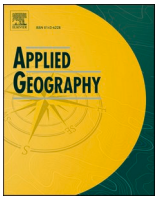
Applied Geography 163 (2024) 103179

Contents lists available at ScienceDirect

Applied Geography

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



# Income estimation based on human mobility patterns and machine learning models

Qi-Li Gao [a,b,c], Chen Zhong [b,*], Yang Yue [c], Rui Cao [d], Bowen Zhang [e]

[a] Shenzhen Audencia Financial Technology Institute(SAFTI), Shenzhen University, Shenzhen, 518060, Guangdong, China
[b] Centre for Advanced Spatial Analysis, University College London, London, WC1E 6BT, United Kingdom
[c] Shenzhen Key Laboratory of Spatial Smart Sensing & Department of Urban Informatics, Shenzhen University, Shenzhen, 518060, Guangdong, China
[d] Department of Land Surveying and Geo-Informatics & Smart Cities Research Institute, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[e] Department of Geography, King's College London, London, WC2R 2LS, United Kingdom

ARTICLE INFO

Handling Editor: Dr. Y.D. Wei

*Keywords:*
Income estimation
Human mobility patterns
Machine learning
Public transit

ABSTRACT

Sustainable and inclusive urban development requires a thorough understanding of income distribution and poverty. Recent related research has extensively explored the use of automatically generated sensor data to proxy economic activities. Notably, human mobility patterns have been found to exhibit strong associations with socioeconomic attributes and great potential for income estimation. However, the representation of complex human mobility patterns and their effectiveness in income estimation needs further investigation. To address this, we propose three representations of human mobility: mobility indicators, activity footprints, and travel graphs. These representations feed into various models, including XGBoost, a traditional machine learning model, a convolutional neural network (CNN), and a time-series graph neural network (GCRN). By leveraging public transit data from Shenzhen, our study demonstrates that graph-based representations and deep learning models outperform other approaches in income estimation. They excel in minimising information loss and handling complex data structures. Spatial contextual attributes, such as transport accessibility, are the most influential factors, while indicators related to activity extent, temporal rhythm, and intensity contribute comparatively less. In summary, this study highlights the potential of cutting-edge artificial intelligence tools and emerging human mobility data as an alternative approach to estimating income distribution and addressing poverty-related concerns.

## 1. Introduction

Obtaining socioeconomic status (SES) is of great significance in social research, urban policy and transportation management for poverty reduction and social inclusion alleviation (Ding, Huang, Zhao, & Fu, 2019). For example, decisions regarding resource allocation and the study of inequality are grounded in the way poverty and wealth are distributed geographically (Blumenstock, Cadamuro, & On, 2015). However, the lack of high-resolution quantitative socioeconomic data, particularly in developing nations, has become a significant obstacle faced by policymakers and researchers. Traditionally, obtaining SES data is either organised by national statistical institutes, usually every ten years by the mandated census, through a large number of household interviews, or by researchers via questionnaires; all are expensive, time-consuming, and labor-intensive (Xie, Xiong, & Li, 2016). Besides, small survey samples fail to accurately capture the socioeconomic status of the whole population (Lu & Pas, 1999; Wu et al., 2019). To deal with this challenge, researchers have realised that new sources of data and novel approaches are required to generate the maps of the geographic distribution of wealth and identify those areas in most need of policy intervention.

To enable the monitoring of the relevant aspects of socio-economic phenomena in quasi-real time, the last few years have witnessed a growing interest in estimating SES from big data sources, which collections are much cheaper and faster (Blumenstock, 2016; Ledesma, Garonita, Flores, Tingzon, & Dalisay, 2020; Smith-Clarke, Mashhadi, & Capra, 2014; Steele et al., 2017). Initially, satellite imagery emerged as a focal point for estimating economic activity. Nightlight images garnered

* Corresponding author.
*E-mail addresses:* qlgao@szu.edu.cn (Q.-L. Gao), c.zhong@ucl.ac.uk (C. Zhong), yueyang@szu.edu.cn (Y. Yue), rucao@polyu.edu.hk (R. Cao), bowen.zhang@kcl.ac.uk (B. Zhang).

