentrification using Instagram Data

1st Sumin Han
School of Computing
KAIST
Daejeon, South Korea
suminhan@kaist.ac.kr

2nd Dasom Hong
Department of Geography
Seoul National Univ.
Seoul, South Korea
aboutcity@snu.ac.kr

3rd Dongman Lee
School of Computing
KAIST
Daejeon, South Korea
dlee@kaist.ac.kr

***Abstract*—Commercial gentrification refers to the replacement of low-value businesses like small local stores into high-value businesses like boutiques and chain stores. A handful of research efforts have been made to identify gentrification and their change by leveraging social media. However, their approaches lack in inferring how much commercial gentrification is developed in a target area and how long it has taken for the area to get to that phase. In this paper, we propose a novel scheme to estimate the commercial gentrification status of a target area and its development in terms of time and geographic dispersion using Instagram data. For this, we define our commercial gentrification phase criteria based on the conceptual model from the urban study. Then, we extract social features from both images and texts of Instagram posts, and leverage regression models to infer the commercial gentrification phase of a target area at the monthly timestamp. We also measure how geographical dispersion of geo-tagged Instagram posts matches the boutiques, which is the physical variable that has the strongest correlation with the commercial gentrification. Evaluation results show that our method yields a good quality of estimation compared to the ground truth. This assures that our method could be a meaningful tool for urban planners and policymakers to investigate and manage commercial gentrification.**

***Index Terms*—Commercial Gentrification Inference, Instagram, Social Media, Regression**

I. Introduction

Commercial gentrification refers to the replacement of low-value businesses like small local stores into high-value businesses like boutiques and chain stores [1]. Zukin et al. [2] found that consumption spaces along streets reveal the spatial form of commercial gentrification. They propose a three-phase transition model that explains how commercial gentrification reflects socio-economic transformation, encompassing sophisticated issues of social class, cultural capital, and race as well as changes in the retail landscape. Besides, commercial gentrification induces changes in place identity. Newcomers in gentrified areas provide opportunities to develop a new place identity catered to the tastes of visitors [3].

There have been a handful of research efforts to leverage geo-tagged social media for discovering new senses of places and correlate them with commercial gentrification. Boy et al. [4] cluster Instagram users into eight groups using the Infomap algorithm, and find that a particular group heavily tags places in the gentrifying districts. Gibbons et al. [5] measure the number of location-based and social interactions between users, and find that the density of such interactions is higher in the gentrifying region. Glaeser et al. [6] explore the correlation between the number of establishments of business categories in the Yelp data and the physical gentrification variables such as demographics. Nikhil et al. [7] try to estimate the level of commercialization via regression. However, the studies above mostly focus on the frequency of the social media data correlated with gentrification. This leads to difficulty in showing the level of gentrification and its development over the time and locations that urban planners are keen to know.

In this paper, we propose a novel scheme, which extracts the embedded features from images and texts of the Instagram posts and infers the phase of the commercial gentrification. We also infer various aspects of commercial gentrification with respect to a target area such as development pattern, and geographic dispersion using Instagram data. For this, we extract various social features such as age and gender of people and place types in an image and word2vec frequency in a text, and rank the importance of these features by recursive feature elimination method based on the support vector regressor. Then, we train several regression models to infer the commercial gentrification phase in a given area at a monthly timestamp. Using this model, we trace the time-series transition pattern of commercial gentrification phases and compare it with the ground truth. We also measure how well the geographical dispersion of geo-tagged Instagram posts matches the locations of boutiques, the physical variable that is most correlated with commercial gentrification. For verification of our proposed method, we select three areas in Seoul that urban studies already identified as gentrified regions. Using various physical variables such as the number of local stores, boutiques, and chain stores, we set up the ground truth of commercial gentrification phases by the criteria defined based on Zukin's model. Evaluation results show that our inference model shows less than 0.3 of RMSE and higher than 0.83 of Pearson correlation, and the geographical dispersion also matches more than 70%. This ensures that the proposed method enables urban planners to infer which commercial gentrification phase a target area lies in and how this area gets developed to the next phase both in time and location before an actual field study.

II. Related Work

A. Measuring Gentrification using Physical Urban Data

Zukin et al. [2] define the three-phase transition model for commercial gentrification that entails changes in retail landscapes constituted by three types of stores; small local stores, boutiques, and large chain stores. In their model, the first phase features the dominance of local stores. In the next phase, the number and share of boutiques increase while those of local stores decrease. In the final phase, the number and share of chain stores hike with rising commercial rents in the third phase.

Some scholars have made attempts to capture and measure commercial gentrification on the basis of Zukin's three-phase transition model. Kim and Choi [8] use building use, land price, and pedestrian volume to capture and analyze the three-phase model of commercial gentrification in Seoul. Ryu and Park [9] also analyze the transitions in the types of stores to identify the occurrence time and speed of commercial gentrification in Seoul.

A growing number of researches from urban studies have explored the central role of retail in neighborhood change in the last decade. On the contrary, methods to define and measure commercial gentrification have been understudied. Researches on commercial gentrification are limited to site-specific and ad-hoc description without methodological standardization across different sites [10]. Gentrification is regarded as global phenomena, and understanding of the various gentrification in different sites is emphasized [1]. Hence, it is necessary to develop a rubric of methods that can be applied across various sites to measure commercial gentrification [10].

