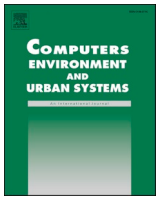
Computers, Environment and Urban Systems 90 (2021) 101709

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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



Discovering income-economic segregation patterns: A residential-mobility embedding approach

Tong Zhang [a,*], Xiaoqi Duan [a], David W.S. Wong [b], Yashan Lu [a]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[b] Department of Geography & Geoinformation Science, George Mason University, Fairfax, VA 22030, USA

ARTICLE INFO

Keywords:
Income
Segregation
Isolation-exposure
Graph embedding
Community type

ABSTRACT

As most studies of segregation rely on the evenness dimension, this current study proposes a graph embedding approach to explore the usefulness of employing the isolation-exposure dimension to evaluate income segregation. While most segregation studies analyzed the static distribution of population subgroups, current study attempts to classify neighborhoods based on house value as a proxy of income, residents' exposure to people of different income levels as constrained by their mobility patterns, and amenities available in the neighborhood. This study exploits the graph embedding method to classify neighborhoods by combining their various attributes, static population distribution and mobility data provided by smart cards to analyze income segregation in Shenzhen, China. Results identify four types of communities with different economic statuses, mobility patterns, and amenity characteristics. They provide rich descriptions about the connections between income segregation patterns, population dynamics, and neighborhood characteristics. The study found that the more segregated communities, which are composed of the poorest and richest groups, are mostly in the peripheral regions of the city while the inner city has lower levels of segregation, mainly due to differentials in transit accessibility. The study demonstrates the great potential of the proposed method to incorporate multiple aspects to evaluate segregation.

1. Introduction

Segregation, which may be broadly described as the spatial separation between different population groups, can be characterized as a multi-contextual and multi-faceted phenomenon (Oka & Wong, 2019). It can be cross-classified by different socio-geographical spaces (e.g., residential, school, work, entertainment, etc.) and population subgroups defined by various socio-demographic characteristics. Among these multiple contextual and social settings, racial-ethnic segregation in the residential space is among the most heavily studied (Massey & Denton, 1988). While "income segregation" is not a frequent term in the literature, its source, "income inequality," has been intensely scrutinized.

Income inequality concerns the disparities in income level among individuals, households-families or population groups, while income segregation refers to the spatial separation of people with different income levels. The causal connections between income inequality and income segregation are complicated and have been approached from multiple perspectives, including health, education, residential structure, and racial inequality (e.g., Kawachi & Kennedy, 1999; Mayer, 2001; Pickett & Wilkinson, 2015; Quillian, 2012; Watson, 2009). Nevertheless, causal relationships between them are likely dependent on the particular societal structure (Reardon & Bischoff, 2011). Despite developing nations face many types of inequality, including health and health care (Makinen et al., 2000; Wagstaff & Doorslaer, 2000) and income (Ravallion, 2014), most studies of income and economic segregation focused on developed nations. Studies of income segregation in developing nations such as China are scarce (Xie & Zhou, 2014).

Many recent studies of measuring segregation are still framed by the five dimensions of segregation: evenness, exposure-isolation, concentration, centralization, and clustering (Massey & Denton, 1988). The well-known Gini coefficient (Gini, 1921), which is an evenness measure, offers a summary measure of income inequality, but is insufficient to capture the nuance income segregation patterns over space and time. Subsequent modifications such as the generalized neighborhood sorting index (Jargowsky & Kim, 2004) attempt to augment the power of Gini coefficient in discerning spatial patterns. Nevertheless, these measures are based on the evenness dimension. Other dimensions, such as isolation-exposure have not been thoroughly explored in studying

* Corresponding author.
E-mail address: zhangt@whu.edu.cn (T. Zhang).

https://doi.org/10.1016/j.compenvurbsys.2021.101709
Received 11 March 2021; Received in revised form 12 August 2021; Accepted 16 August 2021
Available online 25 August 2021

income segregation.

On the other hand, studies of income segregation mostly focus on the residential space, reducing segregation to purely the static distribution of different population subgroups across neighborhoods (Rey & Folch, 2011). These studies conceptualize segregation as the spatial patterns of disparities, implicitly assume that population subgroups in different residential areas are completely static, neglecting the possible social interactions among these subgroups in other socio-geographical spaces, such as work, school and culture (Wong & Shaw, 2011). In the context of income segregation, even if poor and rich people reside in different parts of a city, the interactions between them, if any, should be considered when evaluating segregation of all income groups. The current static treatment of population distribution used in measuring income segregation also fails to exploit the voluminous human mobility data that may reflect the interactions among population groups.

Studies measuring segregation typically focus exclusively on one population characteristic of concern (e.g., race-ethnicity, occupation, income), putting neighborhood characteristics aside. Study results show that some (sub) areas are more segregated than others, but should it be a concern? The general concern about segregation is that people are separated and unequal. However, traditional segregation measures cannot reflect how unequal that people may experience in different neighborhood environments. The proposed approach in this study attempts to consider neighborhood characteristics.

This article is framed around the objectives to tackle the limitations of existing approaches in measuring segregation enumerated above. Instead of relying on evenness measures as in most segregation studies, this current study explores the usefulness of isolation-exposure dimension to measure income segregation by exploiting human mobility data that depict the potential spatial interaction among people with different income levels. Methodologically, this study employs graph embedding, a data mining-based visual analytics method, to combine static population and mobility data, and neighborhood characteristics to analyze segregation at fine-grained scales. The study uses data from Shenzhen, the fourth largest Chinese city to demonstrate the feasibility and utilities of the proposed approach and method. As income data are generally not available in China, we use house value as a proxy of income level in the current study. However, the proposed framework can use other indicators of economic status to evaluate segregation.

2. Limitations in the current practices of measuring segregation

In this section, we review relevant segregation studies in three areas: (1) Segregation encompassing multiple socio-geographical spaces, (2) segregation indices, and (3) data used in segregation studies. These reviews justify the needs of the proposed approach.

2.1. Segregation encompassing multiple socio-geographical spaces

Most segregation studies focused on the state of population in a single socio-geographical space (where people reside, work, learn, entertain, or socialize) within a given time frame. This tradition either is interested in the segregation condition in only a specific socio-geographical space (e.g., residential space) or assumes that the particular socio-geographical space is overwhelmingly important that situations in other spaces may be ignored. Thus, studies following this tradition are limited to assessing the distributions of population subgroups across areal units over a region using aspatial and spatial measures. This tradition also assumes a static view of the population distribution that segregation is determined by where people are within a specific socio-geographical space (say residential), and whether they interact through other socio-geographical spaces (say social or work) is irrelevant. Such approach in evaluating segregation ignores the potential importance of people's interaction across multiple spaces and undermines the roles of individual's activity space and mobility patterns in influencing one's segregation experience (Kwan, 2013). Conceptually, potential interaction among population groups can be reflected by the mobility patterns of individuals, and thus their activity spaces across all socio-geographical contexts should be accounted for in evaluating segregation.

Although Atkinson and Flint (2004) was likely the first proposing the use of time-space trajectories of individuals beyond their place of residence to evaluate segregation, studies on segregation relying on the activity space concept to include multiple socio-geographical spaces have gradually gained momentum over the past decade (Blumenstock & Fratamico, 2013; Farber, O'Kelly, Miller, & Neutens, 2015; Farber, Páez, & Morency, 2012; Kwan, 2013; Matthews & Yang, 2013; McQuoid & Dijst, 2012; Wang & Li, 2016; Wissink, Schwanen, & van Kempen, 2016; Wong & Shaw, 2011). Many of these studies investigated segregation under the framework of time geography, utilizing individual spatiotemporal trajectories to measure segregation. Some methods include comparing spatiotemporal trajectories (Atkinson & Flint, 2004), community-based random walk analysis (Dannemann, Sotomayor-Gómez, & Samaniego, 2018), developing single segregation measures (Wong & Shaw, 2011), and applying regression-based measures (Li & Wang, 2017). However, these studies considering multiple socio-geographical spaces focused on racial-ethnic segregation exclusively. Income segregation involving multiple socio-geographical spaces has been very much ignored.

Besides using activity-space as the framework to evaluate segregation, another recent direction in segregation study is the domain approach (e.g., van Ham & Tammaru, 2016). This approach is similar to the activity space approach that acknowledges segregation to be present in multiple socio-geographical spaces. But different from the activity space approach which often collapses multiple socio-geographical spaces in evaluating segregation, the domain approach focuses on the connections of segregation situations across domains through space and time, such as the connections of segregation between residential and economic (van Ham, Tammaru, de Vuijst, & Zwiers, 2016), work (Tammaru, Strömgren, van Ham, & Danzer, 2016), school (Boterman, Musterd, Pacchi, & Ranci, 2019), and entertainment (Kukk, van Ham, & Tammaru, 2019) spaces.

The domain approach advances segregation study not only by considering multiple socio-geographical spaces, but to assess relationships of socio-spatial processes among domains, providing a framework to understand segregation from a comprehensive societal perspective. However, the domain approach often requires rich individual-level data about the socioeconomic and demographic characteristics of individuals, beyond simple locations or travel trajectories of individuals. Such rich individual-level data are not available for the current study of Shenzhen, although the proposed method can easily accommodate the rich individual-level socioeconomic and demographic data.

2.2. Limitations of using indices

Traditional data sources for time geography rely on surveys as such travel diaries (Buliung & Kanaroglou, 2006; Li & Wang, 2017; Park & Kwan, 2018; Schönfelder & Axhausen, 2003; Wong & Shaw, 2011), which are usually small in size. Recent data depicting individual trajectories, such as cell phone data (Blumenstock & Fratamico, 2013; Dannemann et al., 2018; Järv, Müürisepp, Ahas, Derudder, & Witlox, 2015; Silm & Ahas, 2014), and location-based social network data (Phillips, Levy, Sampson, Small, & Wang, 2020; Wang, Edward, Small, & Sampson, 2018) are usually large in volumes. Despite the richness of data capturing people's mobility patterns within their activity spaces, most of these studies still rely on global segregation indices, summarizing population characteristics across subunits for the entire region (Blumenstock & Fratamico, 2013; Dannemann et al., 2018; Silm & Ahas, 2014.) Complex interaction patterns among population subgroups are reduced to single segregation indices. Although these indices have a long history in the literature, each measure captures only limited aspects of segregation (Yao, Fu, Liu, Hu, & Xiong, 2018). Thus, using multiple

indices has become more common to ensure more comprehensive results (e.g., Blumenstock & Fratamico, 2013; Silm & Ahas, 2014). It is important to note that these measures have been mostly employed in studying ethnic segregation in the residential space.

In the context of measuring income segregation, traditional measures have a major limitation. Among various measures of income inequality, Gini coefficient is probably the most popular (De Maio, 2007 provides a brief overview). Thus, some measures of income segregation conceptually rely on Gini coefficient, including those deviations measures (in reference to the mean or some central tendency measures) (Wheeler & Jeunesse, 2008), such as the Neighborhood Sorting Index (NSI) (Jargowsky, 1996) and its generalized form (GNSI) (Jargowsky & Kim, 2004), and the Centile Gap Index (CGI) (Watson, 2009). All these measures are conceptually evenness measures (Massey & Denton, 1988), with some of them smoothing values spatially. While using evenness (and/or clustering) measures to evaluate income distribution across neighborhood is reasonable, such approach focuses entirely on the static residential distribution of people labeled by their income levels, ignoring their potential interaction with different income groups in other socio-geographical spaces. In the context of measuring segregation that accounts for the interaction or mobility patterns of population subgroups, measures capturing the exposure dimension have been the preferred choices of measures (Kwan, 2013; Li & Wang, 2017; Wong & Shaw, 2011). Thus, evaluating income segregation based on a single index, or multiple indices of a single dimension has significant limitations if both the income differences and cross income-group interaction need to be considered.