https://doi.org/10.1016/j.apgeog.2023.103179
Received 4 September 2023; Received in revised form 3 November 2023; Accepted 10 December 2023
Available online 21 December 2023

considerable attention due to their potential, as regions with brighter lights were indicative of stronger economic activity (Chen, Xi, & Northaus, 2011). However, the method concentrates on large spatial scales, such as county and district. Subsequently, mobile phone data has been used to predict income levels based on the user's social network of communications and mobility patterns (Soto, Frias-Martinez, Virseda, & Frias-Martinez, 2011). Online geotagged social media data like Twitter shows potential for predicting income by extracting features like topics and emotion (Hinds & Joinson, 2018; Preoţiuc-Pietro, Volkova, Lampos, Bachrach, & Aletras, 2015). Particularly, data capturing human movement across the urban space is promising in estimating SES since daily activities (e.g., visiting different places, making different numbers of trips, and having different travel schedules) are highly associated with personal socioeconomic characteristics (Zhang, Cheng, & Aslam, 2019). As the modern public transport system plays an increasingly significant role in daily life, massive smart card data reveal the activity behavior and mobility pattern of individuals, hence providing an excellent opportunity to explore population demographics at finer scales (Li, Bai, Liu, Yao, & Travis Waller, 202; El Mahrsi, Côme, Oukhellou, & Verleysen, 2017).

From the methodology perspective, there is exciting potential for adapting machine learning (ML) to map wealth, fight poverty, and predict socioeconomic levels (Blumenstock, 2016). ML methods have revolutionised the way we understand urban areas and offer promising opportunities for the sustainable city agenda (Casali, Yonca, Comes, & Casali, 2022). A range of ML-based methods (e.g., Naive Bayes, Support Vector Machine, Random Forest, Boosting, NN-based deep learning, etc.) have been performed to infer demographic-socioeconomic attributes from different kinds of datasets (Allahviranloo & Recker, 2013; Solomon, Livne, Katz, Shapira, & Rokach, 2021). They demonstrated that age, gender, occupation, employment, and marital status are highly correlated to travel behavior and activity patterns (Allahviranloo & Recker, 2013; Li, Bai, Liu, Yao, & Travis Waller, 2021; Solomon et al., 2021; Zhang et al., 2019).

Nevertheless, the existing research on income estimation remains inadequate. Firstly, fine-scale estimation within the city based on large-scale population is rare. Secondly, although ML approaches can better handle complex relationships (e.g., non-linear) between human mobility patterns and socioeconomic attributes, the accurate and comprehensive representation of human mobility has become a critical issue in these works, consequently influencing the prediction results. One key challenge is extracting meaningful indicators to characterise human mobility patterns. Previous work has proposed a number of features that quantify characteristics of travel trajectories from different dimensions and capture different perspectives of activity patterns (Wu et al., 2019). However, some researchers argue that limited indicators may not be able to capture the full picture of human mobility behavior and have limitations in accurately mapping the high-resolute economic status (Zhang & Cheng, 2020). Besides, the interplay between various mobility indicators has been ignored using basic ML methods. Another disadvantage of ML approaches is the lack of interpretability of the classification results, namely in terms of what and how hidden and intermediate variables influence the predicted results. Recent advancement in explainable AI provides transparency with these "black box" models and allows for a much more insightful interpretation of relevant attributes. In this regard, to what extent human mobility can be applied to predict socioeconomic status is not well understood.

Drawing on these studies, this study extends the literature by investigating the linkage between human mobility and income status using massive travel data and reliable income data. In contrast to previous studies using a few mobility indicators and focusing on large-scale prediction (e.g., district, census track), we propose different measurements of human mobility patterns and introduce cutting-edge ML approaches. By doing so, whether and to what extent human mobility data associated with ML techniques could be used for fine-scale income or poverty mapping will be answered. The results can be utilized as the evidence base to facilitate scientific decision-making and rapid assessment of the current income geographic distribution towards urban sustainability. Assuming that the average income level in a small area (e.g., district, census tract) has a strong association with distinctive mobility patterns of people living within it, two questions are addressed in this study. First, how human mobility patterns could be represented effectively? Which combinations of human mobility representation and ML models achieve the best performance in estimating income status? By solving the two questions, this study contributes to the existing literature in two ways.

- This study presents a comprehensive solution that combines large-scale human mobility data with ML techniques to achieve precise income estimation at a granular level.
- Regarding methodology, this study proposes a variety of distinct methodologies designed for intricate human mobility patterns and demonstrates the effectiveness and performance of each method.