B. Measuring Gentrification using Social Data

Boy et al. [4] conducted an in-depth interview to analyze the backgrounds of users and their patterns when they upload the Instagram post in Amsterdam. Then they cluster more than 100 users into eight groups using the Infomap community detection algorithm [11] and analyze the major places tagged in posts from each community. They found one of clusters that they named 'locally oriented gentrifiers' is considered to catalyze the gentrification as they actively tag the places, and the number of posts is concentrated to the rapidly gentrifying districts.

Gibbons et al. [5] define the criteria of demographic gentrification typology: not gentrifiable, gentrifying, and not gentrifying in Washington, DC. Using Twitter data, they measure the density of location-based interaction (LN) which represents two users in the same census block in the same hourly timestamp, and location-based social network (LSN) which means LN with social interaction like following between Twitter users. They conduct negative binomial estimations and find the density of LN and LSN is significantly higher in the gentrifying region. They also visualize word clouds of the Twitter texts to compare the impression of gentrifying and non-gentrifying area.

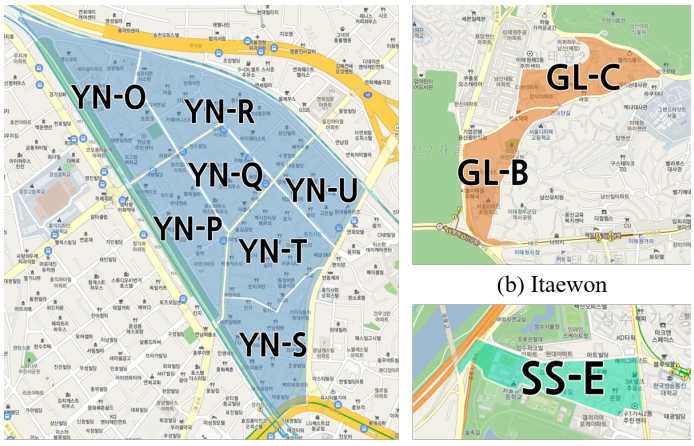
Naik et al. [12] measure the predicted safety scores on a street view image using streetscore CNN-module [7], and find the correlation between streetscore change and the socioeconomic variables change such as population density, level of education, and housing price in 5 US cities. They find a strong correlation between education density and safety, and the relationship between urban physical features of a city and its residents.

Glaeser et al. [6] find the correlation between the changes in the number of each business establishment type in Yelp data and physical variables such as house price, demographics and streetscore [7] using regression. They find the number of Starbucks and cafes are highly correlated to the house price, and also find vegetarian restaurants and wine bars also correlated to other physical changes by streetscore.

III. Datasets

In this section, we first illustrate how we collect physical data and Instagram data. Then, we describe how to establish the ground truth of time-series commercial gentrification phase from physical data by defining the criteria for commercial gentrification phase. Lastly, we explain how we crawl Instagram posts on a given geographic area for our experiment.

A. Physical Data



(a) Yeonnam (b) Itaewon (c) Seoungsu

Fig. 1: Research areas from Yeonnam (YN), Itaewon (GL), and Seongsu (SS), which are representative gentrified area in Seoul.

We choose three gentrified areas in Seoul based on the previous urban researches [13] as depicted in Fig. 1: Yeonnam, Itaewon, and Seongsu. These areas are selected based on the gentrified regions announced by Seoul Metropolitan Government in 2015 to deal with social issues derived from commercial gentrification [14].

In light of the previous work [8], [9], we use the numbers and shares of local stores, boutiques, and franchise stores as physical data to create the gentrification index. We obtain the public store dataset[1] with time period from Jan, 2010 to Dec,

[1] http://www.localdata.kr/

2019, managed by the Ministry of the Interior and Safety, which contains the type, open and closing dates, and exact location of a store. From the dataset, we first select offline sales businesses among all types of businesses. In the case of mail-order businesses, only companies with offline stores are included. Local stores and boutiques are distinguished by the type of building uses under the Building Act, or the ethnicity of food and beverage stores [8], [9], [15]. However, there exist ambiguous cases in food and beverage stores and beauty shops. We use the type of promotion [2], such as the frequency of exposure in Google search and social media, to define boutiques. Finally, we categorize the franchise stores based on the name of shops in the official franchise list[2] managed by the Small Enterprise and Market Service or by the signal of chain stores like Korean suffix *-jum* after the name of the area. e.g., Starbucks Yeonnam-jum.

B. Gentrification Index Establishment

Based on the collected physical data described above, we first define the criteria of the commercial gentrification phase and build the ground truth of a monthly timestamp in each area. Note that the previous work introduced in Section II does not define either specific figures or guidelines for defining each phase of commercial gentrification defined by Zukin [2]. Hence, we define operational guidelines to classify each phase. In addition, we define phase 1.5 and phase 2.5. The former represents the highest growth rate of boutique between phase 1 and phase 2 and helps to analyze the process of gentrification more precisely. The latter is to distinguish between regions where gentrification has been intensified and those that are not. The criteria for commercial gentrification phases in this work is defined as follows:

- Phase 1: The number of local stores is dominant.
- Phase 1.5: The number of boutique stores drastically increased.
- Phase 2 : The number of boutiques and chain stores exceeds the number of local stores.
- Phase 2.5: The ratio of the sum of the boutique and chain store is over 65%.
- Phase 3 : The final phase, when the number and ratio of chain stores have drastically increased.

