Computers, Environment and Urban Systems 90 (2021) 101702

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Variability in and mixtures among residential vacancies at granular levels: Evidence from municipal water consumption data

Yongting Pan [a], Wen Zeng [a], Qingfeng Guan [a,*], Yao Yao [a,b], Xun Liang [a], Yaqian Zhai [a], Shengyan Pu [a]

[a] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430078, Hubei province, China
[b] Department of Data Technology and Products, Alibaba Group, Hangzhou 311121, Zhejiang province, China

ARTICLE INFO

Keywords:
China's new-type urbanization
Residential vacancy
Municipal water consumption
Granular level
Variability
Mixture

ABSTRACT

Unprecedented urbanization in China has directly resulted in residential vacancies, which has seriously stunted sustainable development, a part of China's new-type urbanization plan. Understanding the various types and mixes of residential vacancies is critical for the advancement of our knowledge of speculative urbanism and for devising vacancy-mitigation policies, but this issue remains insufficiently studied. Using municipal water consumption data, this study proposes a feasible and general-purpose framework for providing innovative insights into the variability in residential vacancies at the household level and the mixture of residential vacancies at the building level. This framework was applied to the city of Changshu, China, and four categories of vacant residences at the household level were identified: seasonally vacant residences, long-term vacant residences, newly built residences and occasionally vacant residences. The first category is closely related to tourism and seasonal industries, while the last three exhibit a Matthew effect. In addition to revealing significant and intensifying spatial clustering and three patterns of changes in vacancy mixtures (i.e., emergence, disappearance, and increases or decreases), the results identify particular types of vacant residences at the building level (e.g., extremely low-entropy long-term multihousehold buildings). The insights from this study can contribute to devising customized policies for alleviating residential vacancies.

1. Introduction

Drastic urbanization in China has aroused widespread and increasing attention and is acknowledged as the greatest human-resettlement experiment in history (Bai, Shi, & Liu, 2014) as well as an event with a profound impact on global human civilization and development (Montgomery, 2008). In March 2014, China announced the National New-type Urbanization Plan (2014–2020) (NNUP). As the first official plan to establish new-type urbanization as a national policy (Chen, Liu, & Lu, 2016), the NNUP pinpoints the serious problem that the growth in urbanized land has outpaced the growth in the urban population owing to the previous traditional land-centered urbanization (Feng, Liu, & Qu, 2019; Long, 2014), which results in excessive land resource development (e.g. housing supply) outpacing the residential demand (Jiang, Mohabir, Ma, & Zhu, 2017; Jin et al., 2017; Zheng et al., 2017). Such imbalance has been one of the most important driving forces for residential vacancies in China (Jin et al., 2017; Leichtle, Lakes, Zhu, & Taubenbck, 2019; Zheng, Wang, & Cao, 2014).

Residential vacancies in China have recently attracted the attention of domestic and foreign media as well as of scholars. Fernando (2010) claimed an estimated 64 million vacant apartments and houses in China in 2010 according to Chinese media reports. In 2014, the planned area of the reported twenty-eight wasted cities in China unexpectedly covered 3643 km^2 based on the documented data (He, Mol, & Lu, 2016), of which the calculated gross floor area of total vacant residential buildings up to 765.09 km^2. According to the China Household Finance Survey, in 2011, 2013, 2015, and 2017, the estimated number of vacant residences in towns and cities reached 42, 47.5, 56, and 65 million, respectively. In 2018, based on LJ1–01 night-light data, TanDEM-X data, and Global Urban Footprint data, the identified "ghost neighborhoods" across the urban landscapes of China covered a total area of 353.64 km^2, and the total gross floor area of which ranged between 530.46 km^2 and 707.28 km^2 (Shi et al., 2020). According to the broken window theory (BWT), residential vacancies have extremely adverse effects on society, the

* Corresponding author.
E-mail addresses: panyt@cug.edu.cn (Y. Pan), zengwen@cug.edu.cn (W. Zeng), guanqf@cug.edu.cn (Q. Guan), yaoy@cug.edu.cn (Y. Yao), liangxun@cug.edu.cn (X. Liang), zyq2017@cug.edu.cn (Y. Zhai), pushengyan@cug.edu.cn (S. Pu).

https://doi.org/10.1016/j.compenvurbsys.2021.101702
Received 29 March 2021; Received in revised form 26 July 2021; Accepted 8 August 2021

economy and the environment (He et al., 2016; Stern & Lester, 2021). In addition to decreasing the quality of life (Newman, Gu, Kim, & Li, 2016), property values (Morckel, 2014), community security (Du, Wang, Zou, & Shi, 2018), and the functioning effectiveness of urban system (Jin et al., 2017), residential vacancies also increase the crime rates (Zou & Wang, 2019), depopulation (Kim, Newman, & Jiang, 2020), and urban unsightly aesthetic (Newman, Park, & Lee, 2018). Therefore, understanding vacant residences is critical to facilitate the sustainable development of China's new-type urbanization plan (Lang, Long, & Chen, 2018) and to advance knowledge of speculative urbanism and devise effective urban policies for local governments and planners.

Not all vacant residences are due to excess supply (Huuhka, 2016; Molloy, 2016). It is unfair to treat all vacant residences in the same way (Chi, Liu, Wu, & Wu, 2015). Managing vacant residences does not necessarily demand considerable investments, nor ought blind and wholesale demolitions be conducted (Nam, Han, & Lee, 2016; Silverman, Yin, & Patterson, 2013). Thus, categorizing the various types of vacant residences is the crux of precise management and contributes to developing customized policies (Jiang et al., 2017).

To date the unavailability of the definitions, statistics, or alternative information concerning various categories of vacant residences results from a lack of the all-round and in-depth investigation in China conducted by the official institutional resources (Shi et al., 2020; Williams, Xu, Tan, Foster, & Chen, 2019; Zheng et al., 2017). Even in some countries (e.g., USA), the categories of vacant residences data is collected via field survey, time-consuming, labour-intensive, hardly comprehensive in coverage, and too infrequent to catch up with the rapidly changing (Leichtle et al., 2019; Newman et al., 2018; Zou & Wang, 2019).

Meanwhile, attempts to understand the vacancy categories have been made by some studies. By summarizing the causes of vacancy, cities with extensive vacant houses could be sorted into disaster-induced ghost cities, decline-induced ghost cities, and plan-induced ghost cities (Nie & Liu, 2013). According to the urban development, the cities could be classified into either ghost cities or tourism cities (Chi et al., 2015), or ghost cities characterized by suburban tourism, debt-financed urbanism, and pro-growth strategic urbanism (Jiang et al., 2017), or cities with unbalanced economic systems, resource-exhausted cities, and tourist destinations (Lu, Zhang, Liu, Ye, & Miao, 2018), or declining cities, growing cities, and stagnant cities (Yoo & Kwon, 2019).

Although existing studies have provided insights into the categories of cities with massive vacant residences primarily based on the characteristics of the entire city, most of these studies have not only ignored the complexity and diversity within a city (Qiu, 2012) but also obscured the fluidity and heterogeneity of vacant residences (Newman et al., 2016; Roth, 2019). Challenges still remain.

(1) The variability in residential vacancies has not been sufficiently explored at a granular level. Some studies have found that the types of vacant residences within a city are varied (Molloy, 2016; Mui, Jonessmith, Thornton, Porter, & Gittelsohn, 2017) and even overlap each other (Jiang et al., 2017). For instance, some historical sites may be regarded as long-term vacant residences. Vacant residence categories may also change rapidly (Newman et al., 2016; Yoo & Kwon, 2019), given the recent spatial expansion and internal redevelopment or renewal occurring at rapid speeds within the cities in China (Zhang & Li, 2020). Therefore, exploring the variability in the categories of vacant residences at a granular level (e.g., the household level) is essential for deepening and refining our understanding of residential vacancies in China and is the basis for devising customized mitigation policies; however, it remains insufficiently studied.
(2) The mixture of residential vacancies at a granular level has been inadequately described and analyzed. Buildings are the fundamental components of a city and the basic units for urban planning, and most likely, the vacancy categories within residential buildings are not homogeneous due to the variability in residential vacancies at the household level. Specifically, the vacancy categories for single-household buildings are definitely unique, while those for multihousehold buildings tend to be diverse and can vary considerably in quantity and proportion. Therefore, estimating the mixture of vacancy categories within residential buildings is key for providing spatially explicit insights into the diversity of residential vacancies. This contributes to the implementation of customized policies by providing practitioners with more useful information (e.g., places and priorities) than just categories.

In recent years, municipal infrastructure and service data (e.g., on the consumption of municipal water, electricity, and natural gas) have been widely used to monitor residential vacancies (Haramati & Hananel, 2016; Kumagai, Matsuda, & Ono, 2016; Li, Guo, & Lo, 2019; Nam et al., 2016; Pan et al., 2020). Municipal infrastructure and service data are often provided solely by public service agencies (e.g., local governments and municipal utilities) with the advantages of having fine-grained spatiotemporal resolution, a long time frame, as well as wide coverage of the population and socioeconomic activities in cities and towns, including small or developing ones (Guan, Cheng, Pan, Yao, & Zeng, 2020; Pan et al., 2020). In addition, municipal infrastructure and service data provide possibilities for objectively measuring and modeling the nature and likely causes of the activity patterns of residents (Guan et al., 2020; Lansley & Longley, 2016). Obviously, municipal infrastructure and service data can be used as reliable indicators of residential vacancies at granular levels.

To provide innovative insights into the variability in and mixtures of residential vacancies at granular levels using municipal water consumption data, this study aims to address three unique and challenging questions: (1) How can we identify the various categories of residential vacancies? (2) What are the differences in vacancy patterns among the various vacant residence categories at the household level? (3) What are the spatial and temporal patterns in residential vacancy mixtures at the building level? In addition to examining the patterns and processes of various vacant residences within a city and advancing our understanding of the residential vacancies embedded in the urbanization of China, answering these questions sheds light on policy implications for the mitigation of residential vacancies and further helps decision-makers ensure the sustainability of people-oriented urbanization.

2. Study area and data

2.1. Study area

Changshu (Fig. 1), a typical developing county-level city, is located

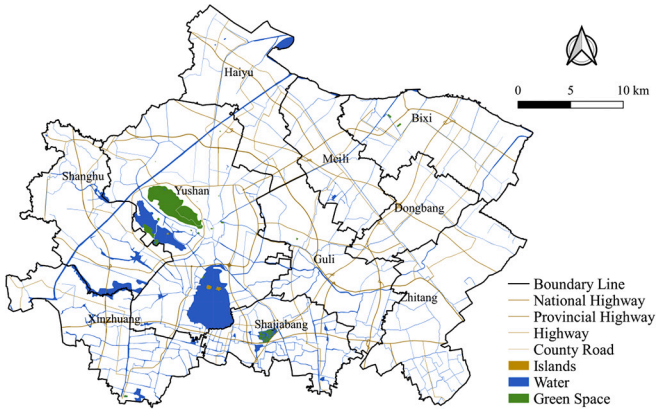


Fig. 1. Case study area: Changshu, Jiangsu, China.

in the southeast of Jiangsu Province as well as in the hinterlands of the Yangtze River Delta, China (31°31′N-31°50′N, 120°33′E-121°03′E). Changshu encompasses nine towns and one district, occupying an administrative area of 1264 km^2. Changshu has experienced accelerated urban development in recent years, with urban areas and urban populations increasing at an unprecedented pace. According to the Urban Construction Statistical Yearbook (2004–2013), the area of urban construction land in Changshu expanded from 62.7 km^2 to 85.15 km^2 between 2004 and 2013. The urban population reached 408,300 by the end of 2013, an increase of 85,500 compared to 2004. Note, however, that the area of urban construction land in Changshu had grown by 3.78% year on year - much faster than its urban population, which grew by 2.78% year on year. The urban population density also declined by 5.35% year on year, from 2131 persons/km^2 to 1197 persons/km^2. Such an imbalance between urbanized land growth and urban population growth implies a considerable number of vacant residences in Changshu. Many studies (Chi et al., 2015; Jin et al., 2017; Pan et al., 2020; Shi et al., 2020) further confirmed their existence, thus making Changshu a suitable location to study residential vacancies.

2.2. Data sets

2.2.1. Municipal water consumption data and residential state time series

The municipal water consumption data for individual customers in Changshu, the core data used in this study, were provided by the municipal water company for the period from January 2004 to December 2013. For each individual customer, the data included the following: (i) the customer ID (account), (ii) spatial coordinates, (iii) the time series for monthly water consumption and (iv) the cancelation date for the customer account. In this study, 286,365 residential customers were analyzed.

