Forecasting commercial real estate indicators under COVID-19 by adopting human activity using social big data

Maral Taşcılar¹ · Kerem Yavuz Arslanlı²

Received: 5 April 2022 / Accepted: 27 July 2022 / Published online: 17 August 2022 © The Japan Section of the Regional Science Association International 2022

Abstract

Dependence of the real estate sector on human activity has been unveiled during the COVID-19 pandemic. In addition, it is assumed that trends emitted from the location-based social networks (LBSNs) successfully reflect human activities, hence commercial property trends. This study examined the use of social media to forecast commercial real estate figures during COVID-19 in Istanbul and determined the potential of social media data for forecasting the future rent/price levels of retail properties. Instagram and Twitter, two major LBSN platforms, were selected as social media data sources. First, 17 million geo-tagged Instagram posts and 230 thousand geo-referenced tweets were collected. Then, the data sets were superposed on COVID-19 key points in Turkey and the relationships observed. Finally, the data sets were combined with the commercial real estate data to monitor increases in the accuracy of rent and price predictions. Beşiktaş District of Istanbul was chosen as the pilot region to test the methodology. The results showed that the LBSN-supported models outperformed baseline models most of the time for price predictions and occasionally for rent predictions. Also, both Instagram and Twitter were found essential to the study and could not be omitted. This study demonstrates the significance and leveraging potential of applying human activities to the decision-making processes of the commercial real estate sector under COVID-19 conditions. This is the first study to adopt LBSN data to forecast commercial property prices.

Keywords Commercial real estate · Social big data · COVID-19 · Urban spatial analysis

JEL classification R30 · C55

Maral Taşcılar
tascilar15@itu.edu.tr

¹ Graduate School, Istanbul Technical University, Istanbul, Turkey
² Department of Urban and Regional Planning, Istanbul Technical University, Istanbul, Turkey
1 Introduction

During the COVID-19 pandemic, the *human* has become the main focus of many industries globally. Likewise, consumers’ changing habits and preferences have made it necessary for the real estate sector to go beyond traditional methods. The lockdowns, change of priorities and the evolution of how people spend time in urban areas have all decreased the assurance of real estate investments. Meanwhile, real estate research has benefited from the opportunities of modern big data in recent years. Considering these two situations, location-based social networks (LBSNs) are assumed to have a high potential to reflect human activity and, therefore, the trends in the real estate sector.

Big data is an exciting yet promising phenomenon that deals with massive and complex data stacks. There are three main aspects, Vs in particular, of big data that differs from traditional analytics: volume, it exceeds even terabytes of data; velocity, it appears as real-time or nearly real-time data; variety, it has a great variety of sources of data. For example, Internet searches, banking transactions, GPS and mobile logs, energy consumption, recorded audio or visual media, social networks, remotely sensed imagery and many other sources could contribute to the big data world. Big data analytics explore those vast quantities of data aiming to discover hidden patterns and correlations within, which reflect crucial information for companies to understand the dynamics leading to their survival in competition better (McAfee et al. 2012; Sagiroglu and Sinanc 2013).

The Institute of Real Estate Management emphasizes that on behalf of the real estate sector’s continuity, big data should be considered a catalyst for industrial growth and should be used to be a part of the competition and adapt to changing trends (Lanning 2017). Winson-Geideman and Krause state that big data applications in real estate research are quite limited; however, the availability of data and the existence of qualified researchers in data analytics promise a rapid growth process for the research area (2016). For example, some studies include the use of big data as a tool for predicting real estate prices (Zamani and Schwartz 2017; Lin and Hsu 2020; Singh et al. 2020). Other studies using “Google Trends” search volume data for both residential and commercial real estate prices and transactions, define it as a leading indicator and highlight its considerable effect on the performance of forecasting models (Dietzel et al. 2014; Wu and Brynjolfsson 2015). Researchers use remotely sensed imagery to map, simulate, or predict urban housing prices and obtain good results (Anselin and Le Gallo 2006; Yu et al. 2007; Yao et al. 2018). However, as residential data are easier to access by public channels, big data’s role in real estate research is considerably restricted with residential markets, and there is a significant absence of studies about dealing with commercial and industrial figures and the orientation of real estate investments (Winson-Geideman and Krause 2016).

Worldwide popular location-based social networks (LBSNs) are considerable sources of big data and they seem to have an excellent potential for real estate studies. Footprints of millions of users called *check-ins* shared on LBSNs, create a significant potential of demonstrating the characteristics of cities and modeling
human activity patterns, which can be an excellent proxy for the future establishment of location-based services and urban planning (Cheng et al. 2011). Despite smaller or rural settlements’ lack of sufficient data, metropolitan areas are more dynamic with adequate data to observe human activity patterns (Zhou and Zhang 2016). Thus, choosing İstanbul as the study area in this study is believed to present good results. Researchers also claim that quantification and visualization of spatial patterns of commercial establishments in metropolitan areas can be reflected from geo-locational social big data (Carpio-Pinedo and Gutiérrez 2020). Therefore, those spatial patterns are studied to identify city characteristics and track human activities (Cheng et al. 2011; Martí et al. 2017; Gupta et al. 2018). However, geo-tagged big data was used on a tiny scale directly on real estate research.

The study uses LBSN big data to obtain a forecasting method that predicts retail properties’ future rent/price levels within a specific location. It also aims to better understand how human activities both pre- and during COVID-19 conditions reflect the city’s characteristics on a temporal dimension by analyzing people’s experience of geo-tagging and integrating two different social media platforms with quite different user profiles.

Section 2 provides a literature review for traditional retail forecasting and the use LBSNs in urban studies. Section 3 first explains the collection and preparation of the data and further analyses the data and develops the forecasting methodology. Section 4 presents and interprets the results for price and rental value forecasts. Lastly, Sect. 5 concludes the study.

2 Background

Over past decades, retail properties have long been studied by a considerable number of scholars. Classical research explores the effects of location, age, tenant types, market conditions, vacancy rates, and distance to Central Business District on retail rents (Sirmans and Guidry 1993; Hui et al. 2007; Liang and Wilhelmsson 2011). Other types of variables on assessing retail property values include accessibility factors such as transportation routes, rail and highway infrastructure, rapid transit lines, skywalk expansions, and pedestrian zone schemes (Xu et al. 2016; Cohen and Brown 2017; Seo et al. 2019; Murakami et al. 2021). Some researchers use footfall (number/volume of pedestrians that pass by) as a demand-side variable to explain the retail rents and find significant positive correlations (Jeong 2015; Koster et al. 2019). Green building incentives and energy performance certificates are also considered regarding their effect on commercial property values and investments by some researchers (Fuerst and McAllister 2011; Onuoha et al. 2018). Although all mentioned valuable works contribute to real estate research to a great extent, they do not benefit from the opportunities of modern-day big data.