Based on the criteria above, we label significant phases at monthly timestamps of an area and connect the points by linear interpolation. The pairs of a monthly timestamp and gentrification phase for each area are listed in Table I, and the histogram of ground truth values after linear interpolation for every timestamp in all areas is described in Fig. 2.

C. Instagram Data

Since Instagram API does not support a geographical region based query, we need to go through the following steps to crawl geo-tagged Instagram. First, we crawl the Facebook places on a region slightly larger than our target areas using the Facebook Graph API[3]. This is because locations on Instagram have some errors (about median 15 meters from actual location) as they are tagged unofficially by individual users. Second, we find the places on Instagram with the same names from the crawled Facebook places. Third, we exclude the places on Instagram whose name makes it difficult to determine an exact location, e.g. 'somewhere in Yeonnam-dong', 'Seongsu-dong alleys', or 'my home in Iteawon'. Fourth, we match the exact latitude and longitude for each Instagram place by searching the place name on the store dataset from III-A or Google map. Finally, we crawl Instagram posts whose exact locations are within target areas using the 3rd party library[4], nearly taking a month. The result is 699,461 Instagram posts from 967 Instagram places, and we use this as our dataset. Since there is a substantially small number of Instagram posts before 2014 in Korea, we investigate gentrification from 2014 to 2019.

TABLE I: Pairs of the monthly timestamp and the phase based on gentrification criteria used to establish gentrification index.

Area	List of (yy.mm, phase) - all starts from (10.01, 1)
YN-O	[(14.01, 1), (16.03, 2), (16.12, 2.5), (20.01, 2.5)]
YN-P	[(14.01, 1), (14.06, 1.5), (15.06, 2), (20.01, 2)]
YN-Q	[(14.01, 1), (15.06, 2), (17.06, 2.5), (20.01, 2.5)]
YN-R	[(14.01, 1), (15.12, 2), (17.02, 2.5), (20.01, 2.5)]
YN-S	[(13.12, 1.5), (14.01, 1.5), (15.12, 2), (19.01, 2.5), (20.01, 2.5)]
YN-T	[(11.08, 2), (14.01, 2), (14.11, 2.5), (20.01, 2.5)]
YN-U	[(14.01, 1), (16.05, 1.5), (17.10, 2), (20.01, 2)]
GL-B	[(12.11, 1.5), (11.04, 2), (14.01, 2), (14.06, 2.5), (20.01, 2.5)]
GL-C	[(13.10, 1.5), (14.01, 1.5), (14.04, 2), (15.10, 2.5), (20.01, 2.5)]
SS-E	[(14.01, 1), (16.02, 1.5), (17.02,2), (19.04, 2.5), (20.01, 2.5)]

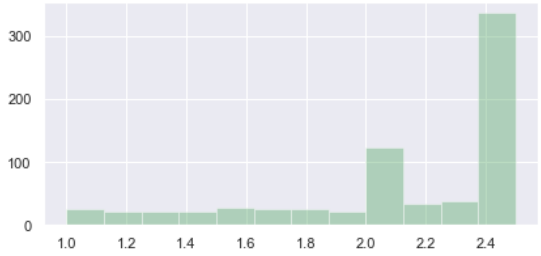


Fig. 2: Histogram of ground truth phase values corresponding to timestamps in all areas. The total number of ground truth data at all monthly timestamps for all areas is $72 \times 10 = 720$.

IV. Method

A. Social Feature Extraction from Instagram Data

In this section, we describe how we extract social features from Instagram data to infer a commercial gentrification phase. The previous researches [5], [6] have shown that commercial gentrification is correlated to demographics, business types, and keywords in social media, thus we extract such features from image and text using recent machine learning based computer vision techniques and Word2Vec-based clustering

[2]http://www.sbiz.or.kr/fcs/list.do

[3]https://developers.facebook.com/tools/explorer

[4]https://instaloader.github.io/

method, respectively. We create a 497-dimensional social feature vector of an area at each monthly timestamp, which consists of 14 age-gender features, 365 place-type features, and 100 w2v-cluster features. Since we consider the data from 2014 to 2019, we produce (72 x 497) matrix in each area, where 72 year-months with each row represent a 497-dimensional social feature vector.

1) Visual Features (Age-Gender, Place-Types): We first filter the posts uploaded in a target area for each time period, and collect their corresponding images. To extract the age-gender feature, we use the age-gender-detection module[5] to estimate the ages and genders of the people in an image. We classify each person into 14 age-gender groups where each age is separated by 0-10, 10-20, ..., 40-50, 50+ years old. We count the number in each age-gender category, and normalize the value by the total number of people. Second, to extract the place-type features, we use the Place365-CNN module[6] [16] with softmax regression. We sum all 365-dimensional feature vectors extracted from the images and normalize it by the number of total posts. The top 10 most frequently appearing Place365 categories in our dataset are *bakery/shop, coffee shop, delicatessen, pizzeria, ice cream parlor, restaurant, patio, beauty salon, butchers shop, and art studio*.