The subsequent sections of this paper are structured as follows: Section 2 comprehensively describes the datasets and introduces the experimental design with an analytical framework and a series of ML-based approaches. Section 3 illustrates the empirical analysis and results. Section 4 discusses the policy implications, limitations, and potential avenues for future research. Section 5 concludes with a summary.

## 2. Experimental design and data

For income estimation, a location-based analytical framework is proposed here since individuals from various datasets are anonymous and are unable to be cross-linked due to privacy concerns. Fig. 1 presents an overview of the analytical framework with four components: human pattern representation, contextual attribute calculation, income estimation, and regression model construction. The proposed methods of depicting human mobility patterns and analytical framework are generalised and could be applied to any geographic area and mobility dataset. The details of approaches for each component are provided in the following subsections.

Note: Inputs to the models (the independent variables) are annotated in the green frame, including human patterns in three different representations, namely human mobility indicators, activity footprints and travel graphs, and contextual attributes of users' residential places, are shown as (1) and (2); outputs form the models (dependent variable) are annotated in the red frame, which are the income indicators of a station catchment area, shown as (3); Finally, with the generic format of inputs and outputs, a variety of well-established regression models are adopted, shown as (4).

The basic spatial unit for analysis is a public transit station catchment area (1 km buffer zone around a station) with the following considerations. First, the spatial scale of station catchment area is much finer than the traffic analysis zones (TAZ) in the household travel survey. Second, transit users' homes and activity places could be easily associated with the most frequently visited transit stations, making the combination of anonymous travel data and survey data logical and feasible.

### 2.1. Human pattern representation

The primary human mobility dataset comes from Shenzhen, which is one of the biggest cities in China, with an area of about 2000 $km^2$ and a population of 17.5 million residents. The data was collected over six days (from Monday to Saturday) from travel card records of public transit (subway and bus) users in November 2016. Subway records contain anonymous card ID, the origin and destination station, as well as the tap-in and tap-out time of each trip. Bus records consist of anonymous card ID, boarding time, bus number without boarding location, and alighting information. To reconstruct the complete trip chains for buses, smart card data was pre-processed to infer the boarding station,

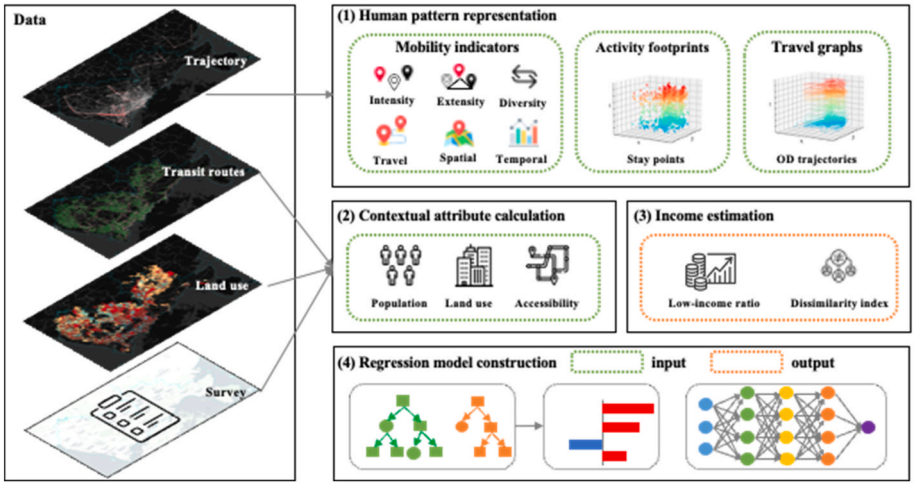



**Fig. 1.** An overview of the methodological framework.

alighting time, and station of bus trips following the method developed by Gao, Yue, Tu, Cao, and Li (2021). This approach uses bus GPS data and boarding time to infer the boarding station, and then it derives the alighting time and station based on the boarding time and location of the subsequent trip. Hence, users with only one record during a day were filtered out since the alighting point cannot be predicted. After processing, we obtained a total of 11,812,888 OD trips and 3,370,583 transit users. The OD trips are illustrated in Table A1 in the **Appendix**. With that, we further define.

- **Human mobility**: the spatial-temporal movement across the urban space, composed of a series of activity places and trips between them.
- **Activity place**: a station catchment area where individuals board and depart for their daily activities. The identification of specific locations for regular daily activities is achieved through the application of the DBSCAN clustering algorithm. Further information regarding the clustering methodology can be found in the work of Gao et al. (2021). It is observed that the residential place is visited most frequently and regularly within a week by an individual.