LBSNs are platforms where people can share location-embedded information. LBSNs blend online services with the physical world using check-ins, requiring internet-connected, location-aware mobile devices (Noë et al. 2016). Geo-tagged-media-based services are one of the sub-categories of LBSNs, where people can add
location information while sharing textual or visual media such as photos and videos (Zheng 2011; Roick and Heuser 2013). The geo-tags could be either added in-situ or later on by browsing the location. Instagram and Twitter are two primary LBSN services that enable geo-tagging by creating geo-tagged posts and tweets, respectively. However, geo-tagging is optional for both platforms and occurs as a secondary function. The semantic representation of a location is called the venue, and people tag those venues instead of using latitudes and longitudes (Chorley et al. 2015). The two LBSNs differ at the venue selection, when someone has to choose from only the venues provided by Twitter, the users can make additions to the venue database of Instagram. In addition, Twitter has been used more than Instagram for research purposes as it provides a public application programming interface (API).

LBSN based applications can be seen in many research subjects, especially in examining human mobility by tracking the geo-tagged contents on a user-oriented basis (Cheng et al. 2011; Salas-Olmedo et al. 2018). In a study where researchers collect 22 million tweets from 220,000 users, they find correlation between radius of gyration of user mobility with the city population density, city average household income, user popularity and user social status (number of followers/number of followings), which means people with higher social status and living in cities with higher geographic and economic constraints, are more active within the city (Cheng et al. 2011). Similar research examines three LBSNs—Twitter, Foursquare and Panoramio—to map tourist activity within Madrid, and the results emphasize the importance of using various data sources and suggest using the maps as an opportunity for retail investments (Salas-Olmedo et al. 2018).

Another use of LBSNs includes zoning spaces, either on classification or popularity (Cranshaw et al. 2012; Zhou and Zhang 2016; Gupta et al. 2018; Landsman et al. 2020). Zhou and Zhang (2016) mine the Twitter and Foursquare data for Boston and Chicago for 2013. They then classify individual locations with historical and real-time dynamics into travel and transport, outdoor and recreation and educational areas, and retail places like shop and service, nightlife, and food and restaurant areas. Likewise, Gupta et al. (2018) collect geo-tagged tweets spatially and temporally; they then visualize the higher activity zones, including commercial property spaces, aiming to describe the pattern of different economic activities within the city of London. Another study using 18 million Foursquare check-ins collected from Twitter, creates distinctly characterized clusters called “Livehoods” within Pittsburgh, PA, and verifies the results by interviewing city residents (Cranshaw et al. 2012).

Urban activity and space density indication of LBSNs generates the basis to our research (Agryzkov et al. 2016; Martí et al. 2017; Carpio-Pinedo and Gutiérrez 2020; Song et al. 2020). Martí et al. (2017) use Foursquare check-in data in order to illustrate the public plaza usage of people, when Agryzkov et al. (2016) compare traditional network classification algorithm with Foursquare–Twitter check-in frequency, and find correlation between two methods in analyzing public spaces (plazas), and also suggest that the urban activity within a city is related to the food-retailers and small-scale retail stores. Another study collects photographs posted on Instagram and Flickr in urban parks of Singapore and compares the photo-user-days (PUD) number of social media data with the nationwide household surveys, and
demonstrates that the number of PUD reflects the park popularity better than traditional surveys (Song et al. 2020). Another study integrates traditional retail data with modern-day social big data with the mixed use of variables which are number of cadastral commercial units or properties, total cadastral floor area, total number of registered businesses and total number of Foursquare check-ins, where all variables are found to be significant in understanding commercial environments and almost equally important so that none could be omitted (Carpio-Pinedo and Gutiérrez 2020).

There has been relatively limited research on the direct impact of LBSNs on real estate values, primarily commercial properties. However, after extensive literature research, it is found that LBSNs have more applications in demonstrating city characteristics and human activities. Here, a few rare examples in recent years are mentioned. A study using Twitter language as a residential real estate foreclosure rate and price change predictor demonstrates that Twitter model is individually as good as traditional socioeconomic (income, unemployment rate and bachelor’s degree holders) and demographic (age, sex, race, marital status and population density) variables, it also shows that forecasts become more accurate when two methods are combined (Zamani and Schwartz 2017). Another study investigates the correlation between housing prices and Twitter sentiment using 1.7 million original tweets mentioning 39 districts of İstanbul, and the results indicate that Twitter sentiment is negatively correlated. However, the percentage of check-in tweets are positively correlated with housing prices and price appreciation within the relevant districts (Hannum et al. 2019). A similar study applies Twitter sentiment to housing prices in the U.S., and conversely, finds a positive correlation between them (Tan and Guan 2021). Perhaps, the most similar research to ours is conducted by Lifang et al. (2020), where the researchers use Sina Weibo (Chinese Twitter) check-in data to map people’s activity within the city of Wuhan, and taking the outcomes as a base for the study, they investigate the spatial changes in urban rental housing prices. In this study, we aim to discover the direct relationship between commercial real estate figures and human activities which are based on geo-tagged social media content, and contribute to the real estate research by illustrating a forecasting methodology to predict future retail rents/prices in specific locations.

3 Methodology

3.1 Data collection and preparation

Data sources of the study can be divided into two: LBSN data, and real estate data. LBSN data cover the data coming from Instagram and Twitter, which are gathered by data scraping. Real estate data cover the online listings related to commercial properties within İstanbul. This section will explain the gathering process and cleaning of these data.

Python 3.8 was used as the programming language for every coding process during the study and Jupyter Notebook as the programming platform. R, a customized
statistical computing language, and R-studio programming platform were used for the following analysis and forecasting processes.

### 3.1.1 Instagram data

Instagram is a popular photo and video-sharing social media platform launched in 2010. Turkey is one of the biggest consumers of the platform with 46 million monthly active users and 2.8 h of average daily time spent on Instagram. Also, the country ranks number one with a penetration rate of 89.5%, which corresponds to the share of Instagram users among all internet users between age 16–64 (Kemp 2020). Therefore, the platform is admitted as a significant data source to observe human activity. The data from Instagram are gathered with two data-scrapers coded distinctively for this study. Therefore, the whole process can be divided into two phases: location-scraping and post-scraping.