2) Textual Features (word-cluster): To process texts into word-cluster features, we first preprocess each text from a post into tokenized words using a Korean tokenizer [17]. Then we train Word2Vec model based on neural networks with the window size of 5 and map each word into a 100-dimensional vector. We cluster 100-dimensional word vectors into 100 clusters using k-means algorithm, and create a word-to-cluster dictionary. To produce word-cluster features corresponding to a target area and a given year-month, we filter the corresponding posts and tokenize all the sentences into words, and count the frequency for words in each cluster. We divide each cluster frequency by the number of posts, resulting in a 100-dimensional w2v-cluster feature. This feature represents the average occurrence rate of words belonging to each cluster.

B. Ranking of Social Feature Importance

Before we construct the actual inference model for commercial gentrification using Instagram data, we rank the importance of social features by the effectiveness to infer the ground truth on our dataset. We leverage SVR-RFE which is a feature elimination method that combines linear-kernel support vector regression (SVR) on a recursive feature elimination (RFE) method [18], [19]. First, we set up a pair of a 497-dimensional social feature vector and the corresponding ground truth commercial gentrification phase value for each area and each monthly timestamp. As there are 72 monthly timestamps on 10 areas, this results 720 pairs of data to be used for SVR-RFE. We repeatedly train linear SVR using this dataset. At every iteration, we eliminate the least important feature. This process repeats until only one feature is left. Then we can rank the importance of features in a reverse order of iterations. Using this feature importance rank, we select multiple sets of features from rank 1 to z, and find the optimal z for our regression models in the experiment.

[5]github.com/yu4u/age-gender-estimation
[6]https://github.com/CSAILVision/places365

C. Commercial Gentrification Phase Regression

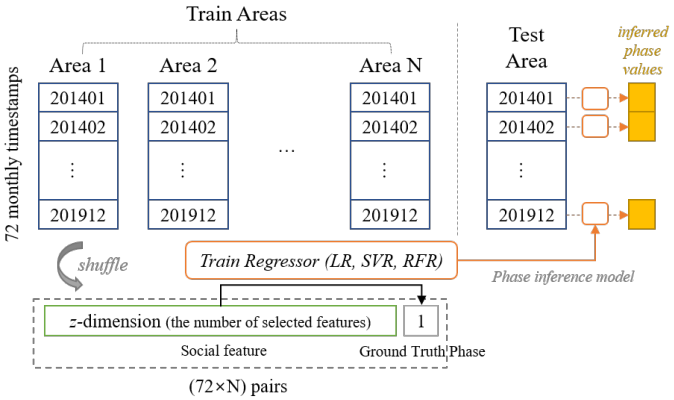


Fig. 3: The training scheme of regression models.

Our ground truth of a commercial gentrification phase value at each timestamp of an area ranges from 1 to 2.5 in real numbers. We leverage a regression approach to infer gentrification phases using the social features mentioned above from Instagram data. Fig. 3 explains the details of how we train our regression models using a social feature vector to infer the commercial gentrification phase for the test area at each timestamp. We leverage three regression methods in our experiment: Linear regression (LR), Epsilon-Support Vector Regression with RBF kernel (SVR), and Random Forest Regression (RFR). The reason for choosing these models is that we can compare how complex models are needed to solve our problem from the simplest LR model, to SVR regression which is a decision boundary based regression, and RFR which is an ensemble method comprised of multiple decision trees. Each regression model trains on pairs of social features and ground truth values collected from each timestamp on every area in the train dataset. Then, the model takes a social feature vector as an input and infers a commercial gentrification phase value as an output.

V. Experiment

We conduct three experiments as follows: 1) we evaluate whether our Instagram-based model can infer the commercial gentrification phase of a target area, closely to its corresponding ground truth; 2) we evaluate how well our model can infer how gentrification evolves in terms of time-series transition patterns, compared with the ground truth; and 3) we evaluate whether our model closely follow the geographic dispersion of gentrification, especially that of boutiques that have strongest correlation with commercial gentrification.

A. Evaluation Setup

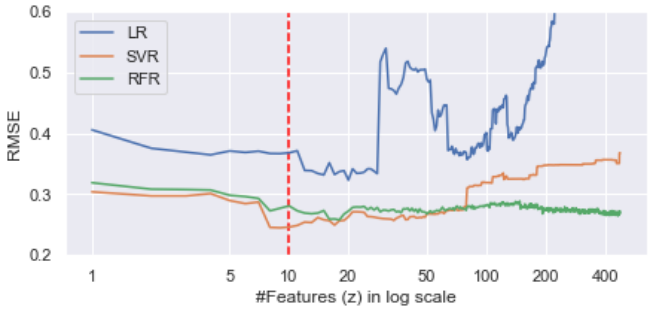
1) Commercial Gentrification Phase Identification: To show how closely our models can infer the gentrification phase of a given area, we conduct cross-regional validation and root mean squared error (RMSE). In our term, cross-regional validation means we use the data from each area as a test set while we train our model based on the data from other areas excluding the test area as described in Fig. 3. We test three different regression models (LR, SVR, RFR) on multiple sets of features from rank 1 to z to find the best inference model and the number of features required. We also evaluate RMSE for each area-group (YN, SS, IT) to validate whether the model has regional tendency.

2) Commercial Gentrification Phase Transition Patterns Estimation: To show how closely the proposed model can estimate how many phase transitions a target area has shifted until now and how long each phase lasts, we measure the Pearson Correlation between an inferred value and its corresponding ground truth, which is one of the commonly used metrics to measure the similarity between two time series data. We also conduct a qualitative evaluation of an inferred transition pattern against its corresponding ground truth.