It is important to note some specific features of the residential water consumption data. (i) The data included when the customer accounts were canceled but not when they were registered. Consequently, for each individual residential customer, it is possible to determine when the water supply service was terminated but not when the service started. (ii) Blanks of various lengths were found in the monthly water consumption time series over irregular time spans. For individual residential customers, the water supply service started and ended at various times and may have been suspended before cancelation for a variety of reasons. In addition, to the best of our knowledge, neither a globally recognized definition of nor a globally accepted standard for vacant residences exist at present. The definition of a vacant residence used in different countries and studies varies according to the standards of usage. Thus, to identify whether a residence is vacant or occupied at a certain time, the conceptual definition of vacant residence and the residential vacancy identification method (RVIM) proposed by Pan et al. (2020) were adopted in this study. Using the RVIM, the numeric residential water consumption time series with blanks and various lengths were converted into nominal residential state time series with unified lengths. At any time during the study period, the state of a residential customer can be one of the following: *Absent* (meaning the customer account has been canceled), *Vacant* (meaning the residence is unoccupied), *Occupied* (meaning the residence is occupied), and *Unknown* (meaning the situation is unclear for some reason, e.g., the customer account has not yet been registered). In this study, the total number of identified vacant residences in Changshu was 227,421, with an increase from 95,188 in 2004 to 134,468 in 2013.

2.2.2. Building footprint geometries

The building floorplans for Changshu in 2015 (Fig. 2) were collected from the local planning authority. In addition to perimeter and area, the building floorplans included the building types associated with residential and nonresidential functions. In line with the design code for residential buildings (GB 50096–2011), buildings with an area less than 22 m^2 were also removed. As a result, a total of 435,188 buildings were used in this study.

3. Methodology

An overview of the proposed framework is depicted in Fig. 3. The framework starts by collecting a corpus of vacant residences derived from the residential state time series and then implements the Doc2Vec and K-means++ methods to identify the categories of residential vacancies, thereby exploring the variability in residential vacancies at the household level and the mixture of vacancy categories at the building level. The framework includes three distinct components: (1) The Doc2Vec algorithm is employed to convert the vacant residence corpus into vectorized document representations, and then the K-means++ algorithm is used to categorize the vacant residences on the basis of the obtained vacant residence vectors. Such a method is inspired by Le and Mikolov (2014), which stated quite definitively that "*The Doc2Vec algorithm can be applied to create vector representations for sequential data and the vectors can be fed directly to conventional machine learning*

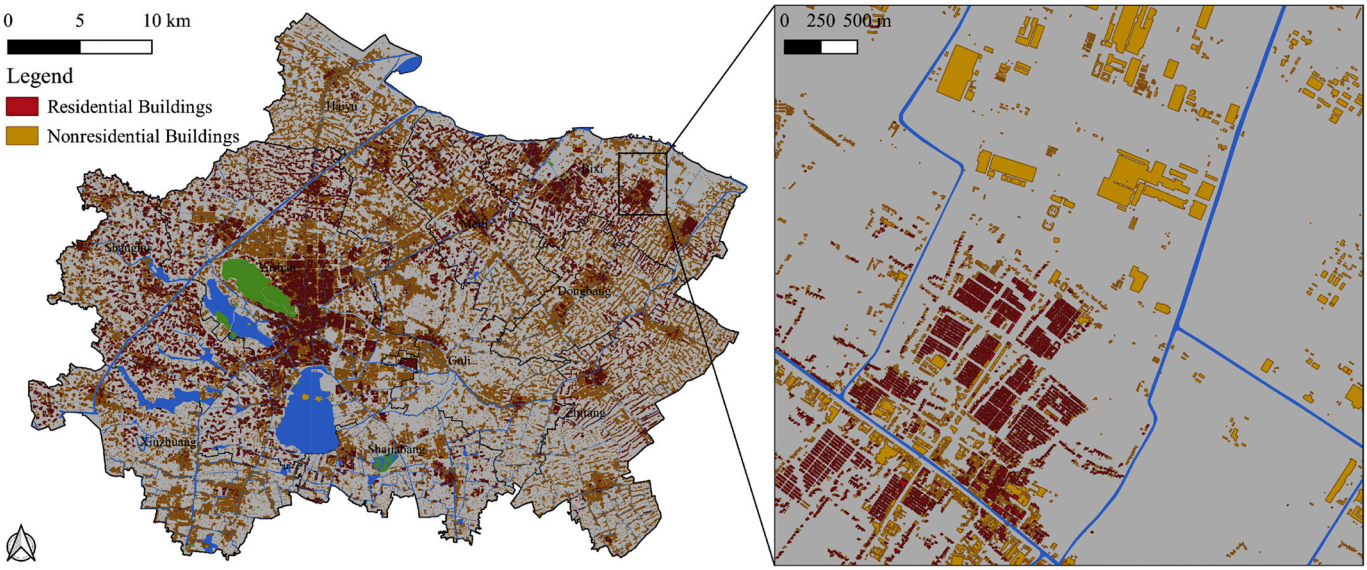


Fig. 2. Building floorplans for Changshu in 2015.

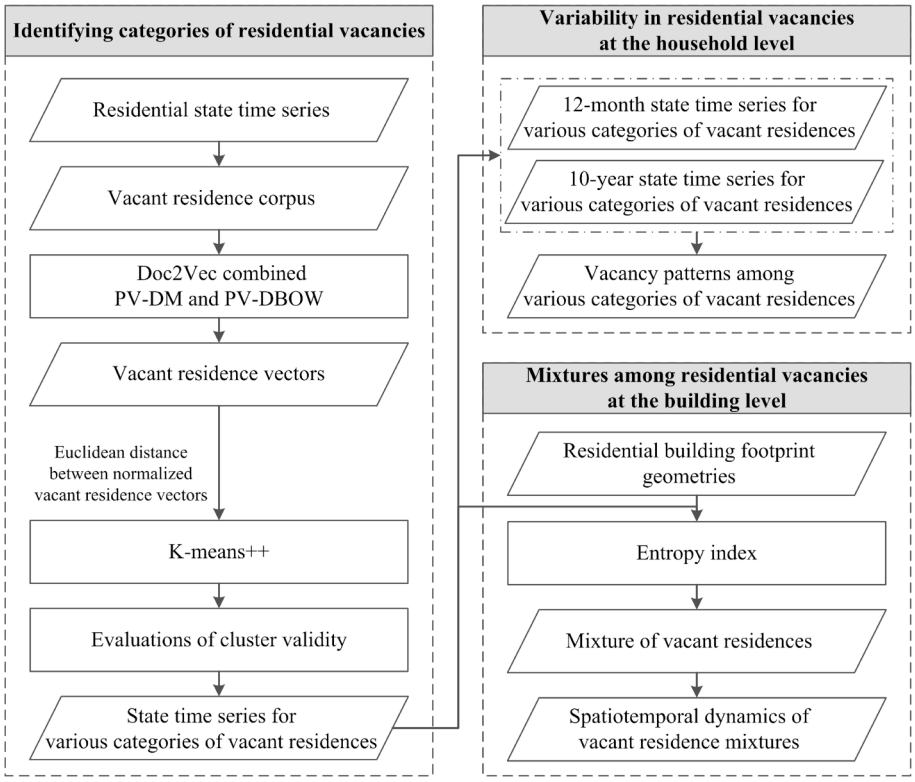


Fig. 3. Overview of the proposed framework.

techniques (e.g., K-means) for performing classification tasks." It is of the utmost importance to note that this study has no a priori knowledge of the categories of residential vacancies but rather allows the cluster validity indices (e.g., the silhouette index and the Davies-Bouldin index) to determine the optimal number of vacant residence clusters and to assess cluster quality. (2) Based on both the categories and state time series of vacant residences, 12-month and 10-year state time series for various categories of vacant residences are generated; thus, the differences in the vacancy patterns among the various categories of vacant residences at the household level are explored. (3) Based on the vacant residence categories at the household level, the mixture of vacant residences at the building level is measured using the entropy index; therefore, both the spatial distributions of and temporal evolutions in the mixtures of vacant residences are identified.

3.1. Preparing the vacant residence corpus

A corpus is a large collection of texts (Jurafsky & Martin, 2010). In this study, the vacant residence corpus is a collection of over a million observations from 227,421 vacant residences in Changshu. The corpus was constructed based on the residential state time series data through the following two steps.

(1) *Absent* and *Unknown* were removed because they provide little information on the residential vacancy categories. For example, for the second vacant residence (#2) in Fig. 4A, its vacant pattern is most likely similar or even identical to that of the first vacant residence (#1) when *Absent* in the state time series is ignored. This seems to align with the actual situation. The same is true for the fourth vacant residence (#4) in Fig. 4B. Hence, this study removed the *Absent* and *Unknown* categories from the residential state time series. It should be noted that this results in the

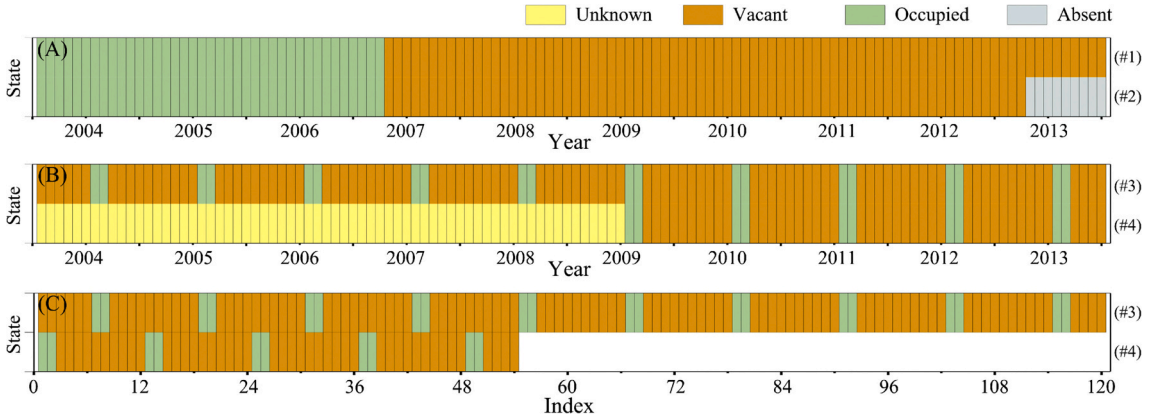


Fig. 4. Examples of residential state time series.

residential state time series varying considerably in length, from 2 to 120 months.

(2) Temporal information - month - was added to the residential state time series, which used nominal values and had varied lengths. Removing *Absent* and *Unknown* may result in the loss of temporal information and then the dislocation of the residential states (see Fig. 4C). Hence, the month information was added to the nominal state of each vacant residence. For example, 12*Vacant* indicates that a certain vacant residence was vacant in December. Adding temporal information to the residential state time series also allows for the identification of periodicity in the residential vacancy states.

3.2. Obtaining document representations of vacant residences

The neural-network-based Doc2Vec (alternatively known as Paragraph Vector) (Le & Mikolov, 2014), an unsupervised algorithm, facilitates the conversion from a variable-length document into a fixed-length numerical representation (vector) while preserving its crucial information (Kim, Seo, Cho, & Kang, 2019). Doc2Vec extends Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and goes beyond the word level to achieve document-level representations (Kim et al., 2019). The architecture of the neural network in Doc2Vec, a paragraph vector with distributed memory (PV-DM) and a paragraph vector with distributed bag of words (PV-DBOW) (Le & Mikolov, 2014), are essentially identical to that of Word2Vec (Kim, Kim, & Cho, 2017). The difference lies in the additional document input into the network, along with the words (Benabderrahmane, Mellouli, & Lamolle, 2018), which works as a form of memory storage that encodes the topic of the document or the information missing from the current words (Le & Mikolov, 2014). In Doc2Vec, each document and each word are mapped to a unique vector, represented by a column in matrices D and W, respectively (Kim et al., 2019). The document vectors and word vectors are trained using stochastic gradient descent, in which the gradient is gained via backpropagation and the objective is to maximize the average log probability:

$$l(\theta) = \frac{1}{T}\sum logP(w_O|w_I, D) \tag{1}$$

$$P(w_O|w_I, D) = \frac{exp\left(v^T_{w_I D} v'_{w_O}\right)}{\sum exp\left(v^T_{w_I D} v'_{w_O}\right)} \tag{2}$$

where T denotes the total number of words, w_I and w_O denote the input and output words, D denotes the document, $v_{w_I D}$ denotes the vector representations of the input words and the input document, and v_{w_o} denotes the vector representations of the output words. Note that the prediction is typically performed via a multiclass classifier (e.g., softmax).