In the first phase, the venues that geo-tagged posts will be scraped later on, are collected. Instagram provides a list of 1000 clusters, including 1000 venues under each, for every country, including Turkey. The first step was to detect the clusters located within the boundaries of İstanbul. However, those clusters do not have a spatial attribute, because they are only semantic representations. Thus, each cluster’s first-coming venue was taken instead of using those clusters directly. As the coordinates of the venues are not directly accessible, a location-scraper was coded to collect and list those venues with their latitudes and longitudes. A collection of 1000 venues with their geo-location was imported into Quantum Geographic Information System (QGIS) software, finding 196 of them located in İstanbul. Then, using the clusters of those 196 venues, a total number of 171,578 venues were listed. The location-scraper was used again, to obtain venues’ coordinates. In the end, a list of 171,578 venues with their latitudes and longitudes.

First, during the cleaning process of venue data, 2755 duplicates were removed, as a venue could have been listed under two or more clusters. Next, 11,970 venues were removed from the remaining 168,823, because they had a total post amount of zero. This no-post situation was thought to have two reasons; either content was removed after posting or still exist but belonged to a private account. Then, 156,853 venues left were transferred to QGIS, filtering those that fall into land frontiers. Excluded 3281 venues were located within the Marmara Sea, possibly created from watercraft. Finally, a manual cleaning procedure was applied, omitting venues such as; country, city-named venues, as the exact location of posts created from is not detectable; misplaced venues which come from outside of İstanbul; venues named of four biggest stadiums which have huge amounts of biweekly repeating posts; and lastly irrelevant venues which have no geo-locational meaning (e.g., “in my heart”, “somewhere on Earth”). By the end of phase I, 152,145 venues were remaining to collect the posts from.

At the beginning of phase II, another scraper, named post-scraper, was coded to collect and list posts geo-tagged from corresponding venues. A total of 17,161,015 posts posted from İstanbul were collected for the time frame June 2019–August 2021. The venues’ post counts were joined by the location of districts. Then, a cumulative density map is created using the 2-year total post counts of venues.
within districts. Figure 1 depicts the city-scale map and the heat map distinctively created for the Beşiktaş district.

3.1.2 Twitter data

Twitter is a micro-blogging platform where users share their personal opinion via tweets. Tweets contain texts limited to 280 characters, which also enable adding photos, links or geo-tags. According to the report Digital 2020: Turkey, the country has 13.6 million monthly active users and 1.1 h of average daily time spent on Twitter. Also, the country has a penetration rate of 72.5%, taking place in the top 5 countries globally (Kemp 2020). Although, the platform stands behind Instagram in terms of geo-tagging, it is still included in the research as it may reflect different users’ activities.

Twitter provides a public API for the developers. However, the free package limits the collection of tweets within the past seven days. Meantime, other scraping methods do not return the geo-locational information of tweets. As a solution, the data gathering process of Twitter is also divided into two phases.

In the first phase, the venues that geo-tagged tweets will be scraped later on, are collected. As, Twitter removed the precise location-sharing feature in June 2019, users can now only geo-tag their tweets using the venues provided. The venue database of Twitter is powered by Foursquare, which is an independent location data platform. Using Foursquare’s Places API, which offers access to a rich global venue database and Postman collection tool, a list of 39,403 venues under various categories and their latitudes and longitudes was acquired.

Fig. 1 Distribution of total Instagram posts
During the cleaning process of venue data, 17,223 duplicates were removed, as a venue could have been listed under two or more categories. Then, 22,180 venues left were transferred to QGIS, filtering those that fall into land frontiers. By the end of phase I, 21,875 unique venues for Twitter were remaining to collect the tweets from.

In the phase II, an open-source tweet-scraper, called snscrape (2021), was used to collect and list tweets geo-tagged from corresponding venues. A total of 227,965 tweets posted from Istanbul were collected for the time frame June 2019–August 2021. The venues’ tweet counts were joined by the location of districts. Then, a cumulative density map is created using the 2-year total tweet counts of venues within districts. Figure 2 depicts the city-scale map and the heat map distinctively created for the Beşiktaş district.

After acquiring both Instagram and Twitter data, the trend graphs of Beşiktaş district are superposed with COVID-19 key points of Turkey as given in Fig. 3. The most significant outcome of this figure is that while the lockdown periods are ending, the social media activities had already started to increase before the government announced the beginning of the official normalization period. This is clear evidence to how well social media data reflect the trend of human activity. Another observed situation is that Twitter check-in activity experienced a sharper decline than Instagram when the first COVID-19 case was seen, and it still could not catch up with its old levels during the monitored period. On the contrary, Instagram post numbers recovered after the first wave and rose above the pre-pandemic levels, then fluctuated in line with the course of the pandemic, and remained at lower levels in the summer season due to seasonal migrations to vacation regions.

![Fig. 2 Distribution of total tweets](image)
The third data set is directly related to commercial real estate to depict social media’s effect on property values. Due to the lack of existing public commercial real estate data in Turkey, online retail property listings have been used for the study. The data set is provided by Emlakjet, which is an online real estate marketplace with 5 million monthly users (Emlakjet, n.d.).

In this study, Beşiktaş district, which can be attributed as the heart of commercial activity in Istanbul, was chosen as the pilot region. The raw data set provided by Emlakjet covers the retail listings for sale and rent in Beşiktaş district for June 2019–August 2021. However, the distribution of the number of listings within the selected timeframe shows that there are several periods where the data are problematic. Therefore, as the low number of listings in a particular period may mislead the rent/price levels of the market, those periods are removed from the analysis. The remaining time frame covers March 30, 2019–June 13, 2021.

As these listings could be freely entered into the platform by the landlords/real estate agents and could not be verified by a competent authority, there were extreme outliers in the data. These outliers mainly were related to the wrongly selected rent or sale options, or perfunctory listings with nonsense figures.

Fig. 3 COVID-19 timeline in Turkey on weekly trend graphs of total number of posts and tweets from Beşiktaş district

### 3.1.3 Real estate data
Based upon unit rents/prices per sq m, both sale and rental listings were cleaned from outliers separately. The threshold point is determined where z-score equals 2 and the 95.45% of the data which remains below this threshold was included in the analysis.

### 3.1.4 Final data set

Although the three data sets obtained have a daily frequency, the real estate data are not sufficient to make a study with a daily frequency in terms of the number of listings. In any case, the nature of the commercial property market does not require monitoring daily fluctuations.