Segregation studies often report how subgroups defined by the chosen population characteristic are separated spatially (Oka & Wong, 2019). A logical follow-up question is whether these separations of subgroups matter if subgroups face similar environments ("separated but equal"). However, many studies argue that separation leads to unequal outcomes, for instance, in education (Logan, Minca, & Adar, 2012) and health (Landrine & Corral, 2009), among others. Inequality associated with segregation is often attributable to the spatial disparities of environmental conditions and resources, which are not reflected by existing segregation measures. This is another limitation of the index-based approach to study segregation.

2.3. Limited data sources

The discontentment of relying only on indices to evaluate segregation is reflected by various analysis tools developed to investigate segregation patterns, trends, or contributing factors in recent years. Wang, Li, and Chai (2012) leveraged a kernel density estimator to examine the social segregation of residents in Beijing. Järv et al. (2015) assessed the exposure among population groups by determining the extent that different groups share the same activity locations. Wang, Edward, Small, & Sampson (2018) utilized a set of analysis tools, including clustering, descriptive statistics, and linear regression to examine the composition of activity spaces of 50 largest American cities. Olteanu, Randon-Furling, and Clark (2019) proposed to evaluate urban segregation at all scales using the Kullback-Leibler trajectories. A subsequent study adopted self-organizing maps to perform cluster-based analytics on segregation patterns (Olteanu, Hazan, Cottrell, & Randon-Furling, 2020).

Several studies reviewed above have partly demonstrated the utility of multiple sources of data that capture both the spatiotemporal dynamics of population and population's socio-demographic characteristics. Most studies as of today rely on either data about population subgroup distributions such as census data, or mobility data from surveys, cell phones or social media. Studies rarely, if any, take advantage of data from multiple sources, capturing both the spatiotemporal dynamics and the socioeconomic characteristics of the population, and the neighborhood characteristics.

2.4. The needs for new methods

We here argue that to evaluate segregation comprehensively requires integrating heterogeneous data that capture different aspects of segregation, including the spatiotemporal dynamics and sociodemographic characteristics of population, and the characteristics of neighborhoods. To accommodate the diverse types of data, a data-intensive and flexible analytical scheme is warranted. Such scheme should be able to utilize heterogeneous data collected at the neighborhood level, and these data may be used to identify local areas sharing certain segregation characteristics in multiple dimensions. While some neighborhoods sharing similar segregation characteristics may be spatially clustered, forming mesoscale subregions, other neighborhoods with similar characteristics may be scattered across the study region, creating a complex spatial pattern with heterogeneous neighborhoods. The resultant segregation structure can facilitate our understanding of how different population groups are spatially distributed and spatiotemporally interacted across the region, potentially revealing the underlying socioeconomic forces driving segregation.

To leverage heterogeneous data to evaluate segregation, methods beyond computing indices are needed. Machine learning-based data-driven methods have advantages over the traditional index-based approach because they have the capabilities to extract segregation patterns and structures from multiple dimensions (e.g., Olteanu et al., 2020). We argue that segregation of an area does not have to be indicated by one single number. We call for an expressive method to capture segregation dynamics with compact representations while at the same time preserving the original socioeconomic, neighborhood and mobility information as much as possible. Representation learning provides a powerful framework to capture discriminative and expressive information from massive data (Bengio, Courville, & Vincent, 2013) and can be employed to represent and extract spatiotemporal segregation structure. In particular, graph embedding meets this need as it strives to preserve graph structure and attribute information with compact vectors (Cai, Zheng, & Chang, 2018).

This study leverages multi-source data to examine the macro-structure of income segregation. In particular, we evaluate income segregation from the isolation-exposure dimension by integrating income and mobility data, going beyond the use of evenness-based measurement on static population distributions and accounting for potential interactions between different income groups in different socio-geographical spaces. We leverage graph embedding to develop a comprehensive representation scheme at fine-grained scales to analyze the massive smart card, amenity and housing data of Shenzhen City, China. Results deliver a holistic depiction of income segregation, including the spatiotemporal and neighborhood aspects of segregation.

3. A residential-mobility embedding approach to study segregation patterns

3.1. Graph and graph embedding

In the context of graph analysis, nodes can represent areal units such as census tracts, zip codes or grid cells, whereas the weights of links between nodes quantify the movement intensities between nodes. These nodes representing areal units can be characterized by various socio-economic and demographic attributes. Thus, the entire graph offers a holistic description of the spatial system, capturing both the characteristics of places (nodes) and the interaction of places. This graph-based representation provides a convenient tool to incorporate heterogeneous and comprehensive data relevant to segregation studies.

Machine learning methods have been employed to analyze graphs, performing a wide variety of learning tasks such as link prediction (Wang, Cui, & Zhu, 2016), node classification (Zhang, Yin, Zhu, & Zhang, 2016), community detection (Cavallari, Zheng, Cai, Chang, & Cambria, 2017), and visualization (Zuo et al., 2018). However, it is

imperative to develop automatic latent representation methods to encode complex graph structures and dynamics into compact low-dimensional vectors in the Euclidean space, which significantly facilitate the use of many downstream machine learning algorithms (Hamilton, Ying, & Leskovec, 2017). This encoding or projection scheme is termed "*graph embedding*". As the original graph data capture rich structural and attribute characteristics of nodes and links, the embedded representation should preserve the original topological structure, content, and attributes (Gao & Huang, 2018; Pan et al., 2018).

Fig. 1 illustrates embedding nodes of a graph into low-dimensional vectors, which preserve the original graph structure and attribute information. Node colors (red, yellow, and brown) in Fig. 1(a), for instance, represent different socio-economic standings of corresponding areal units, and different link widths represent the numbers of inter-nodal trips or the interaction levels between pairs of nodes. Then graph embedding constructs a *d*-dimensional vector representation for nodes in $\boldsymbol{R^d}$. Fig. 1(b) shows the output embeddings in $\boldsymbol{R^2}$ that preserve information based on graph structure and node attributes. For example, nodes 7, 9, and 10 should be embedded closely because they are strongly inter-connected with similar socio-economic profiles. Recently, the use of graph embedding in spatial networks has garnered increasing cross-disciplinary attention. Human mobility data and local attribute information are jointly embedded to discover urban structures or functions (Wang, Fu, Zhang, Li, and Lin, 2018; Yao et al., 2018).

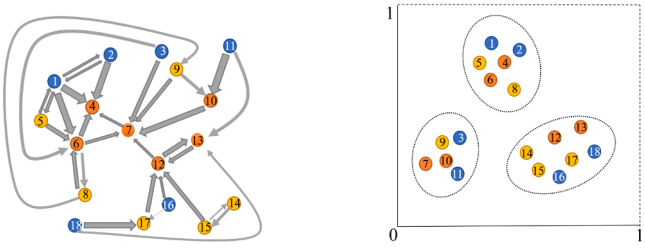
(a) An attributed graph. (b) Graph embeddings in 2D space.

3.2. Assessing income segregation using graph embedding

To study income segregation, the proposed approach considers not just the income status of the population, but also 1) the interaction among different income groups captured by mobility data from which the activity patterns of residents can be extracted, and 2) the characteristics of neighborhoods, which may include socioeconomic variables of the population. Neighborhood characteristics are regarded as attributes of nodes in the proposed framework. These data can be depicted by graphs efficiently, and graph embedding, a deep representation learning method, will be used to identify communities of different experiences to advance our understanding of income segregation patterns.

A weighted attributed graph $\mathcal{G} = \boldsymbol{(V, E, A)}$ is defined over *N* areal units of the study region. The set of graph nodes $\boldsymbol{V}$ corresponds to areal units and the set of graph edges $\boldsymbol{E}$ denotes inter-nodal interaction or connection. Each node i is associated with an f-dimensional attribute vector $\boldsymbol{H}_i \in \mathbb{R}^f$. Embedding of $\mathcal{G}$ is created by learning a low-dimensional vector representation $Y \in \mathbb{R}^d$ for each node in V where $d << f$. These nodal embeddings are then used to examine segregation.

Fig. 2 presents the proposed deep learning-based graph embedding and community detection method that unifies representation and analysis in a complete pipeline. The study region is partitioned into a regular grid and the grid is then modeled as a directed graph. Each grid cell is represented as a graph node and each link denotes people's trips from a node to another. Each node is described by multiple attributes, which may include population's socioeconomic characteristics and environmental properties. Segregation is partly determined by people's mobility patterns reflected by inter-nodal links, which encode the intensities of interaction or exposure of people between nodes. Segregation is also partly reflected by the differences in attributes between nodes. Exposures to people in other units and similarities in attributes between areal units are captured by two matrices that are constructed from the graph. Together, the two matrices capture the dynamic geographic context that residents may experience during their daily life. They form the bases of our segregation representation and analysis.



(a) An attributed graph. (b) Graph embeddings in 2D space.

Fig. 1. An illustrative example of graph embedding. Node colors (red, yellow, and brown) represent different attribute values of areal units. Widths of arrows represent the interaction intensity levels between nodes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2 depicts an attributed graph embedding scheme used to integrate the links or exposure between nodes with the node attributes based on an autoencoder model (Wang et al., 2016). After nodes are embedded, types of communities are detected by clustering the embedded nodes. A multivariate Gaussian Mixed Model (GMM, Cavallari et al., 2017) is used to enrich the representation of community types, and improve nodal embeddings and community type detection through an iterative optimization process. We improve the existing attributed graph embedding scheme by incorporating a new loss function (which will be discussed in section 3.4 below) to minimize the error between the input and output vectors in each node. We adapt the original iterative optimization method (Cavallari et al., 2017) to detect community types with similar segregation experiences. We combine the graph embedding and community detection schemes into a unified end-to-end training framework by back-propagating the loss of the integrated optimization to the autoencoder to improve the node embedding results. These steps and methods are elaborated below.

3.3. Definitions

Fig. 2 has two matrices: attribute similarity matrix and exposure matrix.

Definition 1. **Attribute similarity matrix, $\boldsymbol{W_{attr}}$,** which reflects the attribute similarity between all pairs of nodes (areal units), is defined by a Radial Basis Function:

$$w_{attr,ij} = exp\left(\frac{-\left|\left|\boldsymbol{H}_i - \boldsymbol{H}_j\right|\right|_2^2}{2\gamma^2}\right) \tag{1}$$

where $\boldsymbol{H}_i$ is a vector of attributes in the i-th area, $\|\cdot\|_2^2$ represents a similarity measured in *L2* norm, and γ is a learnable parameter.

The attributed mobility graph also captures interaction of population between areas (nodes). A mobility matrix is constructed to represent the interaction-exposure characteristic.