Recent deep learning and natural language processing studies have claimed that Doc2Vec outperforms other embedding schemes in document-mining tasks (Benabderrahmane et al., 2018; Kim et al., 2019) due to two major advantages of Doc2Vec. (1) Doc2Vec learns from unlabeled data and thus works well without sufficiently labeled data (Le & Mikolov, 2014). (2) Doc2Vec overcomes the key weaknesses of bag-of-words and weighted averaging word vectors (i.e., the failure to consider the semantics of words and the ordering of words), thus addressing the problem that different documents may have the same representations (Chang, Xu, Zhou, & Cao, 2018; Pradhan & Pal, 2020). To date, Doc2Vec has been successfully applied in various realms and applications, including sentiment analysis (Lee & Kim, 2017), recommendation systems (Benabderrahmane et al., 2018), abnormality detection (Chang et al., 2018), and source code retrieval and comparisons (Nafi, Roy, Roy, & Schneider, 2020).

Based on the vacant residence corpus, Doc2Vec combined with PV-DM and PV-DBOW was employed in this study to obtain the vector representations of all vacant residences in Changshu (i.e., the vacant residence vectors), given that this combination has been strongly recommended by previous studies (Kim et al., 2019; Le & Mikolov, 2014). Our study area, Changshu, was regarded as a corpus, and each vacant residence contained within it was treated as a document, and each monthly state for each residence was treated as a word. For a vacant residence, each monthly state was treated as a target word, and its surrounding monthly states (i.e., the states of the previous months and following months) were treated as contextual words. The hyperparameters were set for training an effective Doc2Vec model, including setting the dimensionality to 10 for the generated document vectors, the context window to 3, the training epoch to 3, and the learning rate to 0.025.

3.3. Categorizing vacant residences

Based on the vacant residence vectors, an unsupervised clustering algorithm, K-means, was adopted to generate a set of clusters that describe the groupings of vacant residences with shared salient characteristics. Owing to its simplicity (Xia, Karimi, & Meng, 2017) and efficiency (Xu, Du, Mao, Zhang, & Liu, 2020), K-means remains by far the most commonly used clustering method (Gašparović, Zrinjski, & Gudelj, 2019; Hincks, Kingston, Webb, & Wong, 2018). K-means partitions the vacant residences into K clusters by minimizing the intracluster variation and maximizing intercluster differences (Waldron, O'Donoghue-Hynes, & Redmond, 2019). The Euclidean distance between normalized vectors, which is similar to the cosine distance between nonnormalized vectors (France, Carroll, & Xiong, 2012), was chosen as the metric for measuring the distances between vacant residence vectors. Considering that the performance of K-means is highly dependent on the initial cluster centroids (Ghodousi, Alesheikh, & Saeidian, 2016), the K-means++ method (Arthur & Vassilvitskii, 2007) was employed to determine the initial cluster centroids. K-means++ chooses an initial cluster centroid with a probability proportional to its distance from the closest centroid already chosen. This also effectively improves the efficiency of clustering, as well as the accuracy. In this study, the hyperparameter K was set to range from a two-cluster to a ten-cluster solution.

The quality of the clustering results can be evaluated using cluster validity indices (Lord, Willems, Lapointe, & Makarenkov, 2017). These indices essentially measure the compactness of the objects in the same cluster and their separation into distinct clusters (Lord et al., 2017). In this study, two internal validity indices were employed: the silhouette index (SI) (Yao et al., 2017) and the Davies-Bouldin index (DBI) (Lord et al., 2017). The SI determines the similarity of each vacant residence with all the other vacant residences within its cluster and its dissimilarity from the vacant residences belonging to other clusters, while the DBI determines the dissimilarity of each cluster to other clusters:

$$\mathrm{SI} = \frac{1}{N}\sum_{r=1}^{N}\frac{b(r) - a(r)}{max\{a(r), b(r)\}} \tag{3}$$

$$\mathrm{DBI} = \frac{1}{K}\sum_{i=1}^{K} max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \tag{4}$$

where N is the number of vacant residences. $a(r)$ represents the average distance of the r-th vacant residence to all the other vacant residences within its cluster, and $b(r)$ represents the minimum average distance to the vacant residences in other clusters. K is the number of vacant residence clusters. σ_i and σ_j represent the average distance of all vacant residences in the i-th vacant residence cluster and the j-th vacant residence cluster to their respective cluster centroids c_i and c_j, and $d(c_i, c_j)$ represents the distance between centroids c_i and c_j. The SI varies between −1 and 1, and a high SI indicates compact and well-separated vacant residence clusters. The DBI is not specifically limited, and a

lower DBI indicates a better assignment of vacant residences to clusters.

3.4. Measuring the mixture of vacant residences

Based on the categories of vacant residences at the household level, this study employed Shannon's entropy-based indicator, the entropy index (Nazarnia, Harding, & Jaeger, 2019), to investigate the degree of mixture in residential vacancies at the building level. Entropy index, a widely adopted mixing measure (Hipp, Kane, & Kim, 2017), has been shown to be capable of analyzing the diversity of spatially-complex urban ecosystems such as housing (Cho & Kim, 2017).

$$\mathrm{VREI}_i = \frac{-\sum_{j=1}^{M} P_{ij} ln P_{ij}}{ln(M)} \quad (5)$$

where $VREI_i$ refers to the vacant residence entropy index (VREI) of the i-th ($1 \leq i \leq$ the quantity of residential buildings) residential building. M is the total number of categories of residential vacancies (in this study, M = the number of vacant residence clusters). P_{ij} is the proportion of the j-th residential vacancy category within the i-th residential building. The VREI ranges from zero (homogeneity, wherein the residential building is dominated by a single category of vacant residences) to one (heterogeneity, wherein the vacancies in the residential building are evenly distributed among all categories). An absolutely homogeneous residential building is defined as an extremely low-entropy residential building, one in which diversity in residential vacancies does not exist.

In addition, this study identified the spatial clustering of vacant residence mixtures at the building level using the robust and frequently used global Moran's I (Wang, Li, Myint, Zhao, & Wentz, 2019), where the patterns of features (i.e., vacant residences at the building level) were evaluated as clustered, dispersed or random distributions based on their locations and associated attributes (i.e., their VREI). The distance band (or threshold distance) for the spatial analysis was set to 500 m, given that the average edge length of a residential land parcel is between 300 m and 500 m.

4. Results

4.1. Vacant residence vector extraction and validity evaluations

To guarantee the reliability of the clustering results, this study performed Doc2Vec and K-means++ 50 times on 227,421 vacant residences in Changshu. As illustrated in Fig. 5, the SI was notably high when vacant residences were grouped into either four clusters or six clusters. This suggests that either four or six vacant residence clusters are potentially appropriate. The DBI became progressively lower as the number of vacant residence clusters increased. A reasonable number of clusters - four vacant residence clusters - was finally chosen using the elbow method (Uddin et al., 2019). Therefore, the four categories of vacant residences were chosen on the basis of the optimal number of clusters in this study.

4.2. Categories of vacant residences at the household level

For each category of vacant residence, the ten-year monthly nominal residential state time series was aggregated into a set of 12-month state time series (Fig. 6A) and a set of 10-year state time series (Fig. 6B) to detect vacancy patterns. Specifically, for each vacant residence, the state time series was converted into four different time series. (1) Two 12-month state time series, one in which each element represents the most frequent residential state (i.e., occupied or vacant) and one in which each element represents the percentage of residential vacancies (i.e., the ratio between the number of vacant states and the total number of states), both for a particular month over the ten years. Both the most frequent residential state and the percentage of residential vacancies are fair indicators of the extent of vacancy as neither the mean nor the standard deviation, nor summaries dependent on those statistics can be calculated for nominal residential state time series. Based on the former, the most frequent residential state (scatters in Fig. 6A(a)) and the percentage of residential vacancies (lines in Fig. 6A(a)) for a particular category of vacant residences were calculated. And the variations in the latter for a particular category of vacant residences were simplified via summary plots (e.g., box plots (Fig. 6A(b)) and ribbon plots (Fig. 6A(c))) (Motlagh, Berry, & O'Neil, 2019) without assumptions on the underlying statistical distributions. These summary plots visually combine important distribution metrics such as the central tendency, the core

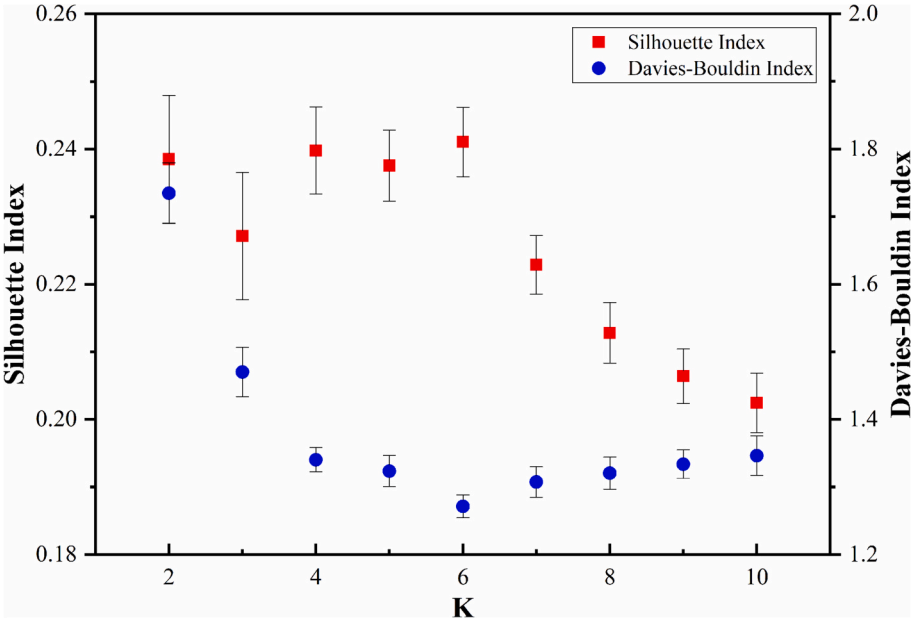
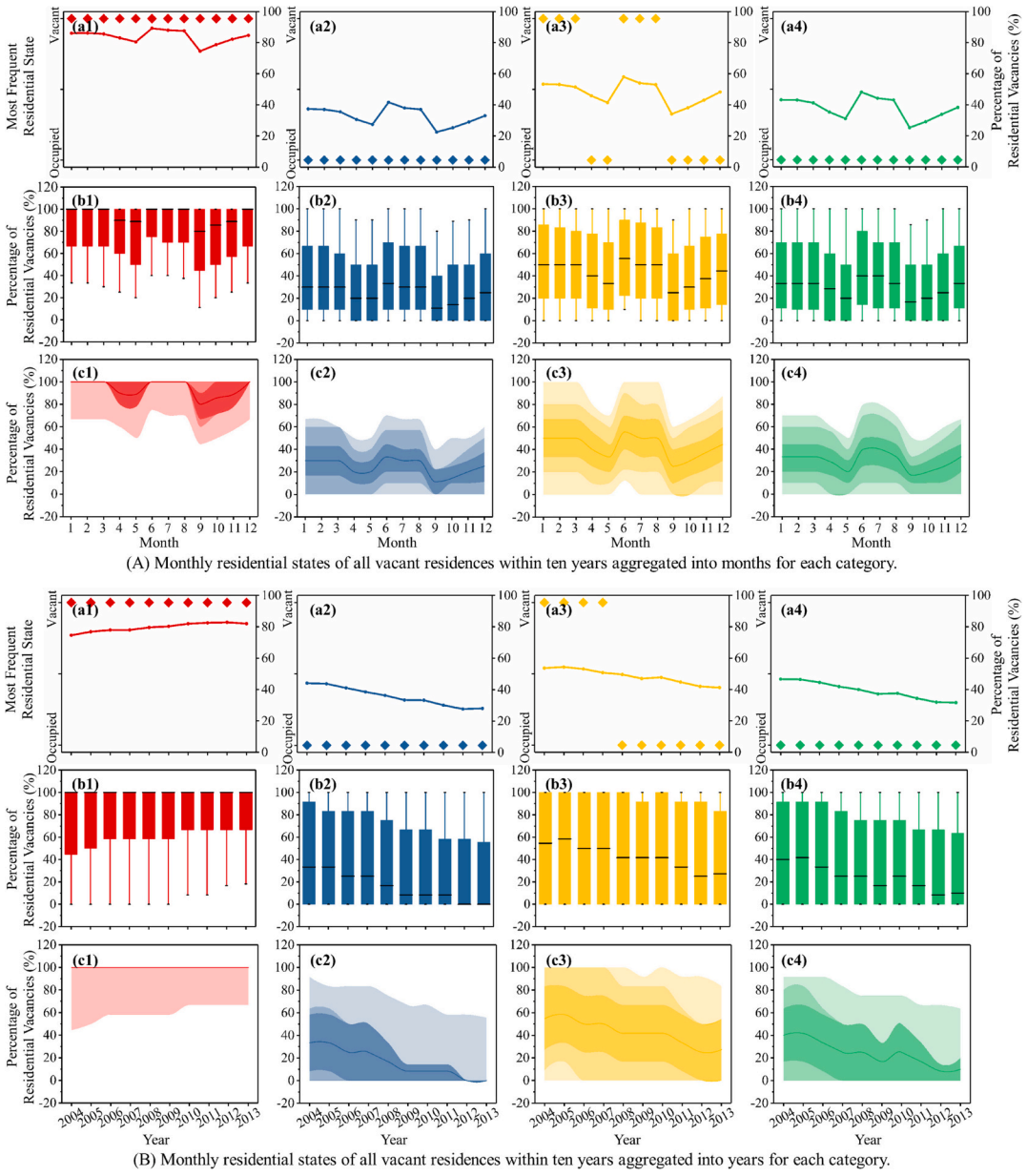


Fig. 5. Variation in the SI and DBI with respect to the number of vacant residence clusters K. Each point and each whisker represent the mean and standard deviation of the index over 50 independent runs.