Based on the identified trips and activity places, we proposed three distinct ways of characterising human mobility patterns.

### 2.1.1. Human mobility indicators

Extracting meaningful indicators from movement trajectories is the most common way to characterise mobility patterns. Drawn on previous research, this study classifies mobility indicators into six categories, including activity intensity, activity extensity, activity diversity, travel efficiency, spatial location, and temporal rhythm, as shown in Table 1. All identified activity places were sorted in descending order by visit frequency, and represented by a place vector $\{(x_1, y_1, p_1), (x_2, y_2, p_2), \ldots, (x_{Npoint}, y_{Npoint}, p_{Npoint})\}$, where $p_i$ is the visit frequency of the $i$th activity places. Based on the place vector, indicators for quantifying activity extensity, activity diversity and spatial location were further calculated according to the metrics in Table 1. Travel rhythm is represented by a vector of trip frequencies that start in a series of time slots, expressed as $\{p(t_1), p(t_2), \ldots, p(t_i), \ldots, p(t_{N_t})\}$ and $p(t_i)$ is the trip frequency in the $i^{th}$ time slot. We divided a single day into $N_t$ time slots (e.g., $N_t = 12$ in this study with each time slot representing a duration of 2 h). Travel entropy was derived from the travel rhythm vector, indicating the temporal heterogeneity of travel behavior. Individuals in the travel data and survey data cannot be directly joined due to the anonymity in identification. Therefore, these mobility indicators were aggregated by average at the station scale by a transit user's home place.

**Table 1**
Human mobility indicators.

| Dimension | Indicator | Calculation |
|---|---|---|
| Activity intensity | number of trips | $N_{trip}$ |
| | number of all points | $N_{point}$ |
| | number of stay points | $N_{stay}$ |
| | number of random points | $N_{random}$ |
| Activity extensity | area of the convex polygon formed by all activity points ($km^2$) | $area = \frac{1}{2}\left\lvert \sum_{i=1}^{N_{point}} (x_i y_{i+1} - x_{i+1} y_i) \right\rvert$, $x_{N_{point}+1} = x_1, y_{N_{point}+1} = y_1$ |
| | radius of gyration ($km$) | $R_g = \sqrt{\frac{1}{N_{stay}} \sum_{i=1}^{N_{stay}} ((x_i - \overline{x})^2 + (y_i - \overline{y})^2)}$ |
| | diameter ($km$) | $diameter = \max\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\right)$, $1 \le i \le N_{point}$, $1 \le j \le N_{point}$ |
| | radius of top k points ($km$) | $R_g^k = \sqrt{\frac{1}{k} \sum_{i=1}^{k} ((x_i - \overline{x})^2 + (y_i - \overline{y})^2)}, k = 2$ |
| | average travel distance ($km$) | $dis = \frac{1}{N_{trip}} \sum_{i=1}^{N_{trip}} distance_i$ |
| | average travel time ($min$) | $time = \frac{1}{N_{trip}} \sum_{i=1}^{N_{trip}} time_i$ |
| Activity diversity | stay point entropy | $E_{activity} = -\sum_{i=1}^{N_{stay}} p_i * \log(p_i)$ |
| | travel entropy | $E_{travel} = -\sum_{i=1}^{N_t} p(t_i) * \log(p(t_i))$ |
| Spatial location | centroid of activity points | $(\overline{x}, \overline{y}) = \left(\frac{1}{N_{stay}} \sum_{i=1}^{N_{stay}} x_i, \frac{1}{N_{stay}} \sum_{i=1}^{N_{stay}} y_i\right)$ |
| | top 1 point | $(x_1, y_1)$ |
| | top 2 point | $(x_2, y_2)$ |
| Temporal rhythm | trip percentage in every time slot | $\{p(t_1), p(t_2), \ldots, p(t_i), \ldots, p(t_{N_t})\}$ |
| Travel efficiency | travel speed ($km/h$) | $speed = dis/time$ |

### 2.1.2. Activity footprints

This study further proposes a way to represent human mobility patterns using frequencies of activity points, which measures the dynamic activity footprints across the whole urban space. For each station, we first extracted all activity points of users living in the station catchment area. All activity points were then divided into different time slots by timestamp. To display the spatial pattern of activities, we partitioned the study area into 2 km grids (21 rows × 42 columns) and counted the number of active points in each grid and in each time slot. Consequently, we obtained time-series 2D images for activity footprints with $N_{station} \times N_t \times N_{row} \times N_{col}$ dimensions, as illustrated in Fig. 2.