Definition 2. **Mobility-based exposure matrix, $\boldsymbol{W_{exp}}$,** is an asymmetric matrix that encodes the level of exposure among population groups based on mobility. In the context of income segregation, we may divide the population into, say four income groups by quartiles. Each element of the exposure matrix can be defined as:

$$w_{exp,i\to j} = I_{i\to j} + O_{i\to j} \tag{2}$$

where $I_{i\to j}$ and $O_{i\to j}$ reflect the first-order exposure and second-order exposure of the j-th area to the i-th area. The two types of exposure can be computed accordingly (Wong, 2002; Wong & Shaw, 2011).

$$I_{i\to j} = \frac{\sum_{g,g\neq k}^{m} P_{j,g} \sum_{k}^{m} Att_{i\to j} P_{i\to j,k}}{\sum_{g}^{m} P_{j,g} \sum_{k}^{m} P_{i\to j,k}} \tag{3}$$

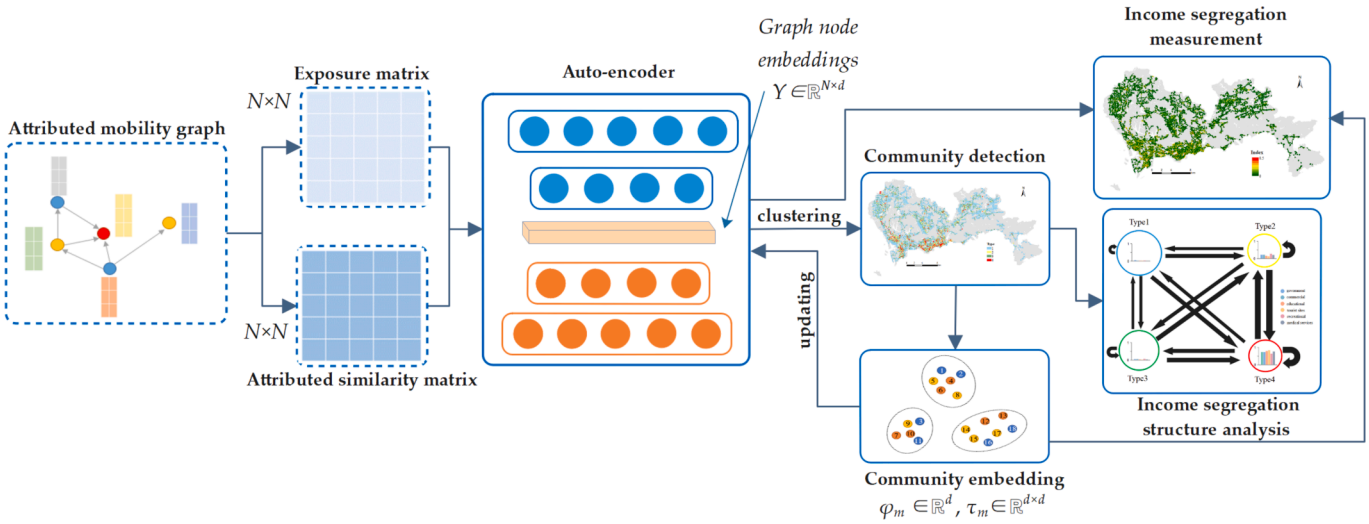


Fig. 2. Illustration of the proposed unified representation and analysis pipeline.

$$O_{i\to j}=\frac{\sum_{t}^{n}\left(\sum_{g,g\neq k}^{m}Att_{j\to t}P_{j\to t,g}\sum_{k}^{m}Att_{i\to t}P_{i\to t,k}\right)}{\sum_{t}^{n}\left(\sum_{g}^{m}P_{j\to t,g}\sum_{k}^{m}P_{i\to t,k}\right)} \tag{4}$$

The direct exposure $I_{i\to j}$ is dependent on the direct trips from i to j, while the indirect exposure $O_{i\to j}$ includes the trips from i to j but through an intermediate location t and there are n intermediate locations. Population is divided into m income groups. In the above eqs. (3–4), $P_{j,\,g}$ is the number of total travelers of income group g that arrive at area j, $P_{i\to j,\,k}$ is the number of travelers of income group k (which is different from g) departing from area i to area j ($P_{j\to t,\,g}$ and$P_{i\to t,\,k}$ are defined similarly). Because exposure is measured between two different income groups ($g \neq k$), travelers from the same income group are not counted in the numerators in eqs. (3–4). The numerators in the two equations quantify the potential interactions between different groups when they arrive at j or t. The denominators enumerate all possible interactions between different groups without the accessibility constraints, i.e., regardless of the occurrence of actual trips. $Att_{i\to j}$ is the attraction to area j from area i ($Att_{j\to t}$ and $Att_{j\to t}$are defined similarly), and it is defined as:

$$Att_{i\to j}=\frac{N_{i\to j}}{N_j}\,sigmoid\left(\frac{t_{i\to j}}{\bar{t}_j}\right) \tag{5}$$

where $N_{i\to j}$ is the number of trips from area i to j, N_j is the total number of trips that end at j, $t_{i\to j}$ represents the average trip time from i to j, $\bar{t}_j$is the average time for trips that end at j. The formulation of the attraction is based on the premise that an area is more attractive than other areas if people are willing to take longer and more trips to this area than to other areas. In other words, attraction is positively proportional to trip number and travel time.

3.4. An auto-encoder model for node embedding

We use an attributed embedding algorithm (Wang et al., 2016) to integrate the exposure dimensions in the activity space and node attribute information to evaluate segregation across areal units (graph nodes). The algorithm employs an auto-encoder to enforce information sharing between the attribute similarity matrix and the exposure matrix at the node level. The auto-encoder consists of an encoder and a decoder. The encoder produces hidden representations for input vectors in graph nodes and the decoder reconstructs data vectors from hidden representations. Assuming that there are k layers in the auto-encoder, we have:

$$\begin{aligned}\boldsymbol{Y}^1&=\sigma\left[\boldsymbol{W}_{\boldsymbol{attr}}\left(\boldsymbol{W}^1\right)^T+b^1\right]\\&\cdots\quad\cdots\\\boldsymbol{Y}^j&=\sigma\left[\boldsymbol{Y}^{j-1}\left(\boldsymbol{W}^j\right)^T+b^j\right]\\&\cdots\quad\cdots\\\boldsymbol{Y}^{k-1}&=\sigma\left[\boldsymbol{Y}^{k-2}\left(\boldsymbol{W}^{k-1}\right)^T+b^{k-1}\right]\end{aligned}$$

$$\boldsymbol{Y}^k=\sigma\left[\boldsymbol{Y}^{k-1}\left(\boldsymbol{W}^k\right)^T+b^k\right]$$

where the initial input is the attribute similarity matrix $\boldsymbol{W}_{\boldsymbol{attr}}$. $\boldsymbol{Y}^j$ denotes the representation of the joint hidden representation for the attribute similarity and exposure matrices of the j-th layer of the encoder, $\boldsymbol{W}^j$ and $\boldsymbol{b}^j$ represent, respectively, the weight and bias vectors for the j-th layer, $\boldsymbol{Y}^k$ is the final output.

The decoder aims to reconstruct the initial input based on the joint hidden representation $\boldsymbol{Y}^k$.

$$\begin{aligned}\widehat{\boldsymbol{Y}}^1&=\sigma\left[\boldsymbol{Y}^k\left(\widehat{\boldsymbol{W}}^1\right)^T+\widehat{\boldsymbol{b}}^1\right]\\\widehat{\boldsymbol{Y}}^2&=\sigma\left[\widehat{\boldsymbol{Y}}^1\left(\widehat{\boldsymbol{W}}^2\right)^T+\widehat{\boldsymbol{b}}^2\right]\\&\cdots\quad\cdots\\\widehat{\boldsymbol{Y}}^j&=\sigma\left[\widehat{\boldsymbol{Y}}^{j-1}\left(\widehat{\boldsymbol{W}}^j\right)^T+\widehat{\boldsymbol{b}}^j\right]\\&\cdots\quad\cdots\\\widehat{\boldsymbol{Y}}^k&=\sigma\left[\widehat{\boldsymbol{Y}}^{k-1}\left(\widehat{\boldsymbol{W}}^k\right)^T+\widehat{\boldsymbol{b}}^k\right]\end{aligned} \tag{7}$$

where $\widehat{\boldsymbol{W}}^j$ and $\widehat{\boldsymbol{b}}^j$ represent the learnable weight and bias vectors for the jth layer in the decoder.

A learnable weighting loss function is developed to regulate the influences of the attribute similarity matrix and the exposure matrix at the node level. The loss function is,

$$min\sum_{i,j}^{n}w_{exp,i\to j}\left\|\boldsymbol{Y}_i^{\,k}-\boldsymbol{Y}_j^{\,k}\right\|_2^2 \tag{8}$$

where $\boldsymbol{Y}_i^{\,k}$, $\boldsymbol{Y}_j^{\,k}$ are the i-th and j-th vectors after embedding. The exposure matrix, $w_{exp,\,i\to j}$, which was defined in Eq. (2), is used as a constraint in the loss function to ensure that both attribute similarity and exposure information are accounted for.

Also, we define the reconstruction loss as:

$$min\sum_{i}^{n}\left\|\boldsymbol{Y}_i^{\,1}-\widehat{\boldsymbol{Y}}_i^{\,k}\right\|_2^2 \tag{9}$$

The similarities of the two matrices between graph nodes are

preserved by minimizing reconstruction errors between input vectors and reconstructed vectors. The overall loss function can be defined as:

$$L_T = \sum_{i}^{n} \left\| \boldsymbol{Y_i}^1 - \widehat{\boldsymbol{Y}}_i^{\,k} \right\|_2^2 + \alpha \sum_{i,j}^{n} w_{exp,i\to j} \left\| \boldsymbol{Y_i}^k - \boldsymbol{Y_j}^k \right\|_2^2 \quad (10)$$

where α is a learnable parameter that regularizes the contribution of the reconstruction loss function (after some tests, we set α = 0.1 since it helps fast convergence and leads to stable community detection results).

To minimize the overall loss, we use the backward propagation method to obtain the joint embedding for each node, relying on the Adam optimizer (Kingma & Lei Ba, 2015) under the PyTorch framework. When the loss stops decreasing, the training process stops and the resultant auto-encoder can be used to produce nodal embeddings, which are normally represented as one-dimensional vectors.

3.5. Segregation-based community detection and representation

After nodes are embedded as compact representations of one-dimensional vectors based on their attribute characteristics and inter-nodal interaction, these node embeddings are clustered to derive community types. Areas sharing similar attribute and exposure profiles form a community type. Areas belong to the same community type have similar patterns of exposure to other community types. The derived community types exhibit salient segregation structure characterized by exposure between income groups and differences in community attributes, thereby facilitating segregation studies and policy making. As one-dimensional vector is used as the default form of node embedding, representing community types by one-dimensional vectors may over-simplify the complex composition of community types since a community type usually contains multiple nodes. Instead, we use a multivariate Gaussian Mixture Model (GMM) to enhance the expressiveness of community embedding (Cavallari et al., 2017). Each community type is characterized by a mean vector and a covariance matrix, which when combined, define a multivariate Gaussian distribution. The mean vector and covariance matrix together provide the overall characteristics and internal heterogeneity of a community type.

However, node-embedding and the subsequent clustering processes may not produce highly compact community types because the initial node embedding process does not sufficiently account for the cohesiveness of community types in the embedding space. The mean vector denotes the center of a community type whereas the covariance matrix encodes the compactness of its node members in related to the center. Because community type detection may be regarded as a typical unsupervised learning problem, we follow the principle of *K*-means clustering and combine node embedding, community detection, and community embedding into an integrated unsupervised optimization model that iteratively derives optimized embeddings of nodes and community types, along with community structure (Cavallari et al., 2017). Node embeddings can be improved by reducing their dissimilarities with their community-type centers, based on the assumption that nodes belonging to the same community types should be embedded closely to the community-type centers. Similar to the K-means clustering method, this process iteratively repeats until convergence. When this joint optimization is performed, loss information is back propagated to the joint embedding scheme (i.e., the auto-encoder) to derive improved node embeddings. Through updated node embeddings, nodes that are expected to belong to the same community types will have more similar embeddings. Then a more consistent community-type structure can be discovered after each iteration.