(A) Monthly residential states of all vacant residences within ten years aggregated into months for each category.

(B) Monthly residential states of all vacant residences within ten years aggregated into years for each category.

Fig. 6. Scatter-line plots (a), box plots (b) and ribbon plots (c) representing the 12-month state time series (A) and 10-year state time series (B) for four categories of vacant residences in Changshu. The scatter-line plots are composed of the most frequent residential state and the percentage of residential vacancies. The box plots consist of the medians, with the boxes encompassing approximately the middle 50% of the data and the lower and upper whiskers at 10% and 90%. The ribbon plots are composed of the central median curves and the surrounding 25%, 50% and 75% central distributions. The four categories of vacant residences are numbered from 1 to 4 and represented by four different colors: red, blue, yellow and green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distribution, outliers and skewness (Motlagh et al., 2019), which are more robust than statistics that are highly sensitive to outliers (e.g., the mean and the standard deviation) (Cox, 2019). (2) Two 10-year state time series, one in which each element represents the most frequent residential state and one with the percentage of residential vacancies, both for a particular year. For illustration, scatter-line plots (Fig. 6B(a)) representing the former as well as box plots (Fig. 6B(b)) and ribbon plots (Fig. 6B(c)) representing the latter for a particular category of vacant residences are depicted.

The first category of vacant residences (colored red in Fig. 6): Both monthly and yearly most frequent residential states show that vacancy accounted for nearly the entire period, and the percentage of residential vacancies was extremely high (approximately 100%). Both monthly and yearly core distributions showed that the percentage of vacancies for most vacant residences were far above 50%. Even monthly extreme valleys showed that the percentage of vacancies among 90% of vacant residences was greater than 40%. Therefore, these vacant residences were classified as **long-term vacant residences**. It is interesting to note that the smaller dispersion of yearly core distributions and the increasing yearly extreme valleys indicate that the extent of vacancy in long-term vacant residences has gradually increased.

The second category of vacant residences (colored blue in Fig. 6): Unlike the long-term vacant residences, both monthly and yearly most frequent residential states show that occupancy accounted for nearly the entire period, and the percentage of residential vacancies was relatively low (approximately 20%). Although the monthly core distributions showed that the percentage of vacancies for most vacant residences was below 70%, the skewness is marked. Specifically, vacant residences

within core distributions were negatively skewed towards a low percentage of vacancies. These vacant residences were considered short-term vacant residences. It is also noteworthy that the dispersion of yearly core distributions markedly decreased, and yearly skewness was quite significant and gradually skewed towards a lower percentage of vacancies. This implies that these short-term vacant residences were most likely **newly built residences**. After registering customer accounts, the newly built residences might be underutilized due to housing logistics (e.g., decoration and moving). Once residents moved in, the extent of vacancy was low.

The third category of vacant residences (colored yellow in Fig. 6): Monthly most frequent residential states show that vacancy accounted for approximately 50% of the entire period and alternated between occupancy as the seasons changed. The dispersion of monthly core distributions and outliers were relatively wide (approximately 10%–80% and 0–100%, respectively), and marked skewness barely exists. Therefore, these vacant residences were classified as **seasonally vacant residences**. Note, in addition, that the seasonally vacant residences progressively decreased in the extent of their vacancies each year, as the most frequent residential states moved from vacancy to occupancy, the dispersion of core distributions decreased, and the gradually marked skewness shifted towards a low percentage of vacancy.

The fourth category of vacant residences (colored green in Fig. 6): Similar to the newly built residences, these vacant residences were also considered short-term vacant residences. However, a higher percentage of vacancies, a greater dispersion of core distributions and less significant skewness indicated that these short-term vacant residences were probably **occasionally vacant residences**.

As shown in Fig. 7, among the four categories of vacant residences, seasonally vacant residences accounted for the largest proportion (nearly 35.3%), while long-term vacant residences exhibited the most rapid growth (8.16% year on year). Quite evidently, there were almost twice the number of seasonally vacant residences as long-term vacant residences, although the long-term vacant residences had roughly doubled between 2004 and 2013.

4.3. *Spatiotemporal dynamics of vacant residence mixtures at the building level*

4.3.1. *Spatial distributions of vacant residence mixtures*

Fig. 8 shows statistically significant positive spatial autocorrelation for vacant residence mixtures at the building level in Changshu (Moran's $I > 0.3$, z-score > 370.0, $p < 0.001$), indicating the presence of a

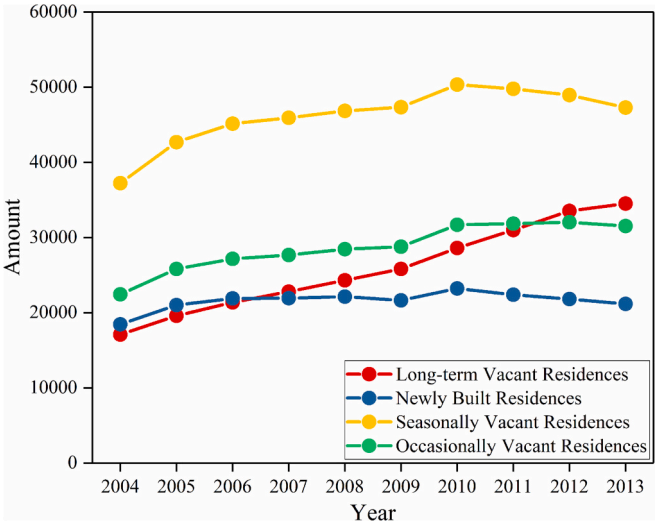


Fig. 7. The number of vacant residences in the four categories from 2004 through 2013 in Changshu.

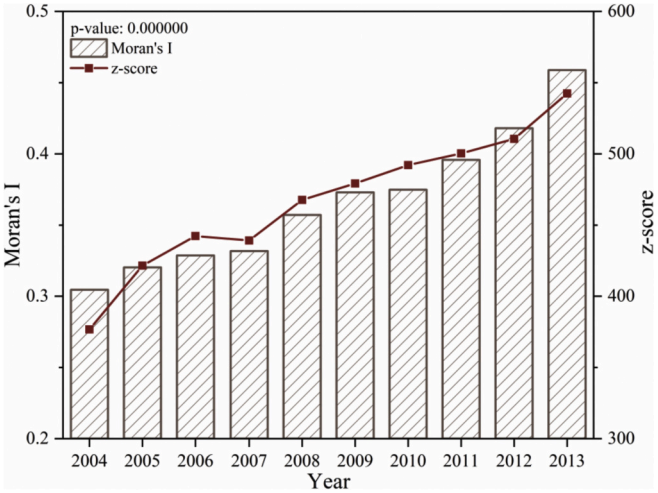


Fig. 8. Global Moran's I statistics for vacant residence mixture at the building level from 2004 through 2013 in Changshu.

significant clustering pattern in the spatial distribution of vacant residences. Moreover, from 2004 to 2013, the steady increase in Moran's I and the z-score (4.70% and 4.19% year on year, respectively) indicates the intensification of clustering in vacant residence mixtures at the building level in Changshu year over year.

The spatial distributions of vacant residence mixtures at the building level from 2004 through 2013 in Changshu exhibit a binary structure, in which the central area of the city had a high level of heterogeneity in residential vacancies, whereas the peripheral areas of the city experienced high levels of homogeneity (Fig. 9). We considered this to be related to residential building types. In the downtown area, multihousehold buildings are dominant, while in the suburban areas, single-household buildings are dominant (Lu, Im, Rhee, & Hodgson, 2014). For multihousehold buildings, the VREI varies from zero to one. It is by definition greater than zero when at least two categories of vacant residences occur in the same building and is one (the highest level of heterogeneity) when all four categories of vacant residences are evenly distributed. For single-household buildings, the VREI is always zero (absolute homogeneity). This was also confirmed through receiver operating characteristic (ROC) analysis. The area under the ROC curve (AUC) ranged between 0.775 and 0.808, indicating that the different VREIs are able to distinguish between residential building types accurately (see the ROC curves in Fig. 9). Furthermore, Fig. 9 allows us to observe annual trends, identifying a gradual increase in the heterogeneity among residential vacancies in the central area of the city, as well as a progressive expansion to the surrounding areas. The cause is described in Section 4.3.2.

Plainly, many extremely low-entropy residential buildings exist. In 2012, for instance, from the perspective of residential building types, such buildings may have been multihousehold buildings or single-household buildings (Fig. 10A). Additionally, from the perspective of the single category of vacant residences contained in, they could have been buildings with long-term vacancies, newly built vacancies, seasonal vacancies, or occasional vacancies (Fig. 10B). Furthermore, certain kinds of extremely low-entropy residential buildings are concentrated in certain areas. In 2012, extremely low-entropy multihousehold buildings in Changshu were mostly located in Yushan town (the central area of Changshu) and the Bixi new district (a subcenter in the northeast, identified as such by Guan et al. (2020)). Although the four categories of extremely low-entropy residential buildings with the single category of vacancies (e.g., extremely low-entropy buildings with long-term vacancies) were distributed across many areas within Changshu, some areas were dominated by a particular category of vacancies.

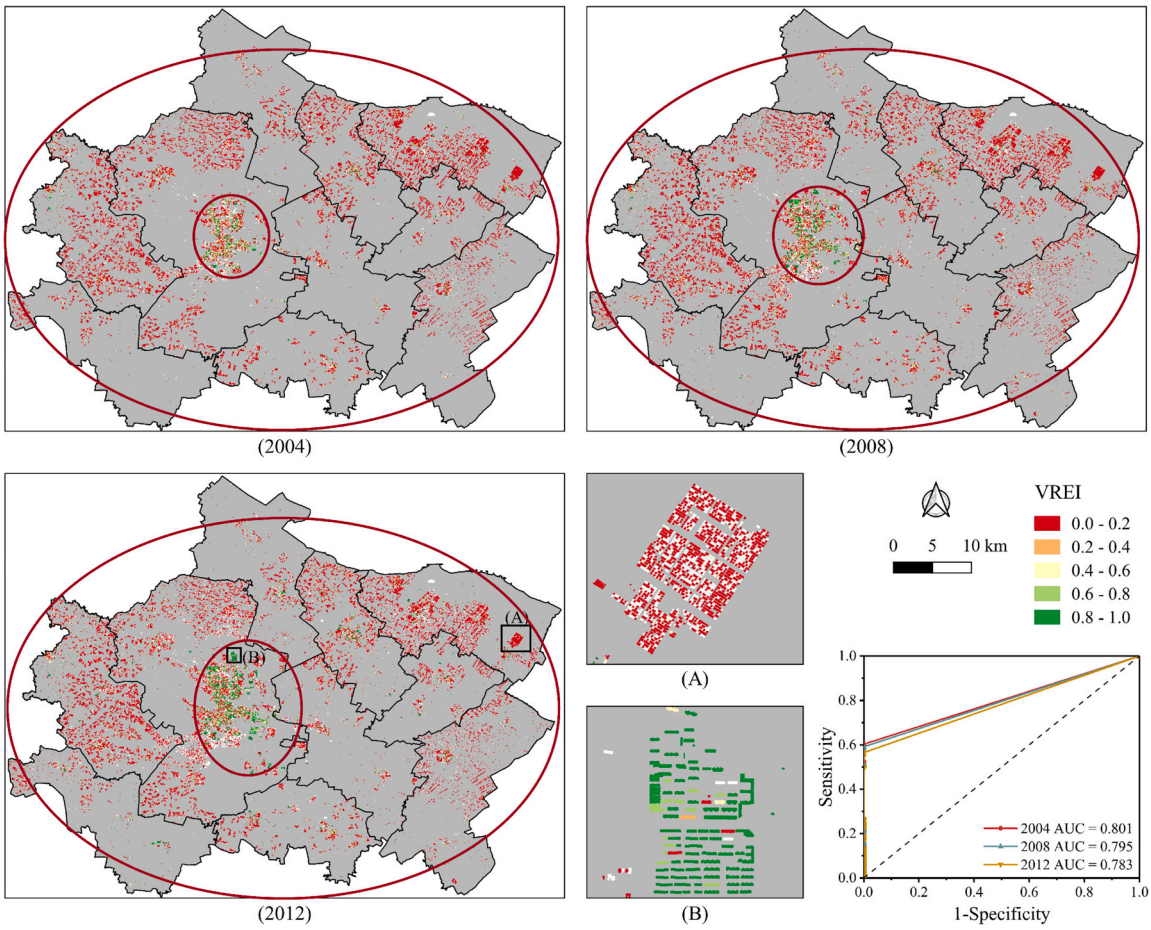


Fig. 9. Spatial distributions of vacant residence mixtures at the building level in 2004, 2008 and 2012 in Changshu. The VREI level was split into equal intervals and increased with increasing VREI values. The yearly red ellipses represent the central area and peripheral area of the city. Note that the VREI of the residential buildings colored white could not be measured because there were no vacant residences. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3.2. Temporal evolutions of vacant residence mixtures

The mixture of vacant residences at the building level in the central area of Changshu experienced positive changes throughout the study period. The regions that experienced positive changes in mixture of vacant residences also experienced expansions of those mixtures outward (Fig. 11b). These factors triggered an increase in heterogeneity in residential vacancies in the central area of the city and an expansion to its surrounding areas (Fig. 11a; please see Section 4.3.1). In addition, some residential buildings experienced negative changes in the mixture of residential vacancies throughout the study period. These buildings were scattered over the entire city (Fig. 11b).