### 2.1.3. Travel graphs

Travel flows between different places are the most intuitive repre-

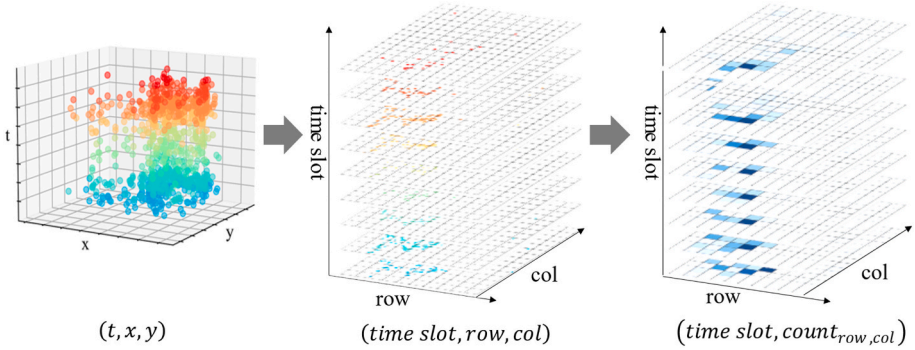



**Fig. 2.** Transformation from raw trajectories to time-series 2D images for dynamic activity footprints: $(t, x, y)$ are the time and location of an activity point; the colours of points represent temporal ranges.

sentation of human mobility patterns. Therefore, we transformed individuals' OD trajectories into travel graphs to capture dynamic spatial interactions across the urban space. As Fig. 3 depicts, individual OD trips were first aggregated at stations and then segmented into the corresponding time slots depending on the start time. The same time period of OD trips can be constructed as a directed graph $G$ ($U_{station}$, $V_{station}$, $E_{count}$), indicating the spatial interaction among different stations. Note: $(t_o, x_o, y_o, t_d, x_d, y_d)$ denote the times and locations of the origin and destination points of an OD trip. For each graph, the transit stations are nodes and the number of OD trips between stations is the edge weight. The geographic coordinates are considered as node attributes, $F_{station}\{f_1, f_2, \ldots, f_n\}$. We totally generated $N_t$ graphs $\{G_1, G_2, \ldots, G_{Nt}\}$. By doing so, we captured mobility patterns without losing much information despite sacrificing some time accuracy.

### 2.2. Contextual attribute calculation

Human mobility patterns not only vary by spatial locations (where to go) but also by the semantics of these places (what kinds of places). Therefore, we introduced some typical features of describing the contextual attributes of activity places. First, the population plays a vital role in portraying urban characteristics (Zhu, Zhao, Wang, & Al Yammahi, 2017). In this study, we used the total number of transit users as the population attribute in the station catchment area, which was derived from the household travel survey data.

Another typical feature characterizing the semantics of places is land use diversity, which is often calculated by using Shannon's entropy. Shenzhen land use data comes from the production released by Gong et al. (2020), which mapped land use at the parcel level (a finer spatial scale than TAZ) based on 10-m satellite images, POI (point of interest), OpenStreetMap, night-time lights and *Tencent* social big data in 2018. Land use is categorized into 12 types: residential, business office, commercial service, industrial, road, transportation stations, airport facilities, administrative, education, medical, sport and culture, park, and green space. Although the time stamp of land use data is different from the smart card data and travel survey, we assume that no dramatic changes in land use patterns have occurred in less than the two-year difference in collections dates.

Moreover, transit accessibility is widely recognized to be an important factor influencing how we engage with activities and access urban resources. In this study, transit accessibility is donated by the density of bus stops and the availability of the subway in a station catchment area. Shenzhen public transport network data was obtained from a map service platform in China, *GaoDe* (https://lbs.amap.com). In 2016, Shenzhen had 1838 transit routes, including 1830 bus routes and 8 subway routes, with more than 80,000 bus stations and 199 subway stations. After filtering duplicate stations and transfer stations with the same coordinates, there were about 20,000 transit stations remaining in the data set.