The research carried out in this field to date has been forecasting long-term periods because the data/indices related to the sector mostly have monthly and quarterly frequencies. In this study, it will be possible to make forecasts for one or more weeks by shortening the forecast horizon, taking into account the convenience of the data sets. For this reason, all three datasets were grouped in 7-day periods using Post Date, Tweet Date, and created_at attributes of the three data sets. Although Instagram and Twitter data have a time frame of 116 weeks, an analysis of 63 weeks was possible due to the availability of real estate data in a more limited range. Previously, 1,369,577 Instagram posts and 21,224 tweets were collected which belong to venues of Beşiktaş district. However, due to the time frame restriction, 730,039 Instagram posts and 5,668 tweets were grouped in 63 weeks and used in the study.

Meantime, although specific periods were excluded from the analysis initially, some missing data were found when the real estate data were grouped weekly. These few empty weeks have been filled using linear interpolation.

The total numbers of posts and tweets obtained from Instagram and Twitter and grouped weekly, entered the final data set as n_insta and n_tweet. From the real estate data, av_price and av_rent variables, the average of unit sales prices/rents per sq m of the listings entered per week, were added to the data set. In addition, the number of listings created for Beşiktaş every week was added to the data set as n_sales and n_rental. Finally, an attribute called last_seen is provided only for the listings for sale. Subtracting created_at attribute from last_seen, the variable av_days, which reflects the average time on market of the for-sale retail listings created per week, is obtained. Table 1 exhibits the variables used in the study with their relevant descriptive statistics.

### 3.2 Analysis

The ARIMA (Autoregressive Integrated Moving Average) has three components, where auto-regressive part indicates that the variable is regressed on its past values, the integrated part indicates the stationarity of the data in terms of the need for differencing, and the moving average part indicates the dependency between an observed value and a residual error from the previous observations (Hyndman and Athanasopoulos 2018). The model is defined by three terms ($p$, $d$, $q$), which are usually very small integers (e.g., 0, 1, or 2) and these terms model the
patterns in the time series data (Tabachnick et al. 2007). Meantime, the nature of this study requires multivariate time series analysis. In multivariate time series models, there are several time-dependent variables that depend not only on their past values but are also affected by others’. There are different methods for such analysis in the literature, and the use of the R package *marima* (Multivariate ARIMA and ARIMA-X Analysis) is suggested for analysis and prediction in this study (Spliid 2017).

As described by Spliid, the creator of MARIMA, the *marima* package is based on their study on a fast estimation method for VARMA models with exogenous variables (1983, 2016). In the most basic model, a k-variate random vector of observations, \( y_t \), and a k-variate random vector of unknown iterations, \( u_t \), exist, where \( t \) denotes time.

\[
y_t = \begin{pmatrix} y_{t,1} \\ y_{t,2} \\ \vdots \\ y_{k,t} \end{pmatrix}, \quad \text{and} \quad u_t = \begin{pmatrix} u_{t,1} \\ u_{t,2} \\ \vdots \\ u_{k,t} \end{pmatrix}, \quad t = \{1, 2, \ldots, N\}. \tag{1}
\]

The model generates a random vector \( y_t \) when the coefficient matrices \( \varphi_1, \ldots, \varphi_p \) and \( \theta_1, \ldots, \theta_q \) all are of dimension \( k \times k \).

\[
y_t + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}. \tag{2}
\]

While the general opinion is that the series of unknown iterations, \( u_t \), is without autocorrelation, the individual coordinates might be correlated. When the exogenous variables are included in the model, as in this study, where \( y_t \) stands for (multivariate) observations, \( x_t \) represents the (multivariate) \( x \)-variables.

\[
y_t + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + x_t + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}. \tag{3}
\]

The reasons for adopting the MARIMA method are that it is very efficient during the model identification phase as various alternative models can be tested with fast results, and it provides a user-friendly interface while formulating the models (Spliid 1986; Kenabatho et al. 2015; Rinchen et al. 2018).

### Table 1 Descriptive statistics

| Variable | N  | Mean | St. Dev | Min   | Q1   | Median | Q3   | Max  |
|----------|----|------|---------|-------|------|--------|------|------|
| av_rent  | 63 | 100.82 | 16.89   | 67.99 | 90.08 | 102.83 | 113.77 | 132.83 |
| av_price | 63 | 25,902.76 | 6,203.79 | 12,166.67 | 22,467.29 | 25,989.90 | 31,305.11 | 40,737.77 |
| n_rental | 63 | 14.25 | 6.59 | 2.00 | 10.00 | 14.00 | 19.00 | 38.00 |
| n_sales  | 63 | 5.54 | 2.83 | 1.00 | 4.00 | 5.00 | 7.00 | 15.00 |
| av_days  | 63 | 94.21 | 55.10 | 12.00 | 54.38 | 80.00 | 129.38 | 237.00 |
| n_insta  | 63 | 11,587.92 | 2,966.14 | 5,273.00 | 9,978.50 | 12,203.00 | 13,680.00 | 15,625.00 |
| n_tweet  | 63 | 89.97 | 44.37 | 31.00 | 49.00 | 91.00 | 123.50 | 200.00 |
To observe the characteristics of the variables to be used in the MARIMA model, first, auto.arima() function of R package *forecast* is applied to all variables of the series separately. Akaike’s Information Criterion (AIC), which is helpful in determining the order of an ARIMA model is adopted while selecting the parameters (Akaike 1974).

However, the existence of different best orders for all variables prevented the determination of a unified order for the MARIMA model. For this reason, the list of ARIMA models considered using the *trace* argument of the auto.arima() function were reported. Table 2 shows the different order options obtained from these lists which minimize the AIC for each variable.

While the *p* and *q* parameters in this list will form the basis for the order selection of the MARIMA model in the next stages, the *d* parameters indicate the need for differencing of the variables. Therefore, first differencing is applied to *n_sales*, *av_days*, *n_insta*, and *n_tweet* parameters and I(0) processes are reached for all variables. Also, the total number of observations were reduced to 62. To ensure stationarity, Augmented Dickey–Fuller (ADF), Phillips–Perron (PP), and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root tests were also applied to the variables (Dickey and Fuller 1979; Phillips and Perron 1988; Kwiatkowski et al. 1992). As given on Table 3, the test results suggested that the time series are stationary.