Further, this study identified three patterns of changes in vacant residence mixtures at the building level: (1) the emergence of mixes (Fig. 11A), (2) the disappearance of mixes (Fig. 11B), and (3) an increase or decrease in mixes (Fig. 11C). For a residential building, the emergence of a mix implies that vacant residences in it began to exist at a certain time, but all residences in it were occupied or the building had not been constructed before that time. Contrary to the emergence of mixes, the disappearance of mixes implies that all residences in the building were occupied or the building was demolished at a certain time, but it had vacant residences before that time. An increase or decrease in mixes indicates that vacant residences have existed in the building throughout the study period, but the number and category of vacant residences at a certain time are different from those before that time.

The first observation is that the changes in most residential building mixtures were zero (i.e., they were unchanged) (Fig. 12). Specifically, for residential buildings with an emergence or disappearance of mixes, 87.2% (87.9% in 2008, 86.5% in 2012) and 96.3% (96.4% in 2008, 96.3% in 2012) of them exhibited changes towards zero, respectively. For residential buildings with an increase or decrease in mixes, 84.3% (85.4% in 2008, 83.1% in 2012) of them remained unchanged. Note that unlike the increases or decreases in mixtures, a change towards zero for the emergence and the disappearance of mixes has great significance for residential buildings, indicating the rise in residential vacancies and the extinction of residential vacancies.

We also observed that the decrease in and even the disappearance of mixtures outpaced the emergence of and increase in mixtures (Fig. 12). Specifically, the rate at which vacancy mixtures emerged in residential buildings in 2012 was much slower (0.7 times) than in 2008. This implies a sharp slowdown in the growth of residential buildings with newly vacant residences. The rate at which vacancy mixtures disappeared from residential buildings in 2012, however, was much faster (1.4 times) than in 2008. Although the rate of increase in the number of residential buildings with an increase in mixtures in 2012 was almost the same as in 2008, the rate of increase in the number of residential buildings with a decrease in mixtures in 2012 was much faster (1.5 times) than in 2008. This implies that the rate of growth in the number of residential buildings in which the number of vacant residences was decreasing, or in which the diversity in residential vacancies was reducing, was picking up.

In addition, in 2008, the number of residential buildings in which mixtures either emerged or increased was much greater (1.6 times) than the number of buildings in which mixtures decreased or even disappeared, but in 2012, the former was much lower than the latter (0.8

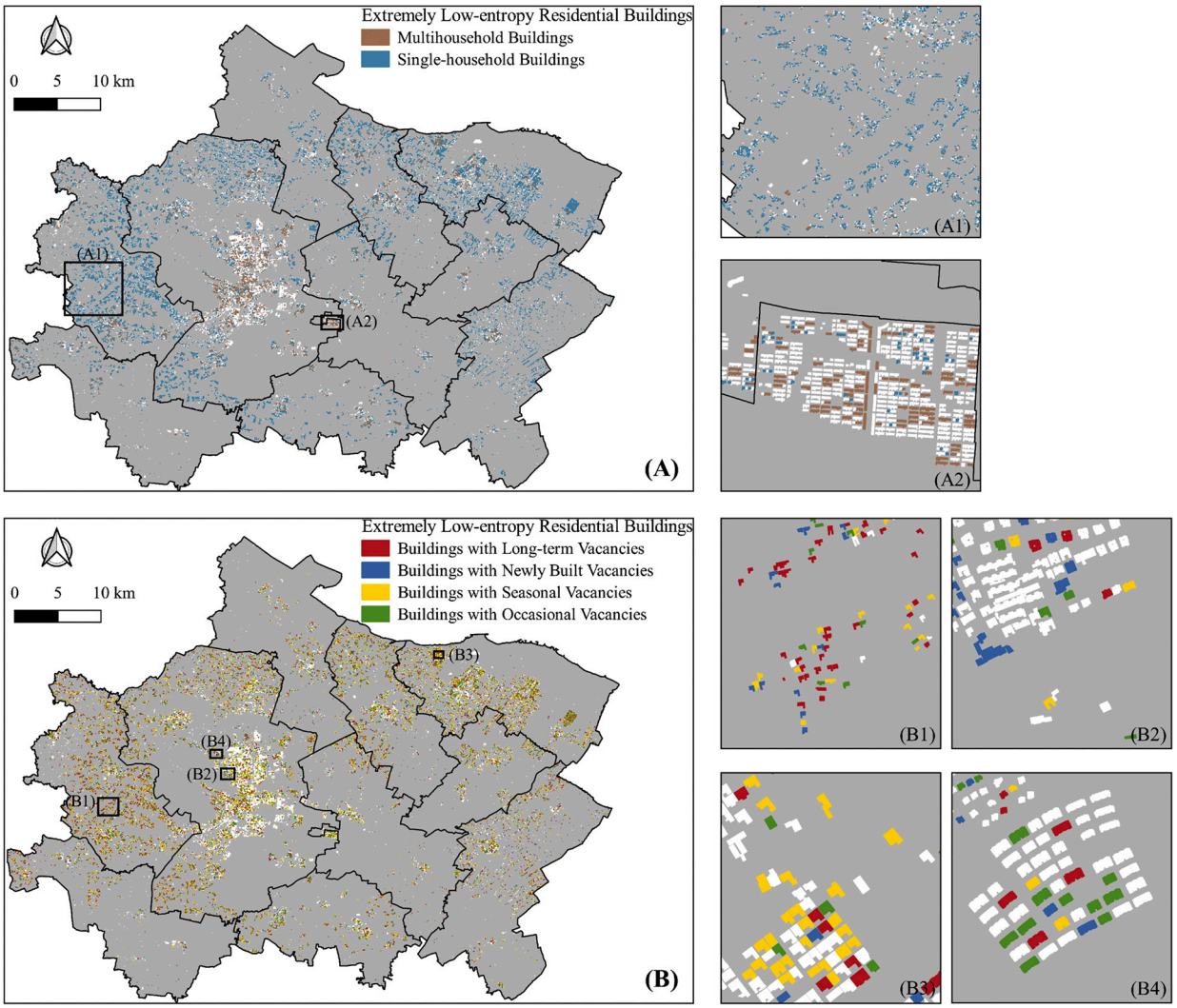


Fig. 10. Spatial distributions of extremely low-entropy residential buildings in 2012 in Changshu from the perspectives of the residential building types (A) and the single category of vacant residences contained in (B). Note that the residential buildings colored white may have been the buildings in which no vacant residences were contained and the VREI of which could not be measured, or the buildings in which some vacant residences were contained and the VREI of which were more than zero.

times) (Fig. 12). This is entirely predictable and the result of mixtures decreasing and even disappearing at an even faster rate than they emerge and increase.

It was also noted that residential buildings in which mixtures emerged and disappeared had VREI changes in the range of 0.4 to 1.0 and − 1.0 to −0.4, and those in which mixtures increased or decreased were in the −0.6-0 and 0–0.6 range (Fig. 12). These findings suggest the presence of both a sudden rise and a sudden decline in various categories of vacant residences in buildings. These residential buildings mainly lie in spatially complex and socioeconomically unstable areas, making it possible for diversity in vacant residence at the building level to appear and disappear abruptly. Additionally, these findings provide convincing evidence that many residential buildings are in reasonably socioeconomically stable areas, leading to sluggish and gradual changes in vacant residence diversity at the building level.

5. Discussion

5.1. Seasonally vacant residences related to tourism and seasonal industries

As a famous historical and cultural city in China, the International Garden City and the first International Wetland City in the world, Changshu had 11 special tourism developments by the end of 2013, as well as more than 70 scenic spots certifications, including one 5A[1] and three 4A certifications, according to the Changshu Statistical Yearbook (2004–2013). The tourism sector in Changshu underwent impressive growth, with an increase in tourist visits of 10.7% year on year, from 5.25 million in 2004 to 17.34 million in 2013, as well as a nearly 23% year-on-year increase in tourism income, from 3.3 billion to over 24 billion yuan. The spillover effects of tourism may generate negative externalities in the housing market. For instance, many people have bought houses in cities with attractive tourism resources for vacation. Almost all these houses are occupied during the peak tourism seasons but remain vacant outside of the tourism seasons (Chi et al., 2015; Wen, Lv, & Liu, 2011). This study presents further empirical evidence on the relationship between seasonally vacant residences and tourism (see Fig. 13). The Pearson coefficients of 0.90 and 0.80 suggest strong, positive linear correlations between seasonally vacant residences and tourist visits as well as tourism income that are statistically significant ($p < 0.01$). The Spearman coefficient on both relationships of 0.85, which is statistically significant ($p < 0.01$), indicates a significant increasing monotonic trend in those relationships.

Apart from tourism, residential vacancies could be immediately

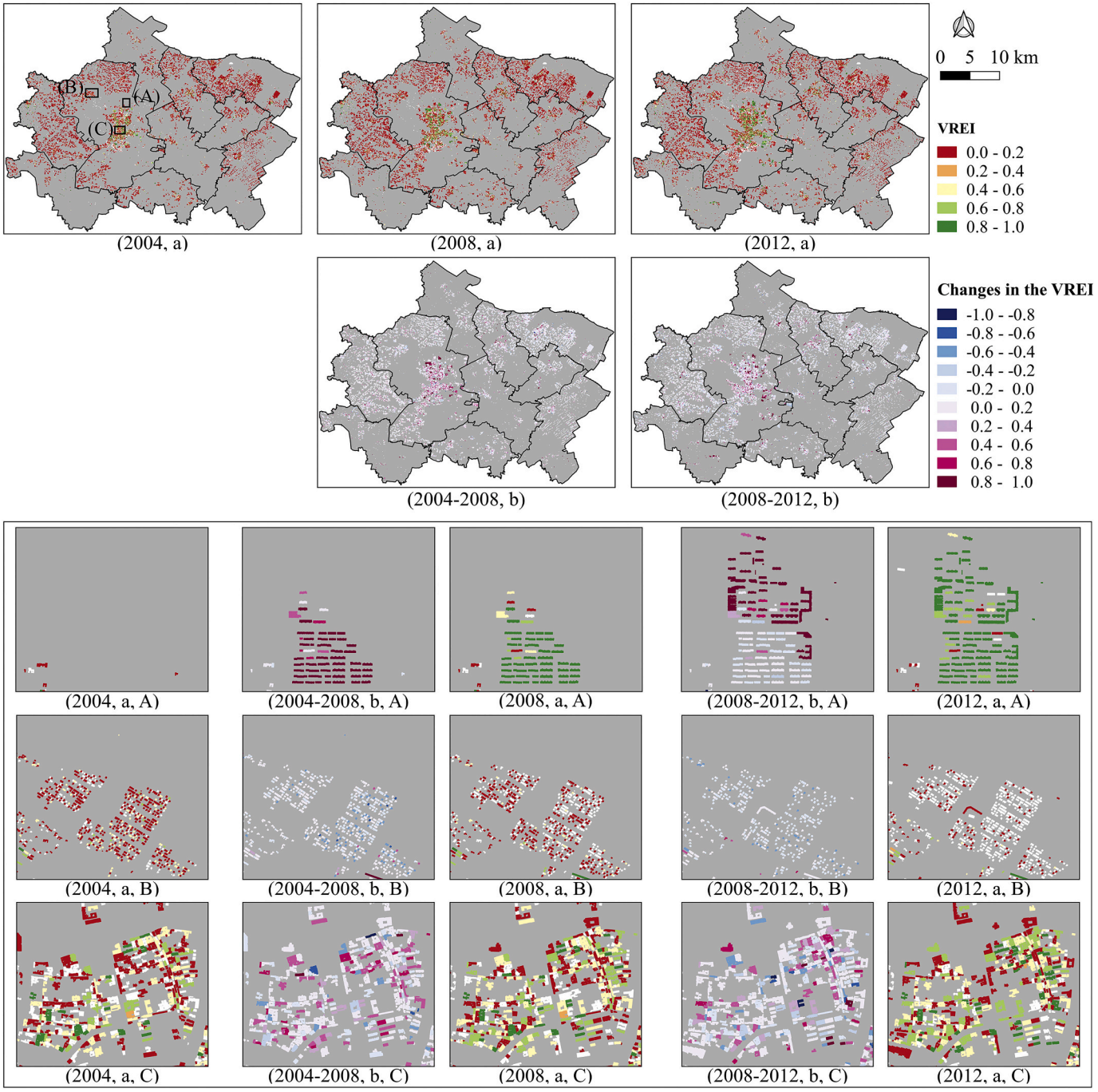


Fig. 11. Spatial distributions of changes in the VREI (b) for vacant residence mixtures at the building level (a) in 2004, 2008 and 2012 in Changshu, where a cooler color (i.e., a negative value) indicates a larger decrease in residential vacancy mixtures and a warmer color (i.e., a positive value) indicates a larger increase: (A) the emergence of the VREI, (B) the disappearance of the VREI, and (C) an increase or decrease in the VREI.