All the attributes for measuring the socioeconomic contexts of residential places are summarized in Table 2. $p(S_i)$ is the areal ratio of the $i$th land use type in the station catchment area.

### 2.3. Income estimation in catchment areas

The ground-truth income data comes from the Shenzhen household travel survey focusing on the travel of Shenzhen residents. The survey was conducted around November 2016, which is the same time period as we used for the smart card data. A total of 68,029 households and individuals were interviewed face to face eliciting the characteristics (e.

**Table 2**
Contextual attributes in a station catchment area.

| Dimension | Attribute | Calculation |
|---|---|---|
| Population | Number of transit users | $pop_{total}$ |
| Land use | Land use diversity | $Diversity_{LU} = -\sum_{i=1}^{12} p(S_i) * \log(p(S_i))$ |
| Transit accessibility | Number of bus stations | $Num_{bus}$ |
| | Availability of subway | $Subway = \begin{cases} 0, & Num_{subway} = 0 \\ 1, & Num_{subway} \geq 1 \end{cases}$ |

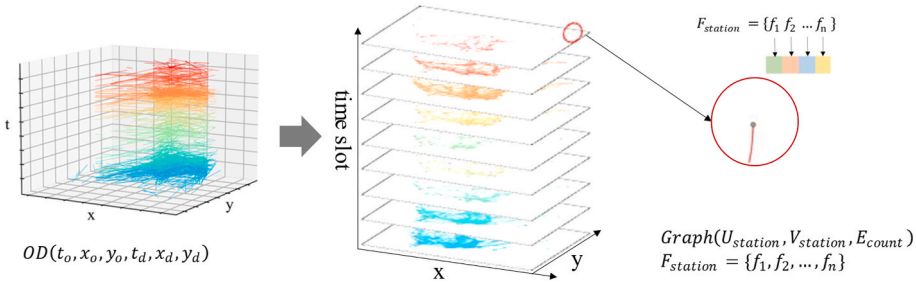



**Fig. 3.** Transformation from raw trajectories to time-series travel graphs for dynamic spatial interactions: colours of lines represent temporal ranges.

g., household annual income, number of family members over 4 years old) of households and travel-related information. Household annual income was graded into five levels (level 1= (0, 100K), level 2= (100K, 200K], level 3= (200K, 300K], level 4= (300K, 500K], level 5= (500K, +∞), currency: CNY). Residences for households were associated with 1112 TAZs within Shenzhen. The survey adopted a stratified sampling method and covered the whole city as well as all kinds of households. As this study mainly focused on those who take public transport, 42,749 transit users of the whole survey population were selected for subsequent analysis, including the calculation of income status. The sampled household survey information is illustrated in Table A2 in the **Appendix**.

Transit stations are usually located along roads and serve people from different zones. As illustrated in the example in Fig. 4, the station catchment area intersects with three TAZs (*A, B, C*) with different socioeconomic levels, and all the people who live in the intersecting areas ($I_A$, $I_B$, $I_C$) are likely to travel from this station.

However, the survey data only records the TAZ in which an individual's home locates (see **Appendix**, Table A2). To estimate the average income status of a station catchment area, we first derived the number of users in each income level for each TAZ and found all the interacted TAZs. Considering that the income level is an ordinal variable, we calculated the low-income ratio (*LR*) and dissimilarity index ($D-index$) to represent the income status of a given station and the balance between different income groups within it. For both indicators, the higher the values, the higher the proportion of low-income transit users. We referred to those whose income levels are no more than 2 (i.e., the median value of income level of all surveyed transit users) as the low-income group, and their counterparts as the high-income group. $N_{low}$ and $N_{high}$ represent the total populations of low-income and high-income transit users across the whole study area. Taking Fig. 4 as an example, the number of interacted TAZs is $N$ and the number of users in income level $i$ in TAZ $A$ is represented as $N_A^i$, thus the total population of the station catchment area is $P = \sum_{i=1}^{5}\sum_{j=1}^{N} N_j^i$. *LR* and $D-index$ were calculated according to the formulas in Table 3.