The Granger causality test has also been conducted between the time series and the results are given on Table 4. The Granger causality test determines whether a time series has a statistically significant effect in forecasting another time series (Granger 1969). To avoid any spurious regressions in the test, it is suggested to use pairs of time series with I(0) processes, which has already been achieved in the former steps. During the lag selection stage of the Granger causality test, a practical issue of working with weekly data has appeared. As Hyndman and Athanasopoulos (2018) stated, it is hard to work with weekly data in time series forecasting as there are 52.18 weeks within a year, which is a non-integer and large number. Most of the methods fail when using such a significant period even though an approximation with 52 can be made. To overcome this situation, the lag order is assumed to be 13, approximately as a period of one quarter of the year.

The results show that, *av_rent* and *av_price* variables can explain past values of each other strongly. Also, the potential exogenous variables generally demonstrate high power in explaining both price and rent values. In the light of this information,

| Variable   | *p* (AR) | *d* (I) | *q* (MA) |
|------------|----------|---------|----------|
| av_rent    | 0/1/2    | 0       | 1/2      |
| av_price   | 0/1/2    | 0       | 0/1/2    |
| n_rental   | 0/1/2    | 0       | 0/1      |
| n_sales    | 0/1      | 1       | 1/2      |
| av_days    | 0/1/2    | 1       | 1/2      |
| n_insta    | 0/1      | 1       | 0/1      |
| n_tweet    | 0/1      | 1       | 0/1      |
it has been deemed appropriate that the models make rents and prices dependent on each other, and that the mentioned exogenous variables are included in the models.

### 3.3 Forecasting methodology

The methodology proposed in this study, rather than offering a single forecasting model, creates different models using the existing real-world data, finds the one with the highest accuracy among these models, and uses the selected model to make forecasts in the desired forecast horizon.

This methodology consists of a series of rental and price values models, which begins with the most basic model and gets ever more complex. These models, which are collected under three main categories, are divided into sub-categories according to the dimension of the model and the exogenous variables it contains. Model A uses univariate ARIMA, which is the purest time series forecasting model, taking rents or prices as dependent variables. Model A is the only one-dimensional model tested, while one of rent or sales values are included as dependent variables. In all other models, both are included as dependent variables together.

---

### Table 3  Unit root tests results

| Variable | Levels | 1st differences |
|----------|--------|-----------------|
|          | ADF    | PP   | KPSS | ADF    | PP   | KPSS |
| av_rent  | -6.8833** (0.0000) | -7.1710** (0.0000) | 0.1956** | -4.6510** (0.0001) | -30.5240** (0.0000) | 0.1256** |
| av_price | -3.6716** (0.0045) | -6.1310** (0.0000) | 0.4226** | -2.7351 (0.0682) | -19.4002** (0.0000) | 0.1962** |
| n_rental | -5.5604** (0.0000) | -5.2671** (0.0000) | 0.0789** | -11.4691** (0.0000) | -18.5940** (0.0000) | 0.0540** |
| n_sales  | -4.7602** (0.0001) | -4.7362** (0.0000) | 0.5197 | -8.7040** (0.0000) | -18.5940** (0.0000) | 0.3684** |
| av_days  | -5.1457** (0.0000) | -4.8871** (0.0000) | 0.4624* | -6.8739** (0.0000) | -17.4501** (0.0000) | 0.2082** |
| n_post   | -2.1619 (0.2204) | -2.1680 (0.2180) | 0.7972 | -8.0209** (0.0000) | -8.0201** (0.0000) | 0.2819** |
| n_tweet  | -1.9326 (0.3169) | -2.2600 (0.1851) | 0.2740** | -9.6361** (0.0000) | -9.6664** (0.0000) | 0.1367** |

* and ** indicate significance at %5 and %1, respectively

---

### Table 4  Granger causality test results

|          | av_rent | av_price | n_rental | Δn_sales | Δav_days | Δn_insta | Δn_tweet |
|----------|---------|----------|----------|----------|----------|----------|----------|
| av_rent  | n/a     | 0.000*** | 0.000*** | 0.000*** | 0.017**  | 0.004*** | 0.001*** |
| av_price | 0.004*** | n/a      | 0.298    | 0.018**  | 0.000*** | 0.000*** | 0.016**  |

Δ indicates first differencing. *, ** and *** indicate significance at %10, %5 and %1, respectively
In A model, which is the only one-dimensional model tested, only one of the rent or sale values is included as the dependent variable, while in all other models, both are added to the model together. B model, which is the simplest of the two-dimensional models, creates a forecast model by making rent and price values dependent on each other. On the other hand, C models consist of a total of seven models in three different categories. These models investigate the effects of different variables to forecast accuracy, apart from the past values of the dependent variables. For example, the C1 model takes \( n_{\text{rental}}, n_{\text{sales}} \) and \( \text{av\_days} \) as exogenous variables, provided by real estate data.

All the models mentioned so far are among the models that can be attributed as traditional estimation methods for commercial real estate values. Many alternative baseline models are considered instead of creating a single one, to present the impact of social media data on forecasting accuracy transparently.

The remaining C2 and C3 are the models which include the social media data. In C2 models, only social media data are the exogenous variables, while C3 models also include the variables from the real estate data. In these models, sub-experiments are also made in which Instagram and Twitter data are either unaccompanied or coexisted. The model categories are shown in Table 5.

After the variables to be included in the models were specified, it was necessary to determine the \( p \) and \( q \) orders that the optimum model could have. For this reason, it was ensured that the methodology creates \( p \) and \( q \) pairs from the list \{0, 1, 2, 1–2\} and runs each model 15 times (except for \( p=0 \land q=0 \)). In this way, 135 models with different orders and variables are created once the process is run.

While the data points in the training subset (in-sample data) are used during the model-building stage, forecast accuracies are measured using the data points in the test subset (out-of-sample data).

Meantime, the method allows selecting the optimum model for the desired forecast horizon by separating the dataset into different train-test subsets. For example, the test period could be determined as three weeks and the model that gives the best result in this process can be used for the prediction when it is desired to make a forecast for a three-week period.

Figure 4 describes the flowchart of data analysis and forecasting stages of the proposed methodology in this study.

| Table 5 | Forecasting model variations |
|---------|-----------------------------|
| Model   | Model dimension | Exogenous variables       |
|         |                 | \( n_{\text{rental}}, n_{\text{sales}}, \text{av\_days} \) | \( n_{\text{insta}} \) AND/OR \( n_{\text{tweet}} \) |
| A       | Univariate      | –                         | –                          |
| B       | Multivariate    | –                         | –                          |
| C1      | Multivariate    | X                         | –                          |
| C2      | Multivariate    | –                         | X                          |
| C3      | Multivariate    | X                         | X                          |
Section 4 compares forecasting results of the baseline and LBSN-supported models.