Assuming that we have M types of communities, each community follows a multivariate Gaussian distribution(φ_u, τ_u), where $\varphi_u \in R^d$ and $\tau_u \in R^{d\times d}$ are the mean and covariance of node vectors in the *u-th* type. Then the objective is to extract community types possessing certain segregation characteristics. Each type consists of multiple areas and areas with the same Gaussian distribution belong to the same type of community. The most suitable number of community types is identified by the gap statistic model (Tibshirani et al. 2001). To unify the community-type detection and embedding into an integrated optimization framework, we need to optimize the following likelihood function (Cavallari et al., 2017):

$$\prod_{i=1}^{N} \sum\nolimits_{u=1}^{M} pr(v_i \in C_u) pr(v_i | v_i \in C_u; Y_i, \varphi_u, \tau_u) \quad (11)$$

where $pr\ (v_i \in C_u)$ denotes the probability of node v_i being classified as the u-th type. Y_i is the node embedding for v_i.

The loss function for the integrated optimization can be defined as:

$$l_G = -\frac{\beta}{M} \sum_{i=1}^{N} log \sum_{u=1}^{M} pr(v_i \in C_u) pr(v_i | v_i \in C_u; Y_i, \varphi_u, \tau_u) \quad (12)$$

where β is a trade-off parameter.

The optimal node and community-type embeddings can be derived by minimizing the loss function in Eq. (12). Meanwhile we can also obtain the optimum $pr(v_i \in C_u)$, with which node embedding can be updated based on Eq. (11) by assuming Y_i is unknown. This iterative node embedding process has the benefit of directing nodes of the same community type to have similar embeddings.

Fig. 3 illustrates the iterative learning for community-type detection and the update of node embedding and the final community embedding. Initially, three community types are detected based on node embeddings (Fig. 3a). Each detected community type can be described by a GMM-based embedding (φ_u, τ_u). In each iteration, node embedding is updated by fixing community-type embeddings and community-type membership probabilities and optimizing Eq. (12). After obtaining improved node embeddings, new community-type structure and updated community-type embeddings are derived (Fig. 3b). Finally, this process converges and community-type structure should be more compact than previous iterations in the embedding space as nodes are embedded closely to the community-type centers (Fig. 3c).

4. Data and results

The methodology described above was applied to detect community types with different characteristics of income segregation in the city of Shenzhen. In the current study, income segregation is characterized by house value differentials, exposure to residents of different community types, and community amenity levels.

4.1. Study region and data

Shenzhen is the largest city in southern Guangdong province in China bordering Hong Kong to the south. While it is one of the largest and fastest growing Chinese cities, it is a relatively new city with highly uneven levels of development. The city is divided into ten administrative districts with the more developed one concentrated in the west and southwest. Data used in the study are described in Table 1.

As income data are in general not available in China, housing price is treated as a proxy variable of income of a household as house value should be a reasonable indicator of a resident's income level. Housing price data were obtained from the website of Lianjia (https://wf.lianjia.com), the largest Chinese real-estate brokerage company (Table 1). As the website indicated house values by point locations, spatial interpolation was used to generate a housing price surface for the entire city. The entire city is divided into 100-m by 100-m grid cells. Using the inverse-distance weighting (IDW) interpolation method, a house value was estimated for each grid cell. After unpopulated cells were removed, cells were further consolidated into grid groups or communities. In operation, transit networks (buses and subways) with stops were overlaid onto the grid. Cells were assigned to the nearest stops and cells sharing the same stops are grouped together (grid groups) to form

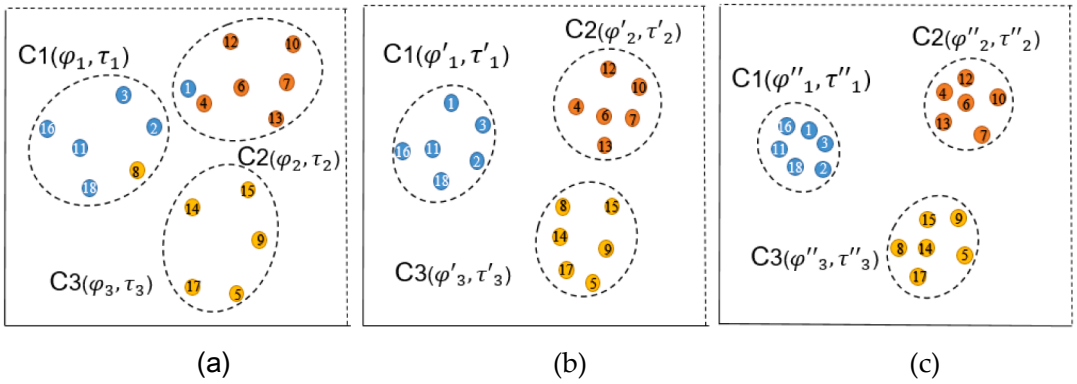


Fig. 3. Iterative community detection and embedding updating: (a) initial results, (b) intermediate results with improved likelihood, (c) final results with the most compact embeddings.

Table 1
Data description.

Data	Description
Smart Card Data (SCD)	Complete boarding-alighting records for subway transactions but only boarding time for bus-based trips (April 3–9, 2017)
Communities	Dividing Shenzhen into 100 m × 100 m grid cells (~ 181,000 cells), and merging grid cells into 18,108 communities (grid groups) with similar transit boarding patterns
Community attributes	Point-of-interest (POI) positions and types (including 5394 government agencies, 54,897 commercial, 3540 educational, 194 recreational, 7520 medical services and 186 tourist sites)
housing price data	Crawling housing price from the *Lianjia* website (https://wf.lianjia.com/, the largest Chinese real-estate brokerage company, similar to Zillow). (data accessed in May 2020)

communities. Thus, these communities are formed by grouping grid cells with similar transit boarding profiles, i.e., the same nearest stop or a list of similar boarding stops (if two cells have no exact nearest stop). For each grid group (i.e., community), its housing price is calculated by averaging over its constituent grid cells. Interpolated housing prices for communities in four categories (low, medium-low, medium-high, and high) are shown in Fig. 4.

Fig. 4 shows that residents with higher-income levels tend to live in the southwestern part of the city (i.e., downtown), which hosts the major employment centers and high-paid jobs. Western part of the city also has much higher average housing prices than the eastern part. At the regional-district level, income segregation based on housing price seems high using the evenness dimension. However, house values vary significantly at the local scales in the downtown area, revealing the highly uneven spatial distribution of wealth.

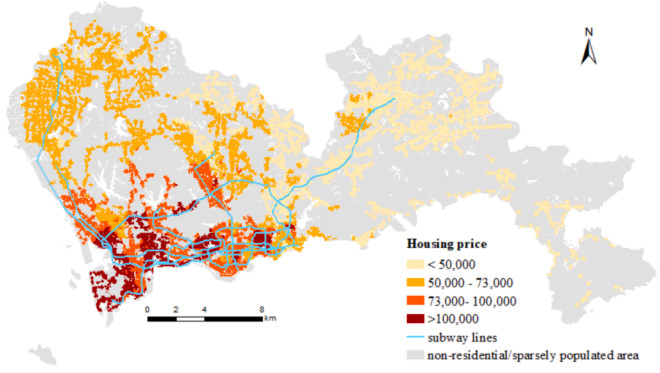


Fig. 4. Interpolated housing price (in Chinese Yuan) for the 100-m by 100-m grid, Shenzhen, China. Housing values are categorized into 4 classes based on natural breaks.

In our analysis framework, housing values and densities of different Point-of-interest (POI) types are regarded as attributes of nodes from which the attribute matrix is derived (eq. 1). Six types of POI are used in the study: government agencies, commercial, educational, recreational, medical services, and tourist sites (Table 1). Although other types of POIs can be used, these six types are chosen mainly to capture the amenity richness of each community and to a certain extent, its attractiveness to visitors. The downtown area includes large numbers of POIs. However, high income is not strongly associated with the numbers of POIs. The numbers of government, commercial, and educational POIs have weak positive correlations with housing values (Pearson correlations of 0.095, 0.189, and 0.107, respectively). Density of a POI type is computed by dividing the number of a POI type by the area of the community. For each community (grid group), a vector is constructed to include housing values and densities of different POI types as attributes.

As in most large and medium cities in China, Shenzhen has a very well-developed transit system served by buses and subways. Transit fares are paid by smart cards. In this study, we used smart card data (SCD) gathered by the transportation authorities of the city during the week of April 3–9 Cavallari et al., 2017. This massive data set contains complete boarding and alighting records of individuals for subway transactions but only the boarding time for bus-based trips. To reconstruct the complete trip chains for both subway and bus trips (including location and time), we reconstructed bus-based trip legs by estimating boarding stop, alighting time, and alighting stop of each trip leg (Zhang et al., 2020). To estimate these leg components, bus trajectory dataset is integrated with the SCD to identify the most probable boarding stop for each leg. Then the alighting stop and time for the leg are inferred based on the common assumption that the most likely alighting stop is closest to the next boarding stop. Finally, separate trip legs are linked together to form complete trip chains if consecutive legs are within 30 min (i.e., we assume the trip made multiple stops). Based on this derived data set, which includes commuting and non-commuting trips, the exposure matrix capturing people's mobility patterns is constructed for the community (grid groups) structure (Fig. 2).

4.2. Segregation analysis

4.2.1. Detection and analysis of segregation structure

The embedding process considering income characteristics, exposure of income groups across communities and neighborhood characteristics produces four types of communities, and their income-group compositions are shown in Fig. 5. The patterns between weekdays and weekends are highly similar. Type 1 communities are dominated by the lower-income groups, while the majority of Type 4 communities have the higher-income groups. Between the two are Type 2 and 3 with Type 3 very similar to Type 1. Type 2 income composition is not very similar to Type 4, but it has the highest proportions of higher-income groups after

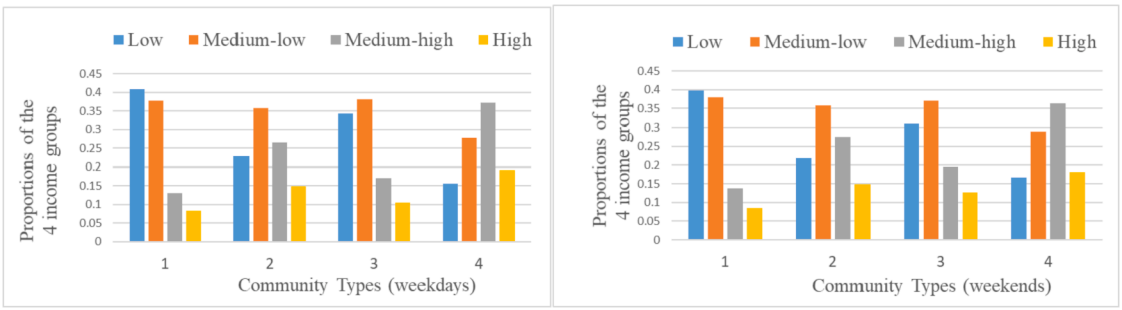
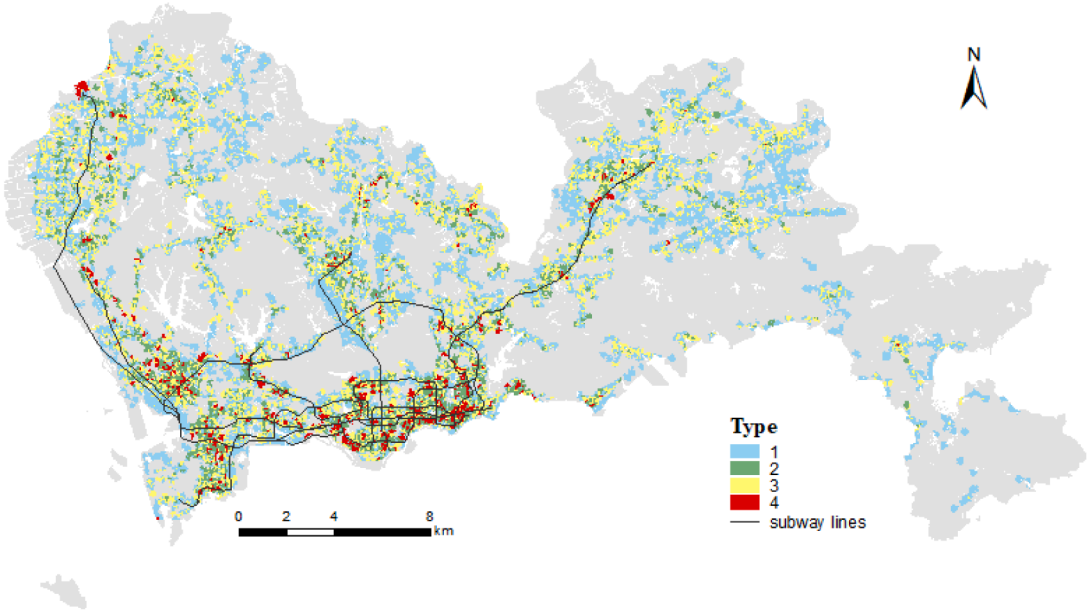


Fig. 5. Proportions of income groups in each of the four types of community types (left: weekdays; right: weekends).