### 2.4. Regression model construction

At this point, our various data sets were in a form whereby we could begin analysis using various regression models. In response to different representations of mobility patterns (i.e., mobility indicators, activity footprint images, and travel graphs), we implemented different models because classical ML regression models are suitable for structural features but not for complicated data types like images and graphs, which are employed in this work. Deep learning models, on the other hand, can handle complex data types. For the structured activity features, this study used the linear regression (LR) model as a baseline and chose the XGBoost algorithm (XGB) to perform regression tasks, which is well known for its effectiveness and performance. For the series of activity footprints that have been transformed into 2D images, we constructed a Convolutional Neural Network (CNN) to model the relationship between human patterns and income status. In terms of the dynamic travel graphs, a Graph Convolutional Recurrent Network (GCRN) was adopted to process the regression task (Seo, Defferrard, Vandergheynst, & Bresson, 2018). The matching between input and output values and types of regression models are summarized in Fig. 5.

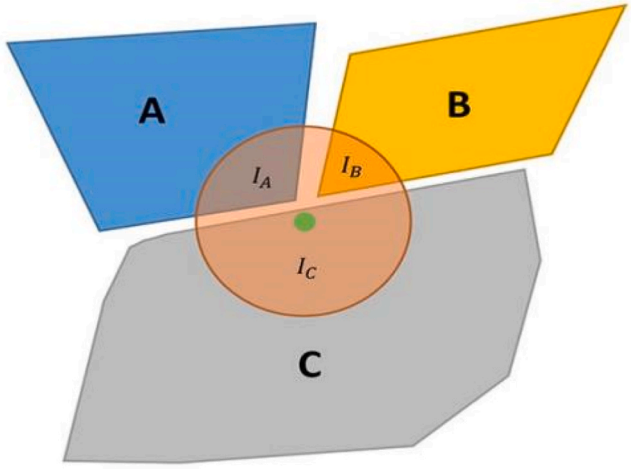



**Fig. 4.** Illustration of a transit station catchment area.

**Table 3**
Indicators of income status of station catchment areas.

| Indicators of income status | Calculation |
|---|---|
| Low-income ratio | $LR = \frac{\sum_{i=1}^{2}\sum_{j=1}^{N} N_j^i}{P}$ |
| Dissimilarity index | $D-index = \left(\frac{\sum_{i=1}^{2}\sum_{j=1}^{N} N_j^i}{N_{low}} - \frac{\sum_{i=3}^{5}\sum_{j=1}^{N} N_j^i}{N_{high}}\right) * 100$ |

For the ML models, the contextual attributes were directly extended to the mobility indicators as the input to the LR and the XGB model. In the CNN model, they served as channels combined with time slots and fed into the convolutional layers. When it comes to the GCRN model, contextual attributes were treated as node features together with the geographical coordinates of transit stations.

## 3. The empirical analysis and results

### 3.1. Feature engineering

Before modelling the relationship between income status and mobility indicators using the ML models, some procedures relating to feature dimension reduction were carried out to remove meaningless features. The first is to identify features with missing values (zero values) where greater than 60% of the data was missing. Consequently, trip percentages for the first three time slots (00:00–06:00 a.m.) were filtered out. Second, Pearson correlation analysis was carried out to identify collinear features. The heatmap of correlation coefficients of the remaining mobility indicators is presented in Fig. 6.

Generally, the indicators that characterise the same dimension of mobility pattern present high correlations (correlation coefficient >0.95). To eliminate correlations, we kept one mobility indicator for each highly correlated group. As a result, we obtained a 20-dimension mobility indicator vector for each station, including the number of stay points ($N_{stay}$), the number of random points ($N_{random}$), radius of gyration ($R_g$), area (*area*), average travel distance (*dis*), average travel time (*time*), the most frequently visited location ($(x_1, y_1)$), stay point entropy ($E_{activity}$), travel entropy ($E_{travel}$), trip percentages in time slots ($\{p(t_4), \ldots, p(t_i), \ldots, p(t_{12})\}$) and travel speed (*speed*). The mobility indicator vector thus became the basis for inputs to the regression models.