affected by seasonal industries. Only during certain seasons or months in the year do seasonal industries manufacture products by employing a considerable amount of migratory labor. During these times, the surrounding houses are occupied by the migrant laborers. Once the industries suspend their production, people are forced to migrate to other locations to survive, and a large number of vacant residences are left behind (Wen et al., 2011). As one of the top ten industrial counties or county-level cities of China (China Academy of Information and Communications Technology, 2020; National Academy of Economic Strategy, 2020), sharply industrialized Changshu (Li, Long, & Liu, 2010; Zhou, Zhang, Ye, Wang, & Su, 2016) has many seasonal industries, including those related to seasonal materials (e.g., sugar production in Fig. 14A) and those whose working conditions are affected by the climate (e.g., brick and tile manufacturing in Fig. 14B). Interestingly, the combination of seasonal industries in Changshu and seasonally vacant residences (colored yellow in Fig. 6) reveals that at least two types of seasonally vacant residences (or seasonal industries) were in Changshu, but the difference between the number of seasonally vacant residences currently vacant (or seasonal industries out of production) and occupied seasonally vacant residences (or seasonal industries in production) was slight each month. Additionally, some adjustments to both the regional industrial structure and industrial production modes might have occurred in Changshu, such as lengthening production times and increases in industries with relatively long periods of production relative

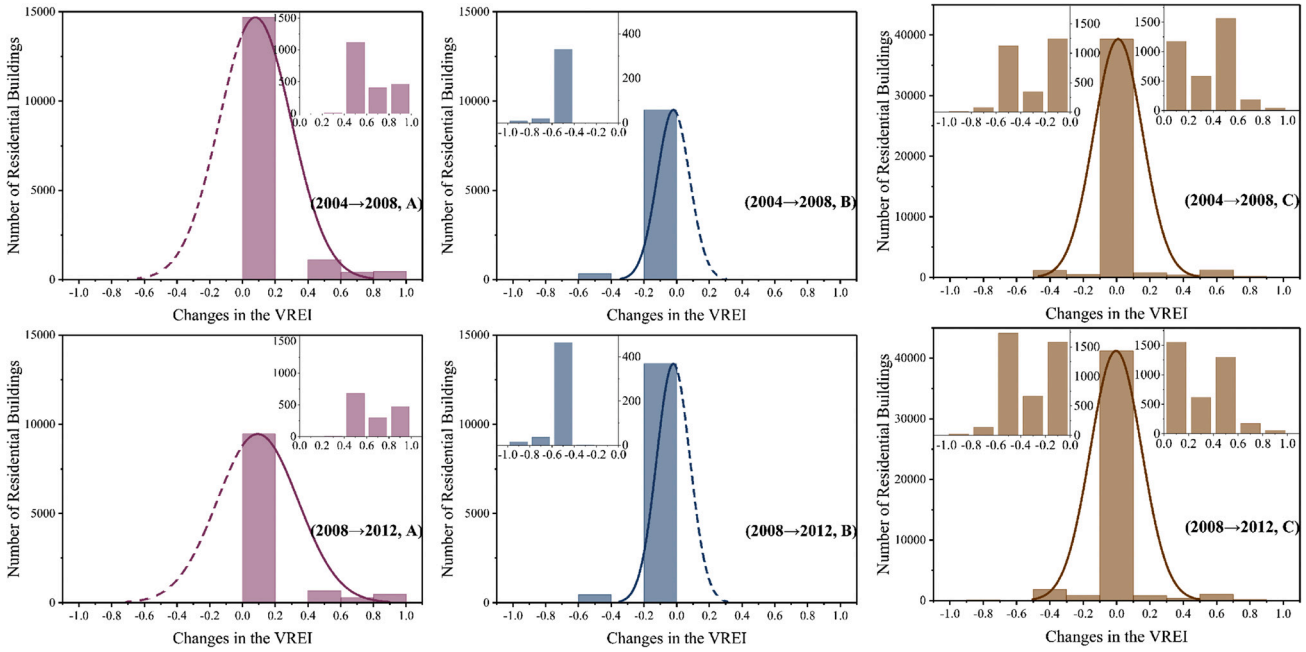


Fig. 12. Summary statistics and the normal distribution of changes in the VREI of vacant residence mixtures at the building level in 2004, 2008 and 2012 in Changshu: (A) the emergence of the VREI, (B) the disappearance of the VREI, and (C) an increase or decrease in the VREI.

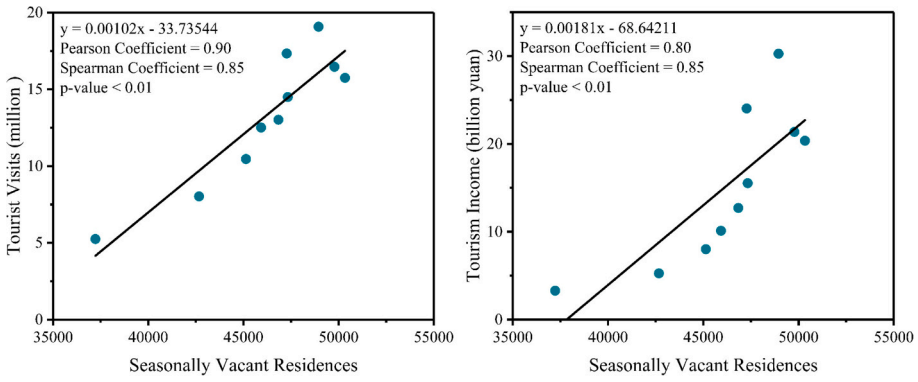


Fig. 13. The relationships between seasonally vacant residences and tourist visits as well as tourism income.

to those with shorter periods of production. These factors make it possible to reduce the extent of vacancy in seasonally vacant residences each year.

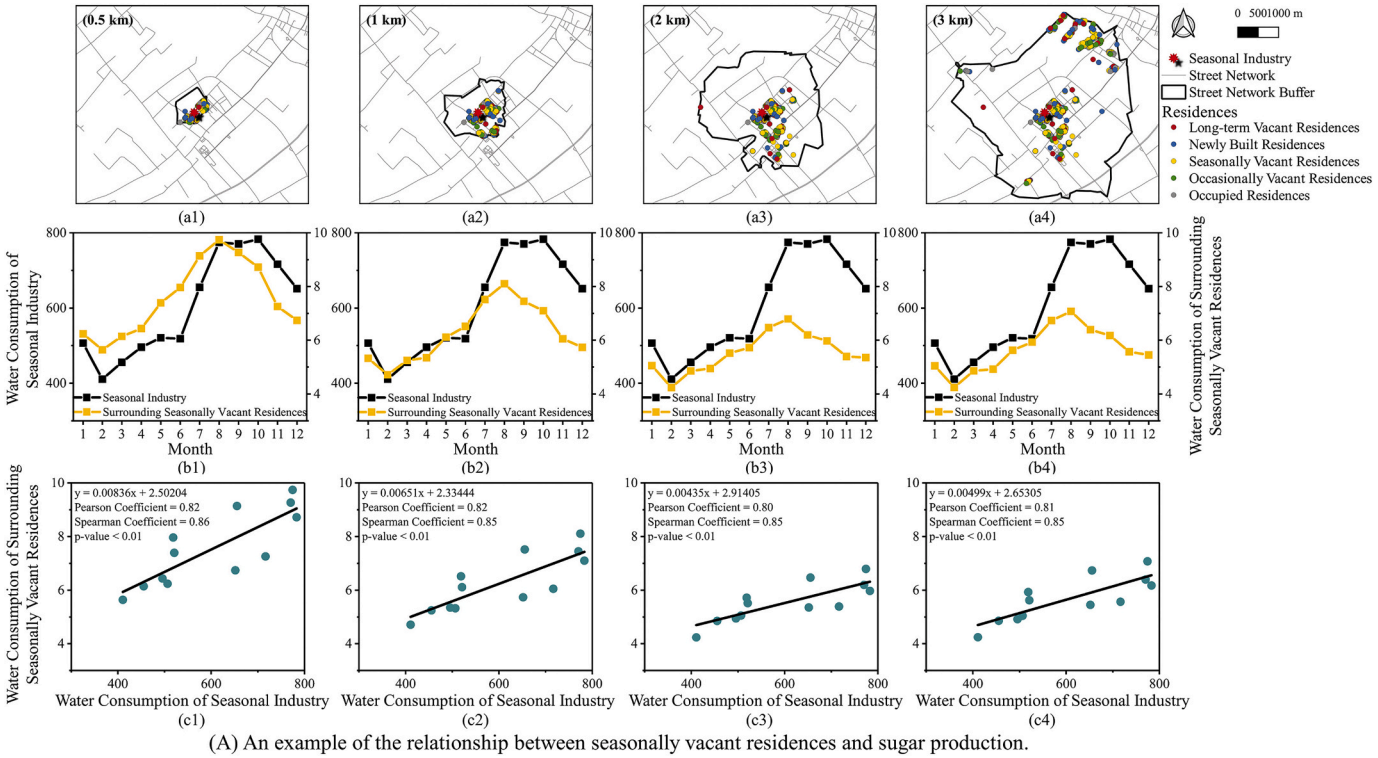
5.2. *Matthew effect in residential vacancy*

Fig. 6 shows that there is a "Matthew effect" (Bol, de Vaan, & van de Rijt, 2018; Merton, 1968) in residential vacancies: the extent of vacancy among short-term residential vacancies (i.e., newly built residences and occasionally vacant residences) may be gradually reducing, while the extent of vacancy among long-term residential vacancies is more likely to be increasing, leading to a wider gap between these types of vacancies. The mitigation of short-term residential vacancies is absolutely heartening news. This happens, unsurprisingly, because currently vacant newly built residences or occasionally vacant residences become occupied relatively quickly (Molloy, 2016). However, deteriorating residences with long-term vacancies are the bane of decision-makers. The idea behind the Matthew effect in this context is that long-term vacant residences will continue to be vacant to a greater extent. This happens in this fashion because of the properties and neighborhoods. On the one hand, long-term vacant residences assigned to transition from an in-use residence to a demolished residence (Huuhka, 2016) are highly correlated with housing market distress and are most likely to reflect excess supply (Molloy, 2016). On the other hand, the BWT (Keizer, Lindenberg, & Steg, 2008; Wilson & Kelling, 1982) suggests that neighborhood disorder (particularly physical disorder, e.g., long-term vacant residences) triggers residential fear and retreat from neighborhoods. The informal social control mechanisms are broken, and the signs of uncontrolled and uncontrollable neighborhoods are evident to the remaining residents and to any outsider, setting in motion the spread and worsening of long-term vacant residences.