4 Results

The performances of the forecasting models are quantified by selecting suitable error metrics. Brooks and Tsolacos (2010) state that a standard metric for the forecast evaluation is the mean absolute error (MAE), which is the average of the differences between the actual and forecasted values in absolute terms, however, while forecasting actual prices/rents instead of price/rent growths, the use of mean absolute percentage error (MAPE) is more sufficient. They also argue that another metric called mean square error (MSE), which measures the average of the squared differences between the actual and forecasted values, can also be produced. However, the root mean square error (RMSE) is more straightforward to comprehend due to its similar scale with the dependent variable. As a result, the forecasting accuracy is determined by reporting MAPE and RMSE metrics, where smaller figures for both indicate more accurate forecast results. The best results from the baseline models (A, B and C1) and from the LBSN-supported model (C2 and C3) are presented for each session.

Lastly, the MAPE and RMSE improvements are calculated to observe the performance of the best LBSN-supported model against the best baseline model. Those figures demonstrate whether the social media data increases the explanatory power of the models and forecasting accuracies.

Fig. 4 Flowchart of the study
All models were run for test periods of one week to ten weeks to showcase results in different forecast horizons. Results from all periods are given in Table 6 for both sales and rental value forecasts.

When the sales forecasts are examined, the main finding is that the LBSN-supported models consistently outperform the baseline models for MAPE and RMSE values. For estimates of up to four weeks, improvements in both metrics ranged from 41 to 74 per cent.

Although it has decreased below 24% in longer-term forecasts, the improvement in MAPE values has always been positive. On the other hand, there is an increase in RMSE values for forecast horizons more than four weeks. Although RMSE improvements reappear when the forecast horizon is extended a little more, it remains low compared to short-term forecasts.

RMSE takes the square of the errors and therefore imposes a greater penalty on data that is exposed to a high number of turning points. On the other hand, MAPE is more robust to outliers in the data. While there are always improvements in MAPE values with social media data, the loss of power in RMSE improvement after a point can be explained by an extreme value of the real-world data at that time.

Looking at the rental value forecasts, there is no consistent evidence that the LBSN-supported models outperform the baseline models. The best result appears at the one week ahead forecast, where both MAPE and RMSE values improve by 37.4%. However, during different forecast horizons, baseline models stand out from time to time, while LBSN-supported models improve estimations in some cases. A possible explanation could be that the rental market of Beşiktaş district is quite well-settled, so it does not necessitate complex forecasting algorithms.

It has been observed that the MAPE values of the baseline models are below the 8% in the forecasts made up to five weeks. This means that in a forecast horizon that can be considered not so short, predictions can already be made with a very low margin of error using only real estate data.

Other findings from these results further investigate the breakdowns of social media data. For prices, the best LBSN-supported model always includes Twitter data, either alone or with Instagram data. When the results are examined in detail, it is seen that the social media-supported model that causes the highest RMSE improvement mostly contains Instagram and Twitter data together. Looking at the rent values, it is seen that there is no consistency in the inclusion of Twitter and/or Instagram as the best explanatory social media data source. As a result, it is not concluded that Twitter and Instagram data can be clearly superior to each other and the existence of both data sources is essential for the creation of successful forecasting models.

5 Conclusion

To the authors’ knowledge, this research demonstrates the first attempt to investigate the direct impact of LBSNs on the commercial real estate sector. Existing research, which is mainly focused on residential properties, offers that social media has great potential to observe human activity within cities. Regarding
| Training subset (weeks) | Forecast horizon (week[s]) | MAPE Baseline | MAPE LBSN-supported | MAPE Improvement (%) | RMSE Baseline | RMSE LBSN-supported | RMSE Improvement (%) | LBSN Included in the Best Model |
|-------------------------|-----------------------------|---------------|---------------------|----------------------|---------------|----------------------|----------------------|-------------------------------|
| 61                      | 1                           | 0.009         | 0.002               | 73.4                 | 284           | 76                   | 73.4                 | Instagram and Twitter          |
| 60                      | 2                           | 0.042         | 0.018               | 56.0                 | 1421          | 661                  | 53.5                 | Twitter                        |
| 59                      | 3                           | 0.141         | 0.069               | 50.9                 | 4523          | 2670                 | 41.0                 | Twitter                        |
| 58                      | 4                           | 0.117         | 0.057               | 51.4                 | 4364          | 2152                 | 50.7                 | Instagram and Twitter          |
| 57                      | 5                           | 0.113         | 0.089               | 21.2                 | 3922          | 4559                 | 16.2                 | Twitter                        |
| 56                      | 6                           | 0.106         | 0.087               | 18.2                 | 4190          | 4492                 | 7.2                  | Twitter                        |
| 55                      | 7                           | 0.107         | 0.085               | 20.0                 | 3898          | 4007                 | 2.8                  | Twitter                        |
| 54                      | 8                           | 0.119         | 0.091               | 23.5                 | 4273          | 4056                 | 5.1                  | Instagram and Twitter          |
| 53                      | 9                           | 0.113         | 0.094               | 17.1                 | 4047          | 3870                 | 4.4                  | Instagram and Twitter          |
| 52                      | 10                          | 0.110         | 0.096               | 12.9                 | 4080          | 3764                 | 7.7                  | Instagram and Twitter          |