3) Commercial Gentrification Geographical Dispersion Estimation: To see how closely the time-series geographical dispersion of Instagram post occurrence corresponds to the ground truth, we first choose the best physical variable among local, boutique, franchise store which has the highest correlation with the ground truth to compare with. We measure the mean of Pearson correlation between the ground truth and each physical features for all area - local store (0.363), boutiques (**0.824**), franchise (0.739) - and choose the location of boutique stores as the representative geographical gentrification index. We plot the heat-map of geo-tagged Instagram post and measure the ratio of the number of Instagram posts within 5 meters within any boutique stores.

B. Results

1) Commercial Gentrification Phase Identification: We first conduct an experiment to find an optimal value of z explained in Section IV-B and the regression model which performs best for our data. We choose z from 1 to 479, and make z-dimensional feature input for each regression model, and choose the best z value which shows least RMSE and best Pearson Correlation values. As shown in the Fig. 4, we choose $z = 10$ as the optimal value which fulfills such condition. This implies that it is necessary to encompass the combinations of various features such as gender, age, store type, and preference (see Table III), not just a few top features . Fig. 4 (a) also shows that SVR performs best compared with RFR and LR in terms of RMSE. The main reason for low performance of LR is because we use the normalized features in IV-A, and it is hard to find a clear linear correlation between these features with the ground truth value. Note that RMSE increases as the number of selected features are more than 50. It implies that some of the features cause bias on training data and result in bigger errors for test data, that is, overfitting.



(a) RMSE

(b) Pearson Correlation

Fig. 4: RMSE and Pearson Correlation as z. We choose $z = 10$ to be optimal number of selected features for inference.

Table II shows the mean RMSE measured for each area-group. First of all, SVR and RFR models show RMSE less than 0.3 at $z = 10$ in overall. In detail, the regression models show minimum RMSE 0.259 for YN (SVR at $z = 10$), 0.174 for GL (SVR at $z = 20$), and 0.267 for SS (SVR at $z = 10$). We ascribe that this is due to the unbalanced number of areas in the dataset, as there are 7 areas in YN, 2 areas in GL, and 1 area in SS.

TABLE II: Evaluation results (RMSE, Pearson Correlation) of prediction result of each model for each area-group.

		RMSE			Pearson Correlation		
Area	z	LR	SVR	RFR	LR	SVR	RFR
YN	raw	0.852	0.407	0.282	0.317	0.727	0.844
	100	0.461	0.358	0.299	0.618	0.666	0.853
	20	0.350	0.288	0.285	0.626	0.765	0.824
	15	0.343	0.275	0.286	0.609	0.762	0.829
	10	0.373	0.259	0.297	0.607	0.809	0.837
GL	raw	0.552	0.225	0.157	0.459	0.830	0.838
	100	0.196	0.203	0.202	0.776	0.719	0.821
	20	0.210	0.174	0.193	0.815	0.858	0.843
	15	0.260	0.190	0.196	0.813	0.882	0.858
	10	0.311	0.189	0.214	0.831	0.904	0.864
SS	raw	1.723	0.379	0.400	0.045	0.844	0.868
	100	0.395	0.280	0.363	0.727	0.916	0.843
	20	0.359	0.267	0.321	0.815	0.889	0.864
	15	0.401	0.273	0.330	0.726	0.879	0.836
	10	0.447	0.269	0.320	0.628	0.873	0.816

2) Commercial Gentrification Phase Transition Pattern Estimation: As shown in Fig. 4 (b), SVR yields the most similar inferred pattern to the ground truth than other methods.

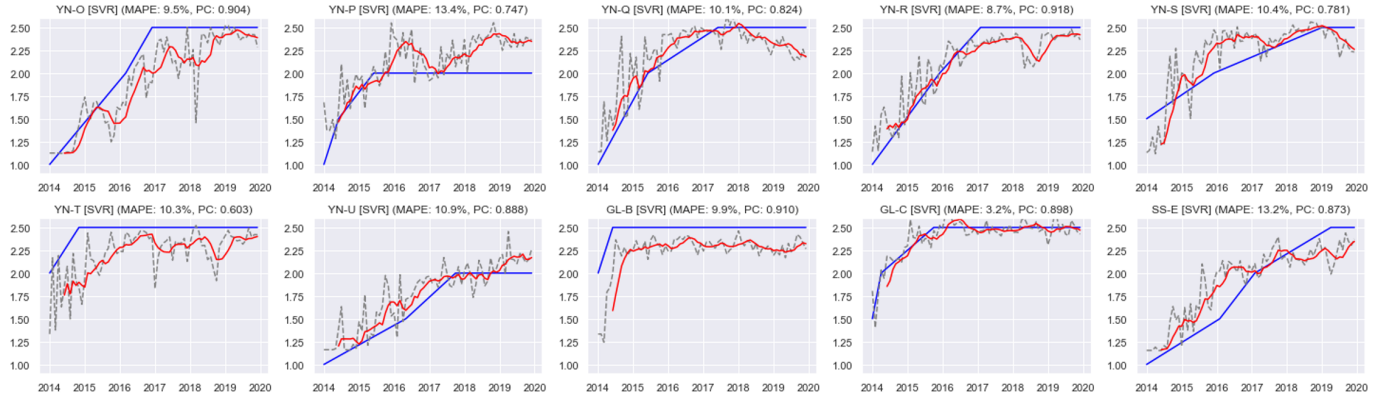


Fig. 5: Inferred transition pattern on three areas by SVR model at $z = 10$. We can compare the ground truth (blue), inferred pattern (gray), and the moving average of the inferred pattern (red).