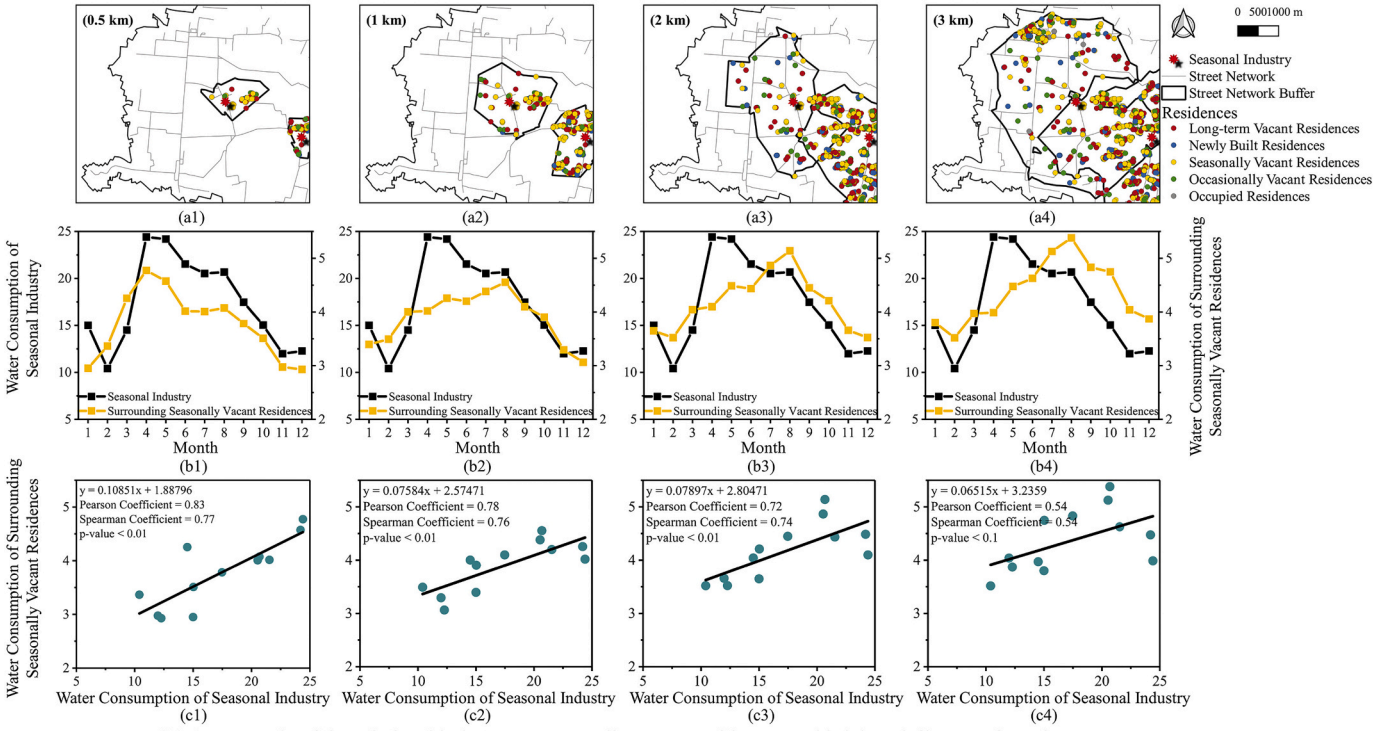
Therefore, this study delivers a clear and important message to decision-makers for mitigating vacancies in long-term vacant residences. First, making an early diagnosis of and intervening in long-term vacant residences are of vital importance when fighting the spread and worsening of such vacancies. Second, to break the cycle, signs of disorder need to be removed (e.g., long-term vacant residences should be demolished), while signs of order ought to be established (e.g., cleaning and providing abundant facilities and services) (Keizer et al., 2008; Pan et al., 2020).

5.3. *Academic contributions and policy implications*

These findings, although some are unique to Changshu, provide



(A) An example of the relationship between seasonally vacant residences and sugar production.



(B) An example of the relationship between seasonally vacant residences and brick and tile manufacturing.

Fig. 14. Examples of the relationships between seasonally vacant residences and seasonal industries: (a) spatial distributions and (b) aggregated 12-month water consumption time series for a certain seasonal industry and for seasonally vacant residences within 0.5 km, 1 km, 2 km, and 3 km street network buffers around the industry, and (c) the correlations between these measures.

interesting and novel lessons regarding what categories of vacant residences at the household level exist in China, what characteristics each one has, and how the residential vacancy mixtures at the building level can change. Unlike existing studies investigating the vacancy categories of entire cities, this study focuses on variability in and mixtures among residential vacancy categories in smaller spatial units within the city (specifically, households and buildings). Some of the results, including the identified categories of vacant residences at the household level and their characteristics, the measured degree of mixture in residential vacancies at the building level and their changes, can be further available microdata to support for more intelligently managing urban living environments, such as small-area population and/or housing prices

estimation, and neighborhood safety assessment.

Another crucial contribution of this study is the feasible and general-purpose framework for providing innovative insights into the variability in and mixtures among residential vacancies at two granular levels using municipal water consumption data. Given that municipal water consumption data are ubiquitous nationwide and worldwide and readily available, especially for decision-makers, the proposed framework is practical when generalizing to other cities and towns including small or developing ones, especially ones with comprehensive municipal water service coverage. Furthermore, considering the methods (e.g., Doc2Vec, K-means++, and the VREI) are almost fully automatic and easily implemented, the proposed framework is less time-consuming and labour-intensive than field survey. This serves the needs of the smart city.

In regards to policy implications, this study suggests two key points. First, this study helps decision-makers identify timely the categories of existing even emerging vacant residences in the absence of any prior knowledge. In addition to leading to a better understanding of the nature of residences, intelligently identifying the categories of vacant residences could reduce the cost and workload of investigation (e.g., a door-to-door field survey).

Second, this study could be of benefit to the government by concentrating its limited resources on specific categories of vacant residences to maximize its interventions and investments. The customized policies involved long-term vacant residences and seasonally vacant residences project to be the future wave of managing residences, while those related to newly built residences and occasionally vacant residences might be not pursued. For areas dominated by long-term vacant residences (e.g., extremely low-entropy multihousehold buildings with long-term vacancies), demolition policies and several recommendations for improving the neighborhoods where such residences are (e.g., enhancing the diversity of facilities and services) could be at decision-makers' disposal. For areas where abundant seasonally vacant residences are concentrated (e.g., extremely low-entropy multihousehold buildings with seasonal vacancies), perhaps adjusting the surrounding seasonal industrial structure would allow these residences to be occupied by different migrant laborers at different times, thus contributing to mitigating seasonal vacancies. Besides, when planning future industrial parks, the diversification of industrial structure should be under consideration to reduce the extent of vacancy in seasonally vacant residences even avoid their emergence.

5.4. Potential biases and future research

Despite the aforementioned informative and invaluable findings, several limitations have not yet been overcome and will be highlighted in our future research. In addition to damage to or the insensitivity of water meters, municipal water consumption data may be biased due to self-supplied residential water consumption (e.g., mineral water or well water). It would be beneficial if future research were to integrate multisource municipal infrastructure and service data to analyze residential vacancies. Additionally, it would certainly be beneficial if future research incorporated additional Chinese cities and then explored the similarities and differences among various types of cities, providing a broader understanding of the residential vacancies embedded in the urbanization of China.

6. Conclusions

The purpose of this study goes beyond revealing the variability in residential vacancies at the household level. Further analyses were conducted to depict the mixture of residential vacancies at the building level. To this end, a framework for analyzing vacant residences at granular levels via municipal water consumption data was proposed. First, the combination of Doc2Vec and K-means++ was used to identify different categories of residential vacancies and further advance our knowledge about the vacancy patterns among the various categories of vacant residences. Second, an entropy index was employed to assess the mixture of residential vacancies and further enable an understanding of the spatiotemporal characteristics of vacant residence mixtures.

The notably high SI and quite low DBI suggested that four categories of vacant residences at the household level was optimal, including seasonally vacant residences related to tourism and seasonal industries, as well as long-term vacant residences, newly built residences and occasionally vacant residences, which exhibit a Matthew effect. From 2004 through 2013 in Changshu, the vacant residence mixtures at the building level presented significant and intensifying spatial clustering (Moran's I from 0.30 to 0.46, z-score from 376.70 to 542.37, $p < 0.001$), a binary structure related to residential building types, and three patterns of changes in the mixtures. Interestingly, the decrease in and even disappearance of mixtures outpaced the emergence of and increase in mixtures. On this basis, some particular vacant residence patterns at the building level were identified, such as extremely low-entropy multihousehold buildings with long-term (or seasonal) vacancies. These findings fill the gap in the residential vacancy analysis literature, in which analyses ignore the variability in residential vacancies, as well as vacancy mixtures, at a granular level. The substantial insights from this study can assist in rethinking residential vacancy, redeveloping policies to combat vacancy, and ensuring the sustainable development of the new-type urbanization in China.

Note

The 5A (AAAAA) scenic spot level represents the highest certification for tourist scenic spots granted by the Ministry of Culture and Tourism of the People's Republic of China, followed by certification levels of 4A, 3A, 2A, and a

Author statement

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. Furthermore, each author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before its appearance in the *Computers, Environment and Urban Systems*.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 41671408, 41801306, 41901332), the National Key R&D Program of China (Grant No. 2019YFB2102903), and Special Fund for Foundation and Frontier of Applications of Wuhan (Grant No. 2018010401011293).

References

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 1027–1035).

Bai, X., Shi, P., & Liu, Y. (2014). Society: Realizing China's urban dream. *Nature, 509* (7499), 158–160. https://doi.org/10.1038/509158a.

Benabderrahmane, S., Mellouli, N., & Lamolle, M. (2018). On the predictive analysis of behavioral massive job data using embedded clustering and deep recurrent neural networks. *Knowledge-Based Systems, 151*, 95–113. https://doi.org/10.1016/j.knosys.2018.03.025.

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences of the United States of America, 115* (19), 4887–4890. https://doi.org/10.1073/pnas.1719557115.

Chang, W., Xu, Z., Zhou, S., & Cao, W. (2018). Research on detection methods based on Doc2vec abnormal comments. *Future Generation Computer Systems - The International Journal of EScience, 86*, 656–662. https://doi.org/10.1016/j.future.2018.04.059.

Chen, M., Liu, W., & Lu, D. (2016). Challenges and the way forward in China's new-type urbanization. *Land Use Policy, 55*, 334–339. https://doi.org/10.1016/j.landusepol.2015.07.025.

Chi, G., Liu, Y., Wu, Z., & Wu, H. (2015). Ghost cities analysis based on positioning data in China. *ArXiv Preprint ArXiv, 1510*, Article 08505.

China Academy of Information and Communications Technology. (2020). Development of Top hundred industrial counties or county-level cities in China (2020). Retrieved from http://www.caict.ac.cn/kxyj/qwfb/bps/202011/P020201211777084001694.pdf.

Cho, S. E., & Kim, S. (2017). Measuring urban diversity of Songjiang new town: A re-configuration of a Chinese suburb. *Habitat International, 66*, 32–41. https://doi.org/10.1016/j.habitatint.2017.05.009.

Cox, N. J. (2019). Stata tip 133: Box plots that show median and quartiles only. *Stata Journal, 19*(4), 1009–1014. https://doi.org/10.1177/1536867X19893643.

Du, M., Wang, L., Zou, S., & Shi, C. (2018). Modeling the census tract level housing vacancy rate with the Jilin1-03 satellite and other geospatial data. *Remote Sensing, 10*(12), 1920. https://doi.org/10.3390/rs10121920.

Feng, W., Liu, Y., & Qu, L. (2019). Effect of land-centered urbanization on rural development: A regional analysis in China. *Land Use Policy, 87*, 104072. https://doi.org/10.1016/j.landusepol.2019.104072.

Fernando, V. (2010). There are now enough vacant properties in China to house over half of America. Retrieved from http://www.businessinsider.com/there-are-now-enough-vacant-properties-in-china-to-house-over-half-of-america-2010-9.

France, S. L., Carroll, J. D., & Xiong, H. (2012). Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. *Information Sciences, 184*(1), 92–110. https://doi.org/10.1016/j.ins.2011.07.048.

Gašparović, M., Zrinjski, M., & Gudelj, M. (2019). Automatic cost-effective method for land cover classification (ALCC). *Computers, Environment and Urban Systems, 76*, 1–10. https://doi.org/10.1016/j.compenvurbsys.2019.03.001.

Ghodousi, M., Alesheikh, A. A., & Saeidian, B. (2016). Analyzing public participant data to evaluate citizen satisfaction and to prioritize their needs via K-means, FCM and ICA. *Cities, 55*, 70–81. https://doi.org/10.1016/j.cities.2016.03.015.

Guan, Q., Cheng, S., Pan, Y., Yao, Y., & Zeng, W. (2020). Sensing mixed urban land-use patterns using municipal water consumption time series. *Annals of the American Association of Geographers*, 1–19. https://doi.org/10.1080/24694452.2020.1769463.

Haramati, T., & Hananel, R. (2016). Is anybody home? The influence of ghost apartments on urban diversity in Tel-Aviv and Jerusalem. *Cities, 56*, 109–118. https://doi.org/10.1016/j.cities.2016.04.006.

He, G., Mol, A. P. J., & Lu, Y. (2016). Wasted cities in urbanizing China. *Environment and Development, 18*, 2–13. https://doi.org/10.1016/j.envdev.2015.12.003.