| Training subset (weeks) | Forecast horizon (week[s]) | MAPE Baseline | MAPE LBSN-supported | MAPE Improvement (%) | RMSE Baseline | RMSE LBSN-supported | RMSE Improvement (%) | LBSN Included in the Best Model |
|-------------------------|-----------------------------|---------------|---------------------|----------------------|---------------|----------------------|----------------------|-------------------------------|
| 61                      | 1                           | 0.041         | 0.026               | 37.4                 | 3.9           | 2.4                  | 37.4                 | Instagram                     |
| 60                      | 2                           | 0.024         | 0.024               | 0.0                  | 3.0           | 3.0                  | 0.0                  | Twitter                       |
| 59                      | 3                           | 0.064         | 0.071               | −11.3                | 12.2          | 13.9                 | −13.6                | Twitter                       |
| 58                      | 4                           | 0.080         | 0.080               | 0.1                  | 12.3          | 9.7                  | 21.4                 | Instagram and Twitter         |
| 57                      | 5                           | 0.072         | 0.068               | 6.6                  | 10.8          | 10.2                 | 5.8                  | Twitter                       |
| 56                      | 6                           | 0.087         | 0.080               | 7.7                  | 13.2          | 12.3                 | 6.6                  | Twitter                       |
| Training Subset (weeks) | Forecast Horizon (week[s]) | MAPE Improvement (%) | MAPE | RMSE Improvement (%) | RMSE | LBSN Included in the Best Model |
|-------------------------|-----------------------------|----------------------|------|----------------------|------|-------------------------------|
|                         |                             | Baseline | LBSN-supported | Baseline | LBSN-supported |                  |                  |
| 55                      | 7                           | 0.133     | 0.136          | –2.2  | 16.6           | 17.6       | –6.1             | Instagram        |
| 54                      | 8                           | 0.151     | 0.153          | –1.4  | 21.0           | 20.6       | 2.0              | Twitter          |
| 53                      | 9                           | 0.141     | 0.118          | 16.0  | 21.2           | 15.6       | 26.3             | Instagram        |
| 52                      | 10                          | 0.133     | 0.111          | 16.4  | 18.8           | 15.3       | 18.5             | Instagram        |
that the human preferences are in the foreground during COVID-19 conditions, these activities are claimed to reflect the demand of users and further the real estate figures, which mainly feed on human behavior. Considering the promising outcomes of these studies, it is foreseen that the LBSN data could be used as a tool to explain the trend of demand for commercial environments within the city, taking into account that the practice of geo-tagging mostly occurs where people engage in commercial activities.

Motivated by the fact that, Turkey has pretty high rankings in the share of users for both Instagram and Twitter among all internet users aged 16–64, these platforms were expected to be reasonably indicative. However, due to the lack/limitations of the public API services provided by these platforms, the data gathering process has built a critical challenge, which has been overcome by coding several specific programs for this study.

The proposed methodology in this paper is constructed by creating baseline models with the variations of real estate related variables, and introducing the social media related variables upon them. The improvements in models that use the social media data over the baseline models are presented using appropriate metrics for sales and rental value forecasts. Trials using different forecast horizons from one to ten weeks are reported. For prices, the results indicate that LBSN-supported models outperform the baseline models most of the time. For rents, it has been observed that while sometimes baseline models come to the fore in forecast accuracies, the LBSN-supported model outperforms the baseline model in some forecast horizons. Another outcome of the study is that both Instagram and Twitter are essential to the methodology and cannot be chosen over each other.

The study’s main limitation is that Turkey lacks decent figures and/or indices specified for the commercial real estate sector. As a result, online retail property listings are used that only cover the asking prices/rents rather than real transactions. Also, the availability of the real estate data in a narrower time frame than the initially targeted resulted in a decretion of the data points.

This study demonstrates the applicability of big data arising from LBSN platforms to commercial real estate markets. The methodology allows the forecasters to better predict the trends in the sector and adapt to potential changes in the market during the volatile nature of COVID-19 conditions. The adoption of proactive strategies would eventually result in improvement in the investment decisions of the professionals.

This study tests the proposed methodology for Beşiktaş, which is a notably commercially vibrant district of Istanbul. Further research could extend the application of this methodology to other districts to examine the explanatory power of social media data on commercial real estate trends.

**Funding** This work was supported by Research Fund of the Istanbul Technical University. Project Number: 43886.
Declarations

Conflict of interest No potential conflict of interest was reported by the authors.

References

Agryzkov T, Martí P, Nolasco-Cirugeda A et al (2016) Analysing successful public spaces in an urban street network using data from the social networks Foursquare and Twitter. Appl Netw Sci 1:1–15. https://doi.org/10.1007/s41109-016-0014-z

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Anselin L, Le Gallo J (2006) Interpolation of air quality measures in hedonic house price models: spatial aspects. Spat Econ Anal 1:31–52

Brooks C, Tsolacos S (2010) Real estate modelling and forecasting. Cambridge University Press

Carpio-Pinedo J, Gutiérrez J (2020) Consumption and symbolic capital in the metropolitan space: Integrating ‘old’ retail data sources with social big data. Cities 106:102859. https://doi.org/10.1016/j.cities.2020.102859

Cheng Z, Caverlee J, Lee K, Sui D (2011) Exploring millions of footprints in location sharing services. In: Proceedings of the International AAAI Conference on Web and Social Media

Chorley MJ, Whitaker RM, Allen SM (2015) Personality and location-based social networks. Comput Hum Behav 46:45–56. https://doi.org/10.1016/j.chb.2014.12.038

Cohen JP, Brown M (2017) Does a new rail rapid transit line announcement affect various commercial property prices differently? Reg Sci Urban Econ 66:74–90. https://doi.org/10.1016/j.regsciurbeco.2017.05.006

Cranshaw J, Schwartz R, Hong J, Sadeh N (2012) The livehoods project: Utilizing social media to understand the dynamics of a city. In: Proceedings of the International AAAI Conference on Web and Social Media

Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74:427–431

Dietzel MA, Braun N, Schäfers W (2014) Sentiment-based commercial real estate forecasting with Google search volume data. J Prop Invest Finance 32:540–569. https://doi.org/10.1108/JPIF-01-2014-0004

Emlakjet (2022) About Emlakjet. https://www.emlakjet.com/hakkimizda/. Accessed 30 Mar 2022

Fuerst F, McAllister P (2011) The impact of energy performance certificates on the rental and capital values of commercial property assets. Energy Policy 39:6608–6614. https://doi.org/10.1016/j.enpol.2011.08.005

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econ J Econ Soc 2:424–438

Gupta N, Crosby H, Purser D, et al (2018) Twitter Usage Across Industry: A Spatiotemporal Analysis. In: 2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, Bamberg, pp 64–71

Hannum C, Arslanli KY, Kalay AF (2019) Spatial analysis of Twitter sentiment and district-level housing prices. J Eur Real Estate Res 12:173–189. https://doi.org/10.1080/02626667.2019.1657230

Hui ECM, Yiu CY, Yau Y (2007) Retail properties in Hong Kong: a rental analysis. J Prop Invest Finance 25:136–146. https://doi.org/10.1108/14635780710733825

Hyndman RJ, Athanasopoulos G (2018) Forecasting: principles and practice. OTexts

Jeong S-Y (2015) Assessing the impact of pedestrian traffic volumes on locational goodwill. J Cadastre Land InformatiX 45:225–240

Kemp S (2020) Digital 2020: October Global Statshot—DataReportal–Global Digital Insights. https://datareportal.com/reports/digital-2020-october-global-statshot. Accessed 19 Jun 2021

Kenabatho PK, Parida BP, Moalafhi DB, Segosebe T (2015) Analysis of rainfall and large-scale predictors using a stochastic model and artificial neural network for hydrological applications in southern Africa. Hydrol Sci J. https://doi.org/10.1080/02626667.2015.1040021

Koster HRA, Pasidis I, van Ommeren J (2019) Shopping externalities and retail concentration: evidence from dutch shopping streets. J Urban Econ 114:103194. https://doi.org/10.1016/j.jue.2019.103194
Kwiatkowski D, Phillips PC, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J Econom 54:159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Landsman D, Kats P, Nenko A, Sobolevsky S (2020) Zoning of St. Petersburg through the prism of social activity networks. Proc. Comput Sci 178:125–133. https://doi.org/10.1016/j.procs.2020.11.014

Lanning M (2017) Big data is the property manager’s friend. Natl Real Estate Invest.