(a) Weekdays

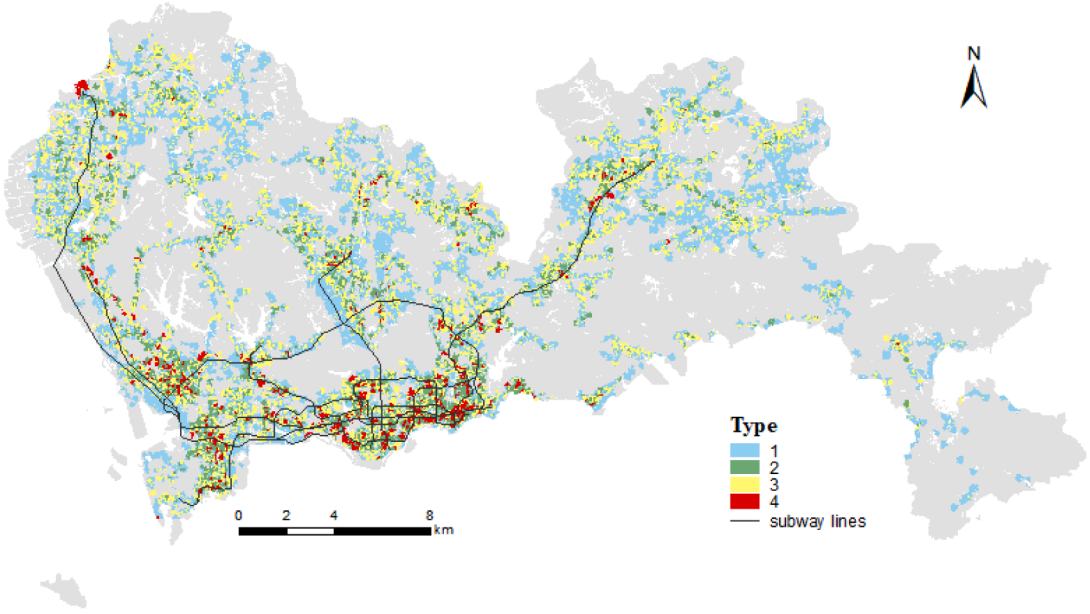


Fig. 6. Detected segregation community type structure.

Type 4.

Spatial distributions of the four types of community are shown in Fig. 6. Type 1 communities are mostly found in the outskirts of the city where subways are poorly served. On the contrary, Type 4 communities are concentrated in and around the downtown region in the south-southwest part of the city. However, it is important to note that city center is not exclusively occupied by Type 4 communities. Many of these communities are also found along subway lines outside the city center and even in terminus areas of the subway lines far away from the city center.

Since the averages of distances to stations and trip lengths (Fig. 7a and b) between weekdays and weekends for each of the four community types are quite similar, therefore, weekdays and weekends are not discussed separately here. Type 1 communities are characterized by very low rates of exposure to the rest of the population regardless of income level as shown by the low ratios of inflow trips to the local populations in Fig. 7 (c and d). The low ratios of inflow trips reflect that the numbers of people in all four income groups from other types of communities are small as compared to the residential population in Type 1 communities. In other words, adopting the exposure dimension of segregation, residents in Type 1 communities are highly segregated from residents in other communities, due to their low exposure levels. They are relatively isolated partly due to the geography that they are mostly located in areas with poor public transit access (Fig. 6a and b). The low transit accessibility levels in these communities are reflected by the longest average distances to the nearest subway stations (Fig. 7a and b). The costs of residents in these areas to access transit services are relatively high. Also, trips originated from these communities are the longest on average, though not by a large margin, further exemplifying the inconvenience of using transit for Type 1 communities. Notably, Type 1 communities cover areas of two income extremes (i.e., very high- and very low-income) when these communities are compared with the housing map in Fig. 4 and the income composition in Fig. 5. Due to the lowest exposure levels to residents in other communities and the highest costs (in terms of distances to access transit) to reach other communities, Type 1 communities maybe regarded as the most segregated community type based on the high costs of access to transit and low inflow ratios. We note that both Type 1 and Type 3 communities are distributed in remote suburb areas with relatively low housing prices. The main difference between the two types is the access to the subway system: Type3 communities generally are close to subway stations, making them easier to travel by transit whereas Type 1 communities have longer walking distances to nearest subway stations. This difference leads to lower exposure levels of Type 1 communities than Type 3 communities since people from other areas are difficult to access Type 1 communities via subways.

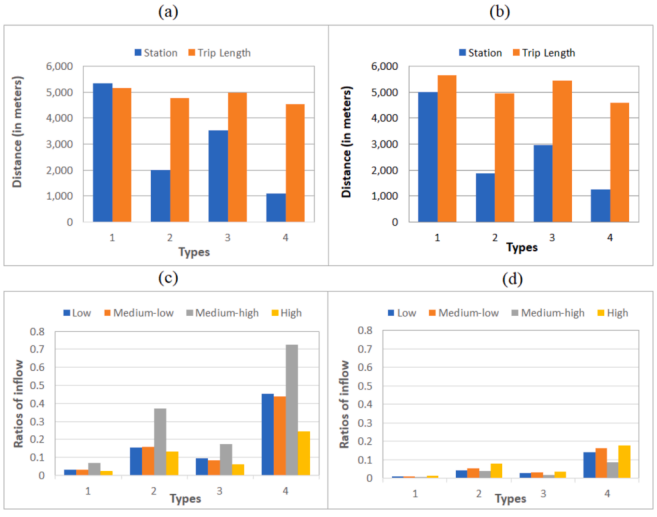


Fig. 7. Four types of community (1, 2, 3, and 4) defined by the average distance to the nearest subway stations and average trip length (a: weekdays; b: weekends), and the ratios of inflow trips from the four income groups (low, medium-low, medium-high, and high) in other communities to the local populations (c: weekdays; d: weekends).

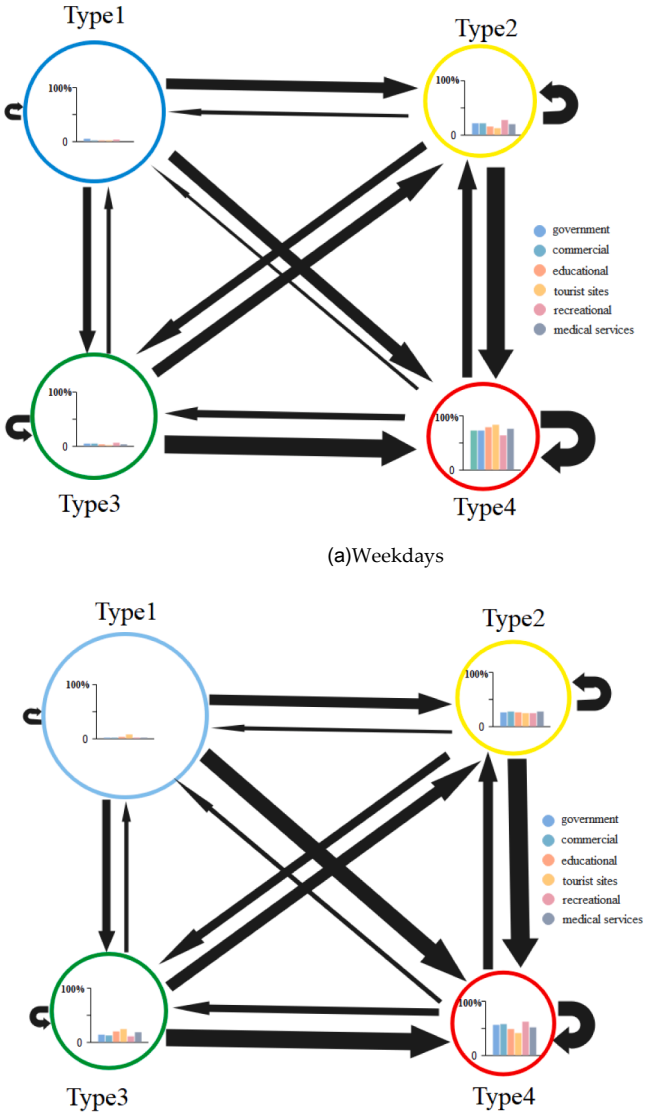
On the contrary, Type 4 communities are close to major subway stations (about 1 km on average, Fig. 7a and b)), enjoying high levels of transit accessibility and interaction with people from different income groups as reflected by the highest ratios of inflow trips to the local population (Fig. 7c and d). Also, Type 4 communities have relatively large numbers of POIs and job opportunities (this will be illustrated later),[1] further attracting people in other communities to visit and thus elevating the exposure levels of local residents. The situations of Types 2 and 3 communities fall between the two extremes of Types 1 and 4. They have moderate levels of exposure to residents across income groups (Fig. 7c and d). A difference between Types 2 and 3 communities is that Type 3 communities are slightly less accessible than Type 2 communities since most Type 3 communities are located in the suburbs connected to one single subway line (Fig. 6) while many Type 2 communities can be accessed conveniently by transit services. Compared to Type 3, residents of Type 2 communities experienced higher exposure levels to residents in other types of communities as indicated by their higher ratios of inflows to local population over those ratios for Type 3 communities (Fig. 7c and d). Fig. 6 also shows that communities in the downtown area are more spatially fragmented than in other areas, indicating that community differentiation is highly localized in the downtown area.

Although accesses to subway stations in terms of distances are clearly different across the four community types, average trip lengths do not vary tremendously across community types. This phenomenon has at least two implications. People belonging to different community types made transit trips of similar distances. Note that community types are not straightly defined by their city center-peripheral settings as trip length is not an attribute in formulating community types. Nevertheless, the rank correlation between the access to station and the average trip length across community types is perfect (cannot evaluate its statistical significance due to a small n of 4). Taking both access to stations and average trip length together as time cost, Type 1 communities pay the highest cost in order to interact with others and Type 4 communities pay the least.