Table II shows the regional analysis on each area-group. SVR model generally produces higher Pearson Correlation values than the other methods, while RFR is also marginally good. Thus, we choose SVR model for qualitative evaluation of the commercial gentrification pattern inference.

Fig. 5 shows the inferred pattern of commercial gentrification for each area, using SVR with $z = 10$. The blue line represents the ground truth pattern, the gray line represents its corresponding inferred pattern, and the red line represents a moving average that shows the trend. Note that inferred patterns fluctuate. It is because social features can be affected by seasonal weather or unpredictable life patterns of people. In YN-O area, the inferred pattern from SVR shows that the area's commercial gentrification drastically develops from 2014 and reaches the top level around 2017. The results are quite reasonable as the Pearson Correlation of SVR in YN-O is over 0.9 though they are not good at inferring the early stage of gentrification.

However, a few cases require additional explanations: YN-P, GL-B, and YN-T areas. The inferred pattern in YN-P reaches phase 2.5 in 2016 while the ground truth remains in the phase 2. This area is located near the central park and in the center of Yeonnam area. It incurs a relatively higher number of Instagram posts than areas with a similar physical setting but no park, which causes our trained model to infer rather more commercially gentrified than the actual value. On the other hand, the inferred pattern in GL-B area shows slower gentrification development than its corresponding ground truth. We conjecture that this area less entices young generation, the major Instagram user group because goods price in this area is relatively high. It results in the number of Instagram posts less than that of actual visitors, which creates a gap between the ground truth and its corresponding Instagram-inferred pattern. YN-T area shows similar to GL-B area but with a different reason. It has a unique characteristic compared with other areas in YN in terms of the saturation of gentrification development - phase 2.5 starts in 2015 which is earlier than the period of Instagram being popular in Korea. That is why the inferred pattern catches up its corresponding ground truth a little bit behind.

3) Commercial Gentrification Geographical Dispersion Estimation: Fig. 6 shows the proportion of Instagram posts within 5 meters of boutiques to see how much Instagram data and commercial gentrification are correlated with each other. We already show that boutiques are the most strongly correlated physical variable to the commercial gentrification based on Pearson Correlation. As such proportions range around 0.7 to 0.9 for YN, SS, and GL, we can see that the majority of geographic dispersion patterns reflect how their corresponding areas are being commercially gentrified in the real map.

We evaluate how closely the proposed method can visualize commercial gentrification development from phase 2 to phase 2.5 in terms of geographic dispersion over the time in a target area, along with the ground truth, that is, the locations of boutique stores in the area. We choose three periods for each area group as those periods are when the most drastic commercial gentrification phase transition happened in those areas. As shown in Fig. 7, the color of circle marker represents the location of boutiques open in different phase: blue represents the before the phase 2, green does the middle of phase 2 and phase 2.5, and red does the after reach the phase 2.5.

The figure clearly shows that our method can give the general trends in the geographical dispersion of commercial gentrification, in sync with the ground truth (over 70% to 90% accuracy as described above). YN-T, GL-B, and GL-C, when the color of the area block changes from green to red, area near the alleyway is being more activated. Given that commercial gentrification spreads inward from large roadsides, we read that the figure corresponds with the regular geographic pattern of commercial gentrification.

VI. Discussion

A. General Applicability of Proposed Method

We analyze top 10 most important features from SVR-RFE: *F30, F20, w2v_78, pizzeria, w2v_48, M20, w2v_69, w2v_19, candy_store, F10*, where *F-* or *M-* represent age-gender feature, *w2v* represent word2vec cluster features, and other names

TABLE III: Features in top 10 ranks from SVR-RFE (terms translated in English).

Rank	W2V_Cluster	Explanation	Rank	W2V_Cluster	Explanation
1	F30	Female 30s	6	M20	Male 20s
2	F20	Female 20s	7	w2v_69	recently, heard, want for, furthermore, like
3	w2v_78	picture, experience, art, pottery, drawing, illustrate	8	w2v_19	I, you, best, life, favor, success, artisan
4	pizzeria	Pizzeria	9	candy_store	Candy stores
5	w2v_48	Korea, Japan, Vietnam, Thailand, France, Italian, speciality store	10	F10	Female 10s

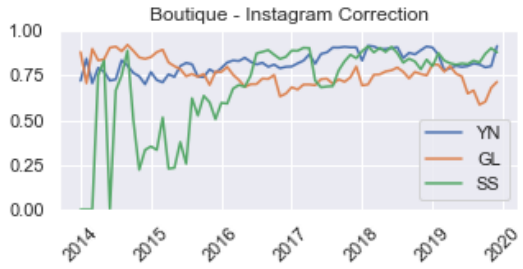


Fig. 6: The ratio of Instagram posts uploaded within 5 meters from boutiques (YN, GL, and SS from left).