Liang J, Wilhelmsson M (2011) The value of retail rents with regression models: a case study of Shanghai. J Prop Invest Finance 29:630–643. https://doi.org/10.1108/14635781111171788

Lifang Z, Ting Y, Yang L, Li Z (2020) Analyses on the Spatial Distribution Characteristics of Urban Rental Housing Supply and Demand Hotspots Based on Social Media Data. IEEE, pp 126–130

Lin TC, Hsu S-H (2020) Forecasting housing markets from number of visits to actual price registration system. Int Real Estate Rev 23:1131–1162

Martí P, Serrano-Estrada L, Nolasco-Cirugeda A (2017) Using locative social media and urban cartographies to identify and locate successful urban plazas. Cities 64:66–78. https://doi.org/10.1016/j.cities.2017.02.007

McAfee A, Brynjolfsson E, Davenport TH et al (2012) Big data: the management revolution. Harv Bus Rev 90:60–68

Murakami J, Villani C, Talamini G (2021) The capital value of pedestrianization in Asia’s commercial cityscape: evidence from office towers and retail streets. Transp Policy 107:72–86. https://doi.org/10.1016/j.tranpol.2021.04.017

Noë N, Whitaker RM, Chorley MJ, Pollet TV (2016) Birds of a feather locate together? Foursquare checkins and personality homophily. Comput Hum Behav 58:343–353. https://doi.org/10.1016/j.chb.2016.01.009

Onuoha IJ, Aliagha GU, Rahman MSA (2018) Modelling the effects of green building incentives and green building skills on supply factors affecting green commercial property investment. Renew Sustain Energy Rev 90:814–823. https://doi.org/10.1016/j.rser.2018.04.012

Phillips PC, Perron P (1988) Testing for a unit root in time series regression. Biometrika 75:335–346. https://doi.org/10.1093/biomet/75.2.335

Rinchen S, Yassine A, Schwartzzentruber K, et al (2018) Integrating small scale green energy into smart grids: Prediction for peak load reduction. In: 2018 International Conference on Computer and Applications (ICCA). IEEE, pp 104–109

Roick O, Heuser S (2013) Location based social networks—definition, current state of the art and research agenda: State of the Art in LBSN Research. Trans GIS n/a-n/a. https://doi.org/10.1111/tgis.12032

Sagiroglu S, Sinanc D (2013) Big data: a review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS). IEEE, pp 42–47

Salas-Olmedo MH, Moya-Gómez B, García-Palomares JC, Gutiérrez J (2018) Tourists’ digital footprint in cities: comparing big data sources. Tour Manag 66:13–25. https://doi.org/10.1016/j.tourman.2017.11.001

Seo K, Salon D, Kuby M, Golub A (2019) Hedonic modeling of commercial property values: distance decay from the links and nodes of rail and highway infrastructure. Transportation 46:859–882. https://doi.org/10.1007/s11118-018-9861-z

Singh A, Sharma A, Dubey G (2020) Big data analytics predicting real estate prices. Int J Syst Assur Eng Manag 11:208–219. https://doi.org/10.1007/s13198-020-00946-3

Sirmans C, Guidry K (1993) The Determinants of Shopping Center Rents. J Real Estate Res 8:107–115. https://doi.org/10.1080/10835547.1993.12090700

Snscrape (2021) https://github.com/JustAnotherArchivist/snscrape

Song XP, Richards DR, He P, Tan PY (2020) Does geo-located social media reflect the visit frequency of urban parks? A city-wide analysis using the count and content of photographs. Landsc Urban Plan 203:103908. https://doi.org/10.1016/j.landurbplan.2020.103908

Spliid H (1983) A fast estimation method for the vector autoregressive moving average model with exogenous variables. J Am Stat Assoc 78:843–849

Spliid H (1986) MARIMA-Estimation of multivariate time series models. User guide for a computer program

Spliid H (2016) Multivariate time series estimation using marima. pp 108–123

Spliid H (2017) marima: Multivariate ARIMA and ARIMA-X Analysis. Version 2.2URL https://CRAN.R-project.org/package=marima

Tabachnick BG, Fidell LS, Ullman JB (2007) Using multivariate statistics. Pearson Boston, MA
Tan MJ, Guan C (2021) Are people happier in locations of high property value? Spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data. Appl Geogr 132:102474. https://doi.org/10.1016/j.apgeog.2021.102474

Winson-Geideman K, Krause A (2016) Transformations in real estate research: The big data revolution. In: Presentation at the 22nd Annual Pacific-Rim Real Estate Society Conference. pp 17–20

Wu L, Brynjolfsson E (2015) The future of prediction: How Google searches foreshadow housing prices and sales. In: Economic analysis of the digital economy. University of Chicago Press, pp 89–118

Xu T, Zhang M, Aditjandra PT (2016) The impact of urban rail transit on commercial property value: new evidence from Wuhan, China. Transp Res Part Policy Pract 91:223–235. https://doi.org/10.1016/j.trap.2016.06.026

Yao Y, Zhang J, Hong Y et al (2018) Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Trans GIS 22:561–581. https://doi.org/10.1111/tgis.12330

Yu D, Wei YD, Wu C (2007) Modeling spatial dimensions of housing prices in Milwaukee, WI. Environ Plan B Plan Des 34:1085–1102

Zamani M, Schwartz HA (2017) Using Twitter Language to Predict the Real Estate Market. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valencia, Spain, pp 28–33

Zheng Y (2011) Location-based social networks: users. In: Zheng Y, Zhou X (eds) Computing with spatial trajectories. Springer, New York, pp 243–276

Zhou X, Zhang L (2016) Crowdsourcing functions of the living city from Twitter and Foursquare data. Cartogr Geogr Inf Sci 43:393–404. https://doi.org/10.1080/15230406.2015.1128852

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.