Although, transit access and trip length patterns between weekdays and weekends are similar (Fig. 7a and 7b), the community exposure patterns to different income groups, which are shown by the ratios of inflow trips from the four income groups in other communities to the local population, have several noticeable differences between weekdays and weekends (Fig. 7c and d). Clearly, residents are more reluctant to travel by transit on weekends than on weekdays. Therefore, exposures to residents from other communities on weekends are lower than the exposures during weekdays. In other words, people across all community types are more segregated on weekends than on weekdays. Assuming that trips made on the weekdays are job-related, then jobs help lower segregation across communities. This claim is further substantiated by the fact that Type 4 communities are densely populated by multiple categories of POIs (Fig. 8), attracting travelers across the city.

Across all four types of communities, the medium-high income group is the largest group of visitors on the weekdays but is the smallest group of visitors on weekends, based on the ratios of inflow trips to local population (Fig. 7c and d). Thus, people in the medium-high income group travelled the most among the four income groups on weekdays, possibly related to their jobs, but on weekends, they travelled the least. On the contrary, people in the high-income group travelled the least on weekdays, but the most on weekends, likely for leisure. From the

[1] According to the company location data from Baidu map, the density of companies for the four types of communities on weekdays are 101.8, 285.6, 201.8, and 360.0 per km^2, respectively.



(a)Weekdays

(b)Weekends

Fig. 8. Mobility networks (a: weekdays; b: weekends) of detected community types. Width of an arrow is proportional to the size of normalized flow. Size of a circle is proportional to the area of a different community type. Colors of circles correspond to the colors of community types in the legend of Fig. 6. Bar chart inside each circle indicates the ratios of POIs over the total POI numbers for the six POI categories. Colors of histograms are annotated in the legend.

segregation perspective, medium-high income population has the highest level of exposure to various communities on weekdays, but the lowest exposure on weekends. High-income population is the most isolated on weekdays, but exposed to other groups the most on weekends.

Among the four types of communities, Type 1 has the lowest exposure to visitors while Type 4 has the highest exposure level. Types 2 and 3 have intermediate exposure levels with Type 2 having slightly high levels. Thus, based on the exposure levels to visitors, Type 4 communities may be regardless as the least segregated or isolated, while Type 1 communities may be the most isolated. However, Fig. 7 fails to describe how populations in the four types of communities interact. Fig. 8 provides some insights about their interactions.

In Fig. 8, width of an arrow is proportional to the size of normalized flow, including the within-type trips. The normalized flows between Type i and Type j communities can be calculated by multiplying the origin frequencies with destination frequencies.

$$NF_{i\to j} = \frac{N_{i\to j}}{N_{i,out}} \times \frac{N_{i\to j}}{N_{j,in}} \tag{13}$$

where $N_{i\to j}$ is the number of trips from community i to j, $N_{i,\,out}$ is the total number of trips leaving i, $N_{j,\,in}$ is the total number of trips arriving at j.

Size of a circle in Fig. 8 is proportional to the total area of the corresponding community type. Bar chart inside each circle displays the ratios of POIs over the total POI numbers for the six POI categories. Arrows in Fig. 8 indicate that Type 4 communities have the highest levels of interaction with other types of communities. In particular, they attract the largest volumes of flows from Types 2 and 3 communities on both weekdays and weekends. Therefore, Type 4 communities have the highest exposure levels as they function as activity hubs for residents from most types of communities across the city. On weekdays, Type 4 communities attract substantial trips from areas with medium-low and low-income groups (Fig. 7c), although the medium-high group constituted the largest inflow stream. These patterns should not be surprising as Type 4 communities have disproportionally large number of POIs (followed by Type 2 communities). Type 4 communities also have the largest normalized within-community flows on both weekends and weekdays.

Due to poor transit accessibility and the small numbers of POIs, trips end within Type 1 communities are comparably less than other community types, thereby validating that they are probably the most segregated areas in the city on both weekdays and weekends. Type 1 communities also occupy the most area on both weekdays and weekends, while Type 4 is the smallest by area. Interaction between Type 2 and 3 communities is substantial but interaction between Type 1 and Type 3 is relatively minor on both weekends and weekdays, suggesting that Type 3 may be less exposed to other types of communities than those of Type 2. On weekends, lowered interaction levels across community types raise segregation overall (Table 2). Such lowering of inter-community interaction is accompanied by the slight increases of intra-community flows for Type 1 and Type 3 communities, reinforcing their isolation from residents in other types of community on weekends.

4.2.2. *Measuring local exposure*

Based on node embeddings, we developed a local exposure measure that quantifies the exposure level of each community or grid group. The idea is to measure the similarity between a learned real-world node embedding and a local mobility-based theoretical embedding of a node. This local theoretical embedding is computed based on a non-parametric radiation model that improved the well-known Newtonian-based gravity model to model spatial interaction (Simini, González, Maritan, & Barabási, 2012).

In this radiation model, the volume of inflow trips of a node is dependent on the populations of the origin and the destination nodes, as well as the population in the vicinity of the origin location (excluding the populations of the origin and the destination locations). Formally, the average inflow trip from i to j can be calculated by,

$$T_{i\to j} = T_i \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})} \tag{14}$$

where m_i *and* m_j represent the populations of i and j, respectively, and s_{ij} is the population within the circular area surrounding i with a radius of

Table 2
Average local exposure measurements over community types. The numbers following ± are standard deviations.

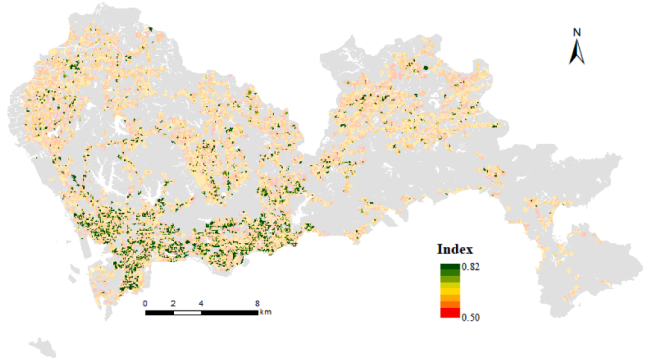
Type	Weekdays	Weekends
1	0.5872 ± 0.0011	0.5509 ± 0.0008
2	0.6360 ± 0.0048	0.6032 ± 0.0009
3	0.5922 ± 0.0015	0.5520 ± 0.0006
4	0.7763 ± 0.0113	0.7271 ± 0.0069

r_{ij} (i.e., the distance between i and j), but excluding the populations of i and j. T_i is the sum of trips originated from i. This radiation form of the mobility model assumes that people's movement is not affected by the characteristics of individuals such as income, or the neighborhood characteristics such as amenities (in the current study, POIs), and the transportation networks. Thus, trip volume estimates provided by Eq. (14) can be regarded as people's exposure purely determined by the geographical distribution of population in the region. Interaction of people is not restricted by landscape factors or negatively affected by increasing distance (the distance-decay effect) between locations. Then, the local exposure measure of a node i can be formulated as the difference in the exposure levels to other income groups based on the level estimated by the theoretical or ideal trip volume using the radiation-based mobility model and the level based on the learned embedding of the node. The proposed local exposure measurement index based on the notion of cosine similarity can be defined as,

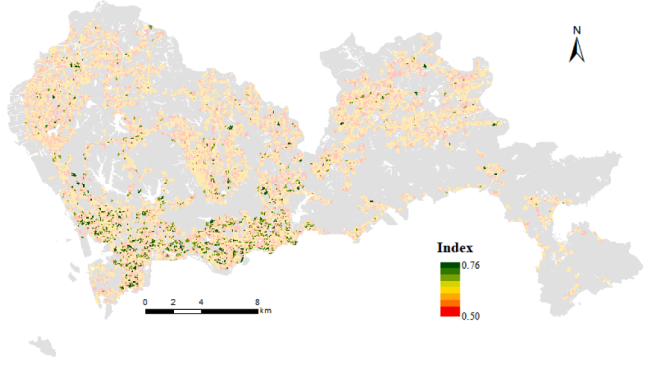
$$e_i = \text{cosine}\left(\boldsymbol{Y}_i \widetilde{\boldsymbol{Y}}_i\right) = \frac{\boldsymbol{Y}_i \bullet \widetilde{\boldsymbol{Y}}_i}{\left\|\boldsymbol{Y}_i\right\| \left\|\widetilde{\boldsymbol{Y}}_i\right\|} \in \left[-1, 1\right] \tag{15}$$

where Y_i denotes the learned real-world embedding vector for the i-th node. $\widetilde{\boldsymbol{Y}}_i$represents the local theoretical mobility embedding vector of node i, which is computed based on the theoretical inflows ($T_{i\to j}$). The local measure has a range of [−1,1], with −1 indicating opposite vector directions while 1 reflects perfect resemblance in direction of the two vectors. Thus, larger e_i reflects closer to the ideal or higher exposure, and vice versa.

According to Fig. 9, e_i ranges from 0.5 to 0.82, showing the real-world embedding vectors resemble the ideal ones at moderate to high levels. The least segregated communities (with the highest exposure levels) are mostly concentrated in the south and southwest parts of the city. Although some of the least segregated communities are found in the suburb and periphery regions, they are highly localized in selected locations. Areas characterized as the most segregated (the lowest exposure levels) are mostly distributed widely in the suburban and peripheral areas. Weekends have lower overall exposure levels than weekdays, as indicated by lower exposure measurements across the study region. On weekdays, high-exposure areas are usually characterized by their high transit accessibility and attractiveness of concentrated job opportunities. On weekends, high exposure levels are evident in commercial centers in downtown areas that are easily accessible by subways.



(a)Weekdays



(b) Weekends

Fig. 9. Embedding-based local exposure measurement. High exposure index values indicate low segregation levels.

For many suburb areas, residents need to take much longer trips than people living in downtown areas in order to access various POIs and opportunities. Even they manage to interact with different population groups, extra transportation costs are significant. When trips are not required (job-related) on weekends, suburb residents tend to make less trips, resulting in much lower exposure levels.

We computed the aggregated exposure measurements for the detected community types on both weekdays and weekends (Table 2) by averaging local exposure measurements over all grid groups with the same community type. Results show that: (1) exposure levels on weekends are slightly lower than those on weekdays; (2) the order of exposure levels is Type 1 < Type 3 < Type 2 < Type 4; (3) weekend exposure levels of different community types are less heterogeneous than those on weekdays, according to the standard deviations of local exposure in Table 2. The downsizing of trips leads to dramatic decreases of exposure for Types 1 and 3.