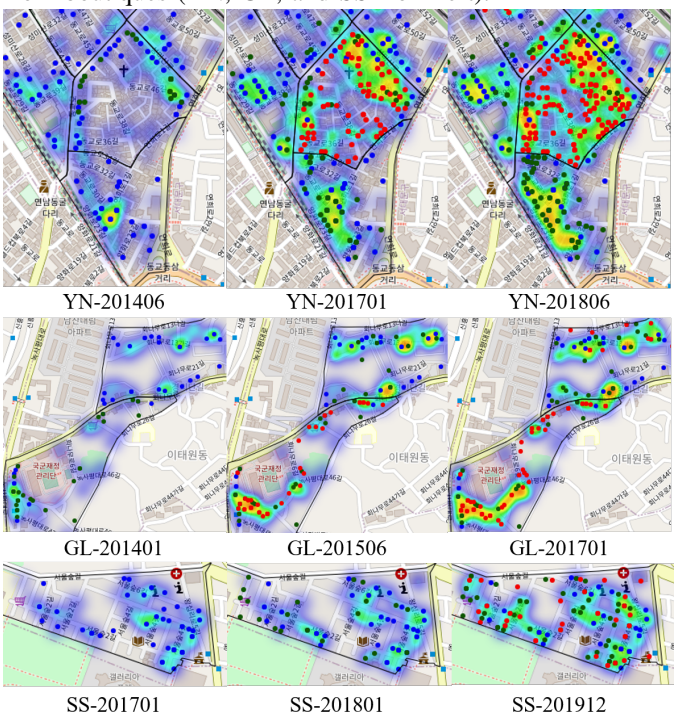


Fig. 7: Geographical dispersion of Instagram posts. Different colors of circle markers represent the boutique stores open in different phase of commercial gentrification: blue (phase $<$ 2), green (2 $\leq$ phase $<$ 2.5), and red (2.5 $\leq$ phase).

represent place-type features. Table III shows more details about the w2v-cluster features with frequent words. Given that words relevant to food and beverage stores, especially foreign-ethnic restaurants, are also frequently mentioned words in the word2vec cluster feature, the results suggest that food and beverage stores bring about commercial gentrification. It corresponds to the results of prior researches in urban studies. According to previous studies that analyze social media data using text mining, the words concerning food and beverage shops and tourism resources dominantly appear in the SNS posts in Korean gentrified areas [20], [21]. Specifically, while detailed store names are often posted on SNS in the early stage of gentrification, more abstract words like cafes and atmosphere appear on SNS as gentrification intensifies [20]. These results are not limited to Korea. As Gibbons et al. [5] argue, the keywords of social media in gentrified areas are mainly composed of restaurants and bars.

In a similar vein, the words related to arts and entertainment repeatedly appear in the word2vec cluster feature. The nouns of w2v_78 indicate the crafts and artistic activities for visitors and tourists. Several scholars in urban studies argued that commercial gentrification features the transformation into spaces for entertainment and consumption for visitors and affluent users [22]. As such, the activities of visitors and tourists are at the heart of commercial gentrification.

Lastly, F20 and F30 are selected as significant features of commercial gentrification, and the distribution plot shows the feature values are higher in a more commercially gentrified area. It reflects that the influx of the younger population is a central facet of commercial gentrification [23], [24]. According to Hardyman [23], the gentrified region become a center for young clientele living outside the area. Some studies on commercial gentrification underlined that the settlement of newly emerging commercial districts is led by groups of innovators, particularly the young urban population, who actively use SNS [25]. This would imply that our method can be applied and used for other areas in general.

B. Limitations of Proposed Method

In this study, we leverage regression models to infer the commercial gentrification phase from social features using Instagram data for each monthly timestamp. However, there are limitations in our experiment.

First of all, it is difficult to define an absolute criteria for commercial gentrification phase. This is because every area has a different physical setting and environment and corresponding gentrification development pattern varies from area to area. Therefore, there needs a more improved model for conceptualizing modern commercial gentrification.

In addition, there could be bias on using Instagram data since the users of Instagram may not represent the whole population. Therefore, the analysis of this study can be limited to the behavior of the users in the social media, and the features we find correlated to commercial gentrification may not be generalizable.

On the other hand, prediction of whether an area will be gentrified more or not can be one of the research interests for

investors or urban planners. For this, we can apply time-series prediction models such as Auto-regressive Integrated Moving Average (ARIMA), Recurrent Neural Network (RNN), or Dynamic Bayesian Network (DBN) to infer the next phase based on the history of social features and commercial gentrification phase transitions. In this case, we can also apply several methods, such as Granger causality or Bayesian Network, which can reveal the causality between social features for a deeper understanding on which features lead commercial gentrification.

VII. Conclusion

In this research, we suggest a commercial gentrification phase inference scheme using Instagram data. First, we establish ground truth based on the number and share of local, boutique, and franchise stores, then adapt linear interpolation to produce a gradual changing pattern. We extract social features such as demographics, place-types, and w2v-clusters from Instagram posts, and set up the dataset of pairs of the ground truth phase values and the corresponding social feature vectors for each area at every monthly timestamp. We train regression models to infer commercial gentrification phase using these social features, and evaluate by RMSE and Pearson Correlation through cross-regional validation.

Our results show that our model can infer the gentrification phase as well as time-series transition patterns of commercial gentrification with respect to a target. We also show that the dispersion pattern of Instagram posts corresponds with the boutique that is measured to be highest correlated to commercial gentrification. Our study implies that the potential of Instagram as a tool for identifying changes in commercial districts and the status of commercial gentrification can be recognized earlier though few studies have been facilitated [25]. We believe that our scheme could be a meaningful tool for urban planners and policymakers who are to investigate and manage commercial gentrification.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01126, Self-learning based Autonomic IoT Edge Computing). We appreciate KB Card Corp. for providing valuable credit card data.