Hincks, S., Kingston, R., Webb, B., & Wong, C. (2018). A new geodemographic classification of commuting flows for England and Wales. *International Journal of Geographical Information Science, 32*(4), 663–684. https://doi.org/10.1080/13658816.2017.1407416.

Hipp, J. R., Kane, K., & Kim, J. H. (2017). Recipes for neighborhood development: A machine learning approach toward understanding the impact of mixing in neighborhoods. *Landscape and Urban Planning, 164*, 1–12. https://doi.org/10.1016/j.landurbplan.2017.03.006.

Huuhka, S. (2016). Vacant residential buildings as potential reserves: A geographical and statistical study. *Building Research and Information, 44*(8), 816–839. https://doi.org/10.1080/09613218.2016.1107316.

Jiang, Y., Mohabir, N., Ma, R., & Zhu, P. (2017). Sorting through neoliberal variations of ghost cities in China. *Land Use Policy, 69*, 445–453. https://doi.org/10.1016/j.landusepol.2017.09.001.

Jin, X., Long, Y., Sun, W., Lu, Y., Yang, X., & Tang, J. (2017). Evaluating cities' vitality and identifying ghost cities in China with emerging geographical data. *Cities, 63*, 98–109. https://doi.org/10.1016/j.cities.2017.01.002.

Jurafsky, D. S., & Martin, J. H. (2010). *Speech and language processing* (Prentice Hall).

Keizer, K., Lindenberg, S., & Steg, L. (2008). The spreading of disorder. *Science, 322*(5908), 1681–1685. https://doi.org/10.1126/science.1161405.

Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences, 477*, 15–29. https://doi.org/10.1016/j.ins.2018.10.006.

Kim, G., Newman, G., & Jiang, B. (2020). Urban regeneration: Community engagement process for vacant land in declining cities. *Cities, 102*, 102730. https://doi.org/10.1016/j.cities.2020.102730.

Kim, H. K., Kim, H., & Cho, S. (2017). Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing, 266*, 336–352. https://doi.org/10.1016/j.neucom.2017.05.046.

Kumagai, K., Matsuda, Y., & Ono, Y. (2016). Estimation of housing vacancy distributions: Basic Bayesian approach using utility data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 41*(B2), 709–713. https://doi.org/10.5194/isprsarchives-XLI-B2-709-2016.

Lang, W., Long, Y., & Chen, T. (2018). Rediscovering Chinese cities through the lens of land-use patterns. *Land Use Policy, 79*(79), 362–374. https://doi.org/10.1016/j.landusepol.2018.08.031.

Lansley, G., & Longley, P. (2016). Deriving age and gender from forenames for consumer analytics. *Journal of Retailing and Consumer Services, 30*, 271–278. https://doi.org/10.1016/j.jretconser.2016.02.007.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning* (pp. 1188–1196).

Lee, S., & Kim, W. (2017). Sentiment labeling for extending initial labeled data to improve semi-supervised sentiment classification. *Electronic Commerce Research and Applications, 26*, 35–49. https://doi.org/10.1016/j.elerap.2017.09.006.

Leichtle, T., Lakes, T., Zhu, X., & Taubenbck, H. (2019). Has Dongying developed to a ghost city? - evidence from multi-temporal population estimation based on VHR remote sensing and census counts. In *, 78. Computers Environment and Urban Systems*. https://doi.org/10.1016/j.compenvurbsys.2019.101372.

Li, J., Guo, M., & Lo, K. (2019). Estimating housing vacancy rates in rural China using power consumption data. *Sustainability, 11*(20), 5722. https://doi.org/10.3390/su11205722.

Li, Y., Long, H., & Liu, Y. (2010). Industrial development and land use/cover change and their effects on local environment: A case study of Changshu in eastern coastal China. *Frontiers of Environmental Science & Engineering in China, 4*(4), 438–448. https://doi.org/10.1007/s11783-010-0273-3.

Long, H. (2014). Land use policy in China: Introduction. *Land Use Policy, 40*(40), 1–5. https://doi.org/10.1016/j.landusepol.2014.03.006.

Lord, E., Willems, M., Lapointe, F.-J., & Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. *Information Sciences, 393*, 29–46. https://doi.org/10.1016/j.ins.2017.02.010.

Lu, H., Zhang, C., Liu, G., Ye, X., & Miao, C. (2018). Mapping China's ghost cities through the combination of nighttime satellite data and daytime satellite data. *Remote Sensing, 10*(7), 1037. https://doi.org/10.3390/rs10071037.

Lu, Z., Im, J., Rhee, J., & Hodgson, M. (2014). Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. *Landscape and Urban Planning, 130*(1), 134–148. https://doi.org/10.1016/j.landurbplan.2014.07.005.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv, 1301*, 3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *ArXiv Preprint ArXiv, 1310*(4546), 3111–3119.

Molloy, R. (2016). Long-term vacant housing in the United States. *Regional Science and Urban Economics, 59*, 118–129. https://doi.org/10.1016/j.regsciurbeco.2016.06.002.

Montgomery, M. R. (2008). The urban transformation of the developing world. *Science, 319*(5864), 761–764. https://doi.org/10.1126/science.1153012.

Morckel, V. C. (2014). Spatial characteristics of housing abandonment. *Applied Geography, 48*, 8–16. https://doi.org/10.1016/j.apgeog.2014.01.001.

Motlagh, O., Berry, A., & O'Neil, L. (2019). Clustering of residential electricity customers using load time series. *Applied Energy, 237*, 11–24. https://doi.org/10.1016/j.apenergy.2018.12.063.

Mui, Y., Jonessmith, J. C., Thornton, R. L. J., Porter, K. P., & Gittelsohn, J. (2017). Relationships between vacant homes and food swamps: A longitudinal study of an urban food environment. *International Journal of Environmental Research and Public Health, 14*(11). https://doi.org/10.3390/ijerph14111426.

Nafi, K. W., Roy, B., Roy, C. K., & Schneider, K. A. (2020). A universal cross language software similarity detector for open source software categorization. *Journal of Systems and Software, 162*. https://doi.org/10.1016/j.jss.2019.110491.

Nam, J., Han, J., & Lee, C. (2016). Factors contributing to residential vacancy and some approaches to management in Gyeonggi Province, Korea. *Sustainability, 8*(4), 367. https://doi.org/10.3390/su8040367.

National Academy of Economic Strategy. (2020). Economic development of top hundred counties or county-level cities in China (2020). Retrieved from http://naes.cssn.cn/cj_zwz/cg/yjbg/zgxyjjfzbg/202012/t20201223_5235779.shtml.

Nazarnia, N., Harding, C., & Jaeger, J. A. G. (2019). How suitable is entropy as a measure of urban sprawl? *Landscape and Urban Planning, 184*, 32–43. https://doi.org/10.1016/j.landurbplan.2018.09.025.

Newman, G., Gu, D., Kim, J., & Li, W. (2016). Elasticity and urban vacancy: A longitudinal comparison of US cities. *Cities, 58*, 143–151. https://doi.org/10.1016/j.cities.2016.05.018.

Newman, G., Park, Y., & Lee, R. J. (2018). Vacant urban areas: Causes and interconnected factors. *Cities, 72*, 421–429. https://doi.org/10.1016/j.cities.2017.10.005.

Nie, X., & Liu, X. (2013). Types of "ghost towns" in the process of urbanization and countermeasures. *Journal of Nantong University (Social Science Edition), 4*, 111–117.

Pan, Y., Zeng, W., Guan, Q., Yao, Y., Liang, X., Yue, H., … Wang, J. (2020). Spatiotemporal dynamics and the contributing factors of residential vacancy at a fine scale: A perspective from municipal water consumption. *Cities, 103*, 102745. https://doi.org/10.1016/j.cities.2020.102745.

Pradhan, T., & Pal, S. (2020). A multi-level fusion based decision support system for academic collaborator recommendation. *Knowledge-Based Systems*, 105784. https://doi.org/10.1016/j.knosys.2020.105784.

Qiu, B. (2012). Complexity science and city transformation. *Urban Studies, 1*, 1–18.

Roth, J. J. (2019). Empty homes and acquisitive crime: Does vacancy type matter? *American Journal of Criminal Justice, 44*(5), 770–787. https://doi.org/10.1007/s12103-019-9469-7.

Shi, L., Wurm, M., Huang, X., Zhong, T., Leichtle, T., & Taubenböck, H. (2020). Urbanization that hides in the dark–spotting China's "ghost neighborhoods" from space. *Landscape and Urban Planning, 200*, 103822. https://doi.org/10.1016/j.landurbplan.2020.103822.

Silverman, R. M., Yin, L., & Patterson, K. L. (2013). Dawn of the dead city: An exploratory analysis of vacant addresses in buffalo, NY 2008–2010. *Journal of Urban Affairs, 35*(2), 131–152. https://doi.org/10.1111/j.1467-9906.2012.00627.x.

Stern, M., & Lester, T. W. (2021). Does local ownership of vacant land reduce crime? An assessment of Chicago's large lots program. *Journal of the American Planning Association, 87*(1), 73–84. https://doi.org/10.1080/01944363.2020.1792334.

Uddin, M. N., Islam, A. K. M. S., Bala, S. K., Islam, G. M. T., Adhikary, S., Saha, D., … Akter, R. (2019). Mapping of climate vulnerability of the coastal region of Bangladesh using principal component analysis. *Applied Geography, 102*, 47–57. https://doi.org/10.1016/j.apgeog.2018.12.011.

Waldron, R., O'Donoghue-Hynes, B., & Redmond, D. (2019). Emergency homeless shelter use in the Dublin region 2012–2016: Utilizing a cluster analysis of administrative data. *Cities, 94*, 143–152. https://doi.org/10.1016/j.cities.2019.06.008.

Wang, C., Li, Y., Myint, S. W., Zhao, Q., & Wentz, E. A. (2019). Impacts of spatial clustering of urban land cover on land surface temperature across Köppen climate zones in the contiguous United States. *Landscape and Urban Planning*, 192. https://doi.org/10.1016/j.landurbplan.2019.103668.

Wen, J., Lv, H., & Liu, B. (2011). Definitions and statistics of American house vacancy rate - statistical research report on foreign house vacancy rate (II). *China Statistics, 1*, 44–47.

Williams, S., Xu, W., Tan, S. B., Foster, M. J., & Chen, C. (2019). Ghost cities of China: Identifying urban vacancy through social media data. *Cities, 94*, 275–285. https://doi.org/10.1016/j.cities.2019.05.006.

Wilson, J. Q., & Kelling, G. L. (1982). Broken windows. *Atlantic Monthly, 249*(3), 29–38.

Xia, H., Karimi, H. A., & Meng, L. (2017). Parallel implementation of Kaufman's initialization for clustering large remote sensing images on clouds. *Computers, Environment and Urban Systems, 61*, 153–162. https://doi.org/10.1016/j.compenvurbsys.2014.06.002.

Xu, L., Du, Z., Mao, R., Zhang, F., & Liu, R. (2020). GSAM: A deep neural network model for extracting computational representations of Chinese addresses fused with geospatial feature. *Computers, Environment and Urban Systems, 81*, 101473. https://doi.org/10.1016/j.compenvurbsys.2020.101473.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science, 31*(4), 825–848. https://doi.org/10.1080/13658816.2016.1244608.

Yoo, H., & Kwon, Y. (2019). Different factors affecting vacant housing according to regional characteristics in South Korea. *Sustainability, 11*(24), 6913. https://doi.org/10.3390/su11246913.

Zhang, X., & Li, H. (2020). The evolving process of the land urbanization bubble: Evidence from Hangzhou, China. *Cities, 102*, 102724. https://doi.org/10.1016/j.cities.2020.102724.

Zheng, H., Wang, X., & Cao, S. (2014). The land finance model jeopardizes China's sustainable development. *Habitat International, 44*, 130–136. https://doi.org/10.1016/j.habitatint.2014.05.008.

Zheng, Q., Zeng, Y., Deng, J., Wang, K., Jiang, R., & Ye, Z. (2017). "Ghost cities" identification using multi-source remote sensing datasets: A case study in Yangtze River Delta. *Applied Geography, 80*, 112–121. https://doi.org/10.1016/j.apgeog.2017.02.004.

Zhou, R., Zhang, H., Ye, X., Wang, X., & Su, H. (2016). The delimitation of urban growth boundaries using the CLUE-S land-use change model: Study on Xinzhuang town, Changshu City, China. *Sustainability, 8*(11), 1182. https://doi.org/10.3390/su8111182.

Zou, S., & Wang, L. (2019). Individual vacant house detection in very-high-resolution remote sensing images. *Annals of the American Association of Geographers, 110*(2), 449–461. https://doi.org/10.1080/24694452.2019.1665492.