4.2.3. Comparison to traditional income segregation measures

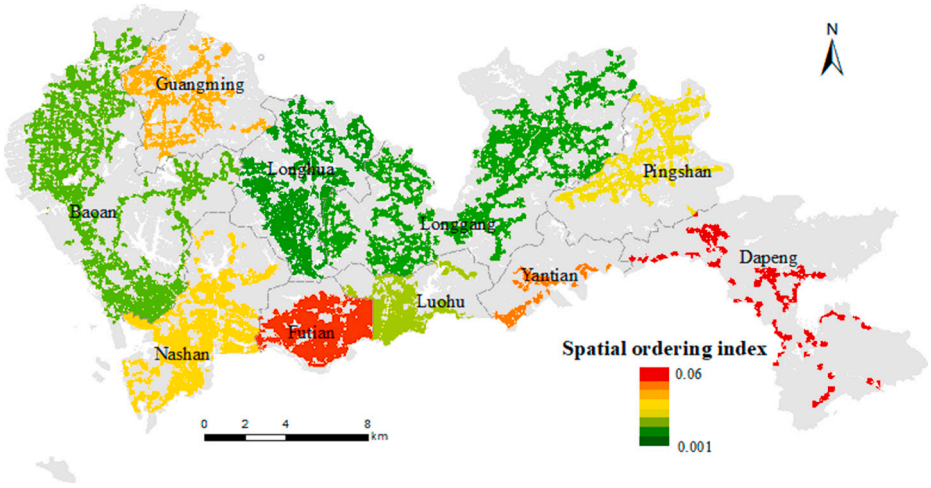
We further compared the proposed exposure-based measures with the Gini-based spatial ordering index (Dawkins, 2007) at the administrative district level. The spatial ordering index is "a comparison of the original neighborhood income parade with the spatially ordered income parade." (Dawkins, 2007, 260) It is computed from the ratio between the spatial Gini index calculated from a nearest neighbor and a Gini index of between-neighborhood income segregation. It is inherently a spatial evenness measure based on Gini or rankings. Fig. 10(a) shows the spatial ordering index for the ten districts in Shenzhen. Based on the proposed local exposure measure, we computed the district-level segregation index as follows,

$$s_k = 1 - \sum_{i=1}^{I} e_i \Big/ I_k \tag{16}$$

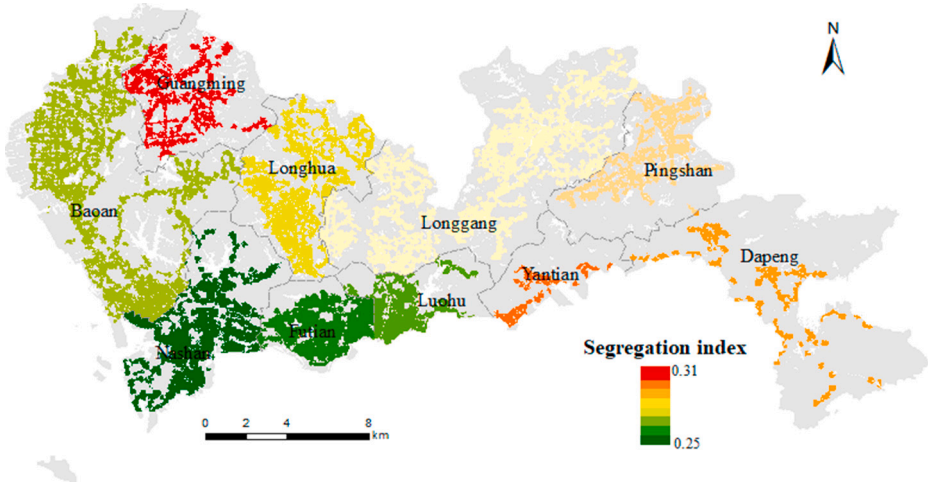
where e_i denotes the local exposure of node i. $I_k$is the number of nodes that belong to the kth district.

Fig. 10 shows that results between the spatial ordering index and the exposure-based index are dramatically different. Just to take two extreme cases to illustrate the differences: *Futian* district, part of the downtown area, has the highest segregation level according to the spatial ordering index, but has one of the lowest segregation levels according to exposure. On the contrary, *Guangming* district has moderately high segregation levels according to the spatial ordering index, but it has one of the highest segregation levels according to exposure on both weekdays and weekends. Although the two measures are not exactly opposite in their results, they are in large disagreement in general, except that *Dapeng* district is ranked highly for both indices. However, these inconsistent results should not be surprising. While the spatial ordering index considers only static house values within each district to capture local variability or evenness, which can easily be derived visually by examining the housing price map in Fig. 4, the proposed measure accounts for population dynamics and thus exposure patterns, and identifies variability in segregation levels in different times (i.e., between weekdays and weekends).

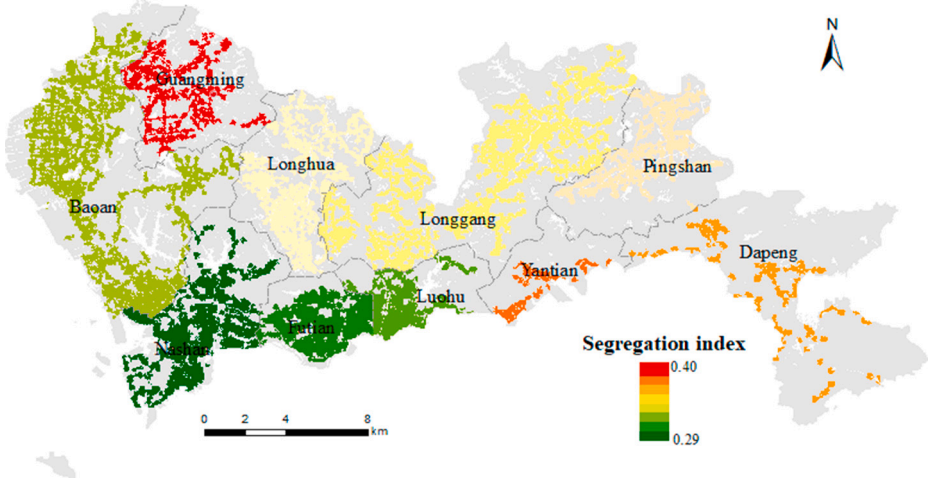
As shown in Fig. 11, downtown districts in the south (*Nanshan*, *Futian*, and *Luohu*) generally have low degrees of segregation while less



(a) Dawkins' spatial ordering index



(b) Proposed exposure-based segregation measurement on weekdays



(c) Proposed exposure-based segregation measurement on weekends

Fig. 10. Comparison of income segregation measurements at district level.

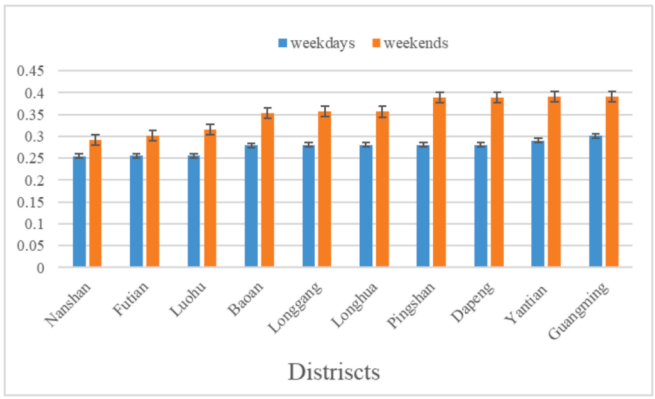


Fig. 11. Segregation levels of the ten districts in Shenzhen, according to the exposure measure for weekdays and weekends. Error bar at the tip of each bar is the standard deviation of the segregation levels within the district. Districts are sorted according to the segregation levels on weekdays.

affluent suburb and exurb areas are more segregated. Public transit accessibility is a key factor: districts with high accessibility are usually associated with low segregation levels. The rank orders of the districts in segregation levels between weekdays and weekends are slightly different, indicating that the proposed exposure measurement approach is capable of capturing the temporal dynamics of segregation.

5. Conclusion and discussion

This study proposes using the graph embedding as a representation and analysis framework to identify community types with similar segregation experiences. These experiences are characterized by the level of amenities in the residents' communities, residents' exposure to other groups defined by the concerned characteristic (in this study, housing value as a proxy to income). The proposed framework considers not just population and community characteristics, but also population dynamics which affect segregation level over time. Contrary to the traditional index-based approach which mostly considers one population characteristic (may that be race-ethnicity, income, or occupation) and ignores the community characteristics of the population, results of the proposed framework provide rich descriptions of segregation experiences of residents and associated environments.

The main contribution of the current study is on the methodological realm, while the understanding of income segregation of Shenzhen is preliminary and exploratory in nature. Shenzhen is a unique Chinese city as it was one of the four earliest designated special economic zones (SEZs) to jump start China's open-door economic policy in the 1980s. The city literally had no history before 1980. Understanding the development of Chinese cities in general and income segregation in specific need to recognize the larger contexts of China's experience in globalization and the burgeoning post-socialist economy (Ma & Wu, 2005). Some of the China's unique urban development phenomena need to be considered include the influx of floating population (migrant workers) and the emergence and removal of urban villages (*chengzhongcun*) (Liao & Wong, 2015). Therefore, comparing the income segregation geography of Shenzhen with that of another city in the Global North and South should account for different urban development processes and histories (van Ham et al., 2016). We are aware of a major limitation of our empirical results. SCD were used to extract transit-based trip patterns. Because wealthier residents are unlikely frequent patrons of transit services, mobility patterns extracted might not represent the situation of this subgroup sufficiently. This is a limitation of the data, not the framework.

Nevertheless, the current study of Shenzhen adds to our knowledge about the socioeconomic landscape of some cities in general. Ignoring the income variability at the highly local scale (which is an intriguing phenomenon in some Chinese cities), in general, Shenzhen's residents with higher income are mostly found in the more developed and highly accessible districts in the southwest and south, while those resided in the less developed east side of the city are mostly lower-income people (Fig. 4). This simplified center-rich periphery-poor urban structure is contrary to the structures of many North American cities (Glaeser, Kahn, & Rappaport, 2008), but to a certain degree, are similar to the structure of some cities elsewhere, including some European cities (Brueckner, Thisse, & Zenou, 1999; Hochstenbach & Musterd, 2018), and some Latin American cities (Griffin & Ford, 1980). Higher-income groups reside closer to the city centers while poorer groups are concentrated in the less accessible peripheral areas.

Applying the principles to assess income segregation developed for Shenzhen in this study to other cities, both the spatial distributions of different income groups and their neighborhood accessibility should be taken into account. Many cities in the developing world are flooded with informal settlements, most of which are occupied by the poorest population in the least accessible areas (e.g., Huchzermeyer, 2002), a double whammy of poverty and isolation. If the poorest sections of the city have access to efficient transportation, such as those selected locations beyond the center districts of Shenzhen but are near the transit stations, their segregation levels may be lowered (Fig. 9). Thus, transportation infrastructure can be a means to reduce segregation in general.

However, the most segregated communities (Type 1) in Shenzhen are not exclusively constituted by the lower-income residents. They also housed some of the wealthy people, possibly the super-rich. The rich, who likely have personal transportation and therefore their mobility data are not included in the current study, may choose to be self-segregated, residing in enclaves such as gated communities. These residential locations may be remote and not easily accessible by public transportation so that they can set certain physical distances from the rest of the society. Such phenomenon exists in Chinese cities including Shenzhen and also elsewhere around the world (Atkinson & Ho, 2020). In this case, segregation or isolation is intended.

The employed graph embedding representation is a highly flexible data-driven machine learning method that encapsulates characteristics of nodes and their interactions or linkages into an integrated representation and analysis framework. In the context of segregation study, a desirable feature of this framework is to capture the movement of people, an important aspect of exposure in evaluating segregation. The current study employs transit trip information extracted from SCD. Other data sources can be used. Similarly, the current study uses house values as proxies of income levels. The proposed framework can accommodate other attributes to classify population into subgroups. Neighborhood characteristics are defined by six types of POIs. Apparently, other neighborhood characteristics can be used. In other words, the proposed framework is highly flexible and adaptable and will open new avenues for segregation studies.

Declarations of Competing Interest

None.

Author statement

Tong Zhang: Conceptualization, Methodology, Writing - Original draft, Formal analysis, Supervision, Project administration, Funding acquisition.

Xiaoqi Duan: Software, Validation, Investigation, Data curation, Visualization.

David W.S. Wong: Writing - Original draft, Writing- Reviewing and Editing, Formal analysis, Validation.

Yashan Lu: Software, Validation, Data curation.

Acknowledgements

This work was jointly supported by the National Key R & D Program of China (International Scientific & Technological Cooperation Program) under Grant 2019YFE0106500, the National Natural Science Foundation of China under Grant 41871308, and the Fundamental Research Funds for the Central Universities.