References

[1] Lees, Loretta, Hyun Bang Shin, and Ernesto López-Morales. Planetary gentrification. John Wiley Sons, 2016.

[2] Zukin, Sharon, et al. "New retail capital and neighborhood change: Boutiques and gentrification in New York City." City Community 8.1 (2009): 47-64.

[3] Zukin, Sharon. "Reconstructing the authenticity of place." Theory and society 40.2 (2011): 161-165.

[4] Boy, John D., and Justus Uitermark. "Reassembling the city through Instagram." Transactions of the Institute of British Geographers 42.4 (2017): 612-624.

[5] Gibbons, Joseph, Atsushi Nara, and Bruce Appleyard. "Exploring the imprint of social media networks on neighborhood community through the lens of gentrification." Environment and Planning B: Urban Analytics and City Science 45.3 (2018): 470-488.

[6] Glaeser, Edward L., Hyunjin Kim, and Michael Luca. "Nowcasting gentrification: using yelp data to quantify neighborhood change." AEA Papers and Proceedings. Vol. 108. 2018.

[7] Naik, Nikhil, et al. "Streetscore-predicting the perceived safety of one million streetscapes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014.

[8] Kim, Hee Jin and Choi, Mack Joong. "Characteristics of Commercial Gentrification and Change in Perception of Placeness in Cultural Districts : The Case of Samcheong-dong and Sinsa-dong Streets in Seoul." Journal of Korea Planning Association 51.3 (2016): 97-112.

[9] Ryu, Hwa-Yeon and Park, Jin-a. "A Study on the Variation Process of Commercial Gentrification Phase in Residential Area in Seoul : Focused on Business Type of Commercial Characteristics." Journal of Korea Planning Association 54.1 (2019): 40-51.

[10] Kosta, Ervin B. "Commercial Gentrification Indexes: Using Business Directories to Map Urban Change at the Street Level." City Community 18.4 (2019): 1101-1122.

[11] Rosvall, Martin, and Carl T. Bergstrom. "Maps of information flow reveal community structure in complex networks." arXiv preprint physics.soc-ph/0707.0609 (2007).

[12] Naik, Nikhil, et al. "Computer vision uncovers predictors of physical urban change." Proceedings of the National Academy of Sciences 114.29 (2017): 7571-7576.

[13] Yoon, Yoonchae, and Jina Park. "Stage classification and characteristics analysis of commercial gentrification in Seoul." Sustainability 10.7 (2018): 2440.

[14] Seoul Metropolitan Government.Comprehensive Measures for Gentrification in Seoul. Government Document, 2015.

[15] Yoon, Yoon-chae and Park, Jin-a. "The Rate of Commercial Gentrification in Seoul focusing on Changing Type of Business" Seoul Studies 17.4 (2016): 17-31.

[16] Zhou, Bolei, et al. "Places: A 10 million image database for scene recognition." IEEE transactions on pattern analysis and machine intelligence 40.6 (2017): 1452-1464.

[17] Park, Eunjeong L., and Sungzoon Cho. "KoNLPy: Korean natural language processing in Python." Proceedings of the 26th Annual Conference on Human Cognitive Language Technology. Vol. 6. 2014.

[18] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers Electrical Engineering 40.1 (2014): 16-28.

[19] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[20] Hyeon Woo Kang and Hee Chung Lee. "A research of speed of gentrification used by text-mining analysis." Journal of Urban Policies 9.3 (2018): 71-87.

[21] Jiyoiung Shin, Suji Park, and Eunkyeong Chae. "Tourism gentrification and changes in place images: Applied text mining on the Ihwa Mural Village." International Tourism Conference 81 (2017): 70-74.

[22] Gant, Agustín Cócola. "Tourism and commercial gentrification." Proceedings of the RC21 International Conference on "The Ideal City: Between Myth and Reality, Representations, Policies, Contradictions and Challenges for Tomorrow's Urban Life", Urbino, Italy. 2015.

[23] Hardyman, Rachel Ann. "Hawthorne Boulevard: Commercial gentrification and the creation of an image." (1992).

[24] Ma, Zuopeng, et al. "The transformation of traditional commercial blocks in China: Characteristics and mechanisms of youthification." City, Culture and Society 14 (2018): 56-63.

[25] Heo, Jayun, et al. "Relationships between SNS and Vitality of Commercial Area." Journal of Tourism Management Research 18.4 (2014): 517-534.

[26] Jenkins, Porter, et al. "Unsupervised Representation Learning of Spatial Data via Multimodal Embedding." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

[27] Huang, Cheng-Chia. Developing a data mining framework to identify a sense of gentrification through social media data: A case study using Instagram posts in Salt Lake City, Utah. Diss. San Diego State University, 2017.

[28] Nishimura, Takuya, et al. "How fashionable is each street?: Quantifying road characteristics using social media." 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2016.

[29] Ye, Minting, and Igor Vojnovic. "The Diverse Role of Women in Shaping Hong Kong's Landscape of Gentrification." Urban Affairs Review 56.2 (2020): 368-414.