References

Atkinson, R., & Flint, J. (2004). Fortress UK? Gated communities, the spatial revolt of the elites and time-space trajectories of segregation. *Housing Studies, 19*(6), 875–892.

Atkinson, R., & Ho, H. K. (2020). Segregation and the urban rich: Enclaves, networks and mobilities. In S. Musterd (Ed.), *Handbook of Urban Segregation* (pp. 289–305). Edward Elgar Publishing.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representative learning: A review and new perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Blumenstock, J., & Fratamico, L. (2013). Social and spatial ethnic segregation: A framework for analyzing segregation with large-scale spatial network data. In *Proceedings of the 4th annual symposium on computing for development (ACM DEV-4 '13. ACM), New York, America, article no 11*.

Boterman, W., Musterd, S., Pacchi, C., & Ranci, C. (2019). School segregation in contemporary cities: Socio-spatial dynamics, institutional context and urban outcomes. *Urban Studies, 56*(15), 3055–3073.

Brueckner, J. K., Thisse, J.-F., & Zenou, Y. (1999). Why is Central Paris rich and downtown Detroit poor?: An amenity-based theory. *European Economic Review, 43*(1), 91–107.

Buliung, R., & Kanaroglou, P. (2006). Urban form and household activity-travel behavior. *Growth and Change, 37*(2), 172–199.

Cai, H., Zheng, V., & Chang, K. (2018). A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering, 30*(9), 1616–1637.

Cavallari, S., Zheng, V., Cai, H., Chang, K., & Cambria, E. (2017). From node embedding to community embedding. In *, 6–10. Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM), Singapore, November* (p. 2017).

Dannemann, T., Sotomayor-Gómez, B., & Samaniego, H. (2018). The time geography of segregation during working hours. *Royal Society Open Science, 5180749*.

Dawkins, C. (2007). Space and the measurement of income segregation. *Journal of Regional Science, 47*(2), 255–272.

De Maio, F. (2007). Income inequality measures. *Journal of Epidemiology and Community Health, 61*, 849–852.

Farber, S., O'Kelly, M., Miller, H., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography, 49*, 26–38.

Farber, S., Páez, A., & Morency, C. (2012). Activity spaces and the measurement of clustering and exposure: A case study of linguistic groups in Montreal. *Environment and Planning A: Economy and Space, 44*(2), 315–332.

Gao, H., & Huang, H. (2018). Deep attributed network embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden.* July 13–19 2018.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal, 31*(121), 124–126.

Glaeser, E. L., Kahn, M. E., & Rappaport, J. (2008). Why do the poor live in cities? The role of public transportation. *Journal of Urban Economics, 63*(1), 1–24.

Griffin, E., & Ford, L. (1980). A model of Latin American City structure. *Geographical Review, 70*(4), 397–422.

van Ham, M., & Tammaru, T. (2016). New perspectives on ethnic segregation over time and space. A domains approach. *Urban Geography, 37*(7), 953–962.

van Ham, M., Tammaru, T., de Vuijst, E., & Zwiers, M. (2016). *Spatial segregation and socio-economic mobility in European cities*. IZA Discussion Papers https://www.econstor.eu/handle/10419/147963. (last accessed 11 May 2021).

Hamilton, W., Ying, R., & Leskovec, J. (2018). Representation Learning on Graphs: Methods and Applications. *IEEE Data Engineering Bulletin, 40*(3), 52–74. arXiv :1709.05584v3.

Hochstenbach, C., & Musterd, S. (2018). Gentrification and the suburbanization of poverty: Changing urban geographies through boom and bust periods. *Urban Geography, 39*(1), 26–53.

Huchzermeyer, M. (2002). Informal settlements: Production and intervention in twentieth-century Brazil and South Africa. *Latin American Perspectives, 29*(1), 83–105.

Jargowsky, P. (1996). Take the money and run: Economic segregation in U.S. metropolitan areas. *American Sociological Review, 61*, 984–998.

Jargowsky, P., & Kim, J. (2004). A measure of spatial segregation: The generalized neighborhood sorting index. In *National Poverty Center Working Paper Series, 05–3*.

Järv, O., Müürisepp, K., Ahas, R., Derudder, B., & Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies, 52*(14), 2680–2698.

Kawachi, I., & Kennedy, B. (1999). Income inequality and health: Pathways and mechanisms. *Health Services Research, 34*(1 Pt 2), 215–227.

Kingma, D., & Lei Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR), San Diego, May 7–9* (p. 2015).

Kukk, K., van Ham, M., & Tammaru, T. (2019). EthniCity of leisure: A domains approach to ethnic integration during free time activities. *Tijdschrift voor Economische en Sociale Geografie, 110*(3), 289–302.

Kwan, M. (2013). Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility. *Annals of the Association of American Geographers, 103*(5), 1078–1086.

Landrine, H., & Corral, I. (2009). Separate and unequal: Residential segregation and black health disparities. *Ethnicity & Disease, 19*, 179–184.

Li, F., & Wang, D. (2017). Measuring urban segregation based on individuals' daily activity patterns: A multidimensional approach. *Environment and Planning A, 49*(2), 467–486.

Liao, B., & Wong, D. W. (2015). Changing urban residential patterns of Chinese migrants: Shanghai, 2000–2010. *Urban Geography, 36*(1), 109–126.

Logan, J. R., Minca, E., & Adar, S. (2012). The geography of inequality: Why separate means unequal in American public schools. *Sociology of Education, 85*(3), 287–301.

Ma, L. J. C., & Wu, F. (2005). Restructuring the Chinese city: Diverse processes and reconstituted spaces. In L. J. C. Ma, & F. Wu (Eds.), *Restructuring the Chinese city: Changing society, economy and space* (pp. 1–20). London: Routledge.

Makinen, M., Waters, H., Rauch, M., Almagambetova, N., Bitran, R., Gilson, L., … RamInequalities, S. (2000). Inequalities in health care use and expenditures: Empirical data from eight developing countries and countries in transition. *The Bulletin of the World Health Organization, 78*(1), 55–65.

Massey, D., & Denton, N. (1988). The dimensions of residential segregation. *Social Forces, 67*(2), 281–315.

Matthews, S., & Yang, T. (2013). Spatial polygamy and contextual exposures (SPACEs): Promoting activity space approaches in research on place and health. *American Behavioral Scientist, 57*(8), 1057–1081.

Mayer. (2001). How the growth in income inequality increased economic segregation. In *JCPR working papers 230*. Center for Poverty Research: Northwestern University/University of Chicago Joint.

McQuoid, J., & Dijst, M. (2012). Bringing emotions to time geography: The case of mobilities of poverty. *Journal of Transport Geography, 23*, 26–34.

Oka, M., & Wong, D. (2019). Segregation: A multi-contextual and multi-faceted phenomenon in stratified societies. In T. Schwanen, & R. van Kempen (Eds.), *Handbook of urban geography* (pp. 255–280). Cheltenham: Edward Elgar Publishing Ltd.

Olteanu, M., Hazan, A., Cottrell, M., & Randon-Furling, J. (2020). Multidimensional urban segregation: Toward a neural network measure. *Neural Computing and Applications, 32*, 18179–18191.

Olteanu, M., Randon-Furling, J., & Clark, W. (2019). Segregation through the multiscalar lens. *Proceedings of the National Academy of Sciences, 116*(25), 12250–12254.

Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., & Zhang, C. (2018). Adversarially regularized graph autoencoder for graph embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, July 13–19 2018*.

Park, Y., & Kwan, M. (2018). Beyond residential segregation: A spatiotemporal approach to examining multi-contextual segregation. *Computers, Environment and Urban Systems, 71*, 98–108.

Phillips, N., Levy, B., Sampson, R., Small, M., & Wang, R. (2020). The social integration of American cities: Network measures of connectedness based on everyday mobility across neighborhoods. *Sociological Methods & Research.* https://doi.org/10.1177/0049124119852386.

Pickett, K., & Wilkinson, R. (2015). Income inequality and health: A causal review. *Social Science & Medicine, 128*, 316–326.

Quillian, L. (2012). Segregation and poverty concentration: The role of three segregations. *American Sociological Review, 77*(3), 354–379.

Ravallion, M. (2014). Income inequality in the developing world. *Science, 344*(6186), 851–855.

Reardon, S., & Bischoff, K. (2011). Income inequality and income segregation. *American Journal of Sociology, 116*(4), 1092–1153.

Rey, S., & Folch, D. (2011). Impact of spatial effects on income segregation indices. *Computers, Environment and Urban Systems, 35*(6), 431–441.

Schönfelder, S., & Axhausen, K. (2003). Activity spaces: Measures of social exclusion? *Transport Policy, 10*(4), 273–286.

Silm, S., & Ahas, R. (2014). Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers, 104*(3), 542–559.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature, 484*(7392), 96–100.

Tammaru, T., Strömgren, M., van Ham, M., & Danzer, A. M. (2016). Relations between residential and workplace segregation among newly arrived immigrant men and women. *Cities, 59*, 131–138.

Wagstaff, A., & Doorslaer, E. (2000). Income inequality and health: What does the literature tell us? *Annual Review of Public Health, 21*(1), 543–567.

Wang, D., Cui, P., & Zhu, W. (2016). *Structural deep network embedding. KDD '16, August 13–17, 2016, San Francisco, CA, USA*.

Wang, D., & Li, F. (2016). Daily activity space and exposure: A comparative study of Hong Kong's public and private housing residents' segregation in daily life. *Cities, 59*, 148–155.

Wang, D., Li, F., & Chai, Y. (2012). Activity spaces and sociospatial segregation in Beijing. *Urban Geography, 33*, 256–277.

Wang, P., Fu, Y., Zhang, J., Li, X., & Lin, D. (2018). Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs. In *ACM transactions on intelligent systems and technology 9(6): Article 63*.

Wang, Q., Edward, P., Small, M., & Sampson, R. (2018). Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences, 115*, 7735–7740.
Watson, T. (2009). Inequality and the measurement of residential segregation by income in American neighborhoods. *Review of Income and Wealth, 55*(3), 820–844.
Wheeler, C., & Jeunesse, E. (2008). Trends in neighborhood income inequality in the US: 1980–2000. *Journal of Regional Science, 48*(5), 879–891.
Wissink, B., Schwanen, T., & van Kempen, R. (2016). Beyond residential segregation: Introduction. *Cities, 59*, 126–130.
Wong, D. (2002). Modeling local segregation: A spatial interaction approach. *Geographical and Environmental Modelling, 6*(1), 81–97.
Wong, D., & Shaw, S. (2011). Measuring segregation: An activity space approach. *Journal of Geographical Systems, 13*, 127–145.
Xie, Y., & Zhou, X. (2014). Income inequality in today's China. *Proceedings of the National Academy of Science, 111*(19), 6928–6933.
Yao, Z., Fu, Y., Liu, B., Hu, W., & Xiong, H. (2018). Representing urban functions through zone embedding with human mobility patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, July 13–19 2018*.
Zhang, D., Yin, J., Zhu, X., & Zhang, C. (2016). Homophily, structure, and content augmented network representation learning. In *In proceedings of the 16th IEEE international conference on data mining* (pp. 609–618).
Zhang, T., Li, Y., Yang, H., Cui, C., Li, J., & Qiao, Q. (2020). Identifying primary public transit corridors using multi-source big transit data. *International Journal of Geographical Information Science, 34*(6), 1137–1161.
Zuo, Y., Liu, G., Lin, H., Guo, J., Hu, X., & Wu, J. (2018). Embedding temporal network via neighborhood formation. In *KDD'18, August 19–23, 2018*. London: United Kingdom.