### 3.2. Spatial distribution of income status

The spatial distribution of income status allows us to better understand the spatial structure of the study area, and interpret the impact of mobility and contextual indicators. Fig. 7**(a)** and **(b)** present the spatial distributions of the low-income ratio and D-index, respectively. Both indices exhibit similar patterns in representing the spatial arrangement of income statuses. Generally, the spatial variations observed in these income indicators correspond to the urban structure of Shenzhen. The urban core centres are predominantly located in the southwestern region of the city. As a whole, there is a gradual decrease in the income indicators as the distance from the city core increases. This implies that

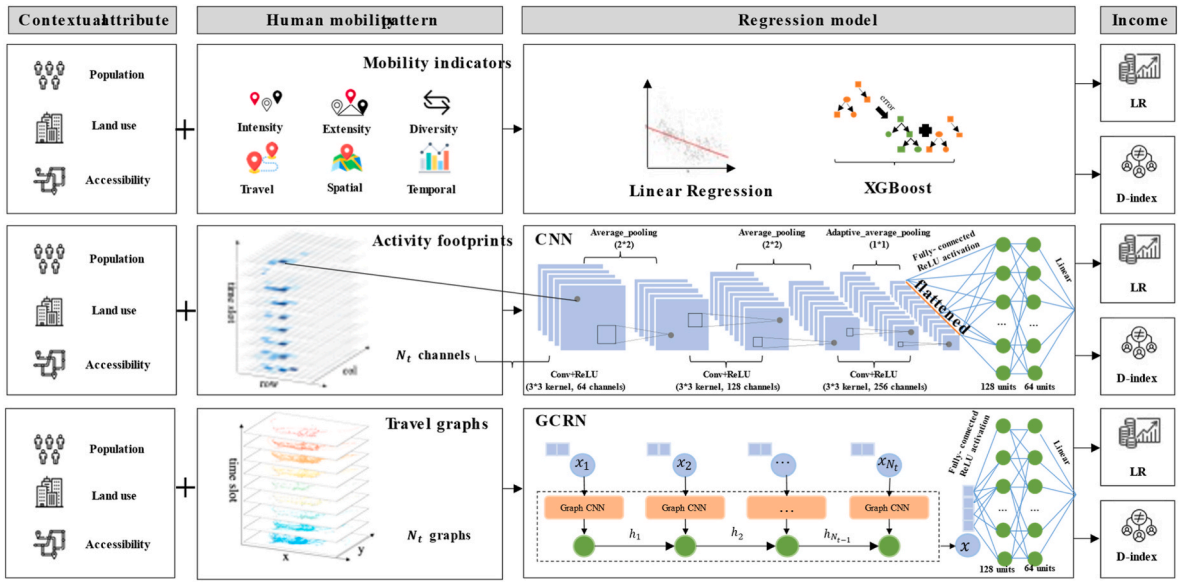



**Fig. 5.** Regression models used in this study.

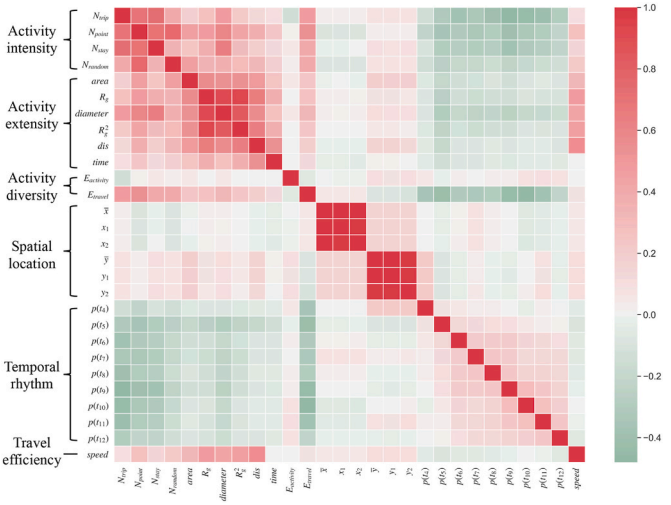



**Fig. 6.** The heatmap of correlation coefficients of mobility indicators, which are ordered by dimension summarized in Table 1.

the proportion of the low-income population tends to be higher in suburban areas compared to those residing within or in close proximity to the city centre. Nevertheless, there are also some areas on the outskirts of the city that display a more balanced income distribution. Notably, the western part of the city exhibits a greater number of areas with a higher proportion of low-income individuals compared to the eastern part, revealing a substantial spatial disparity in income levels across the city.

### 3.3. Comparison of regression results

The linear and XGB models were implemented using the SciKit-Learn packages. The deep learning models were performed using PyTorch. A 5-fold cross-validation approach was employed for all ML models. Considering that the prediction targets, pertaining to income indicators, are continuous numerical variables, the performance of the ML models was evaluated using the coefficient of determination ($R^2$) and mean square error (*MSE*).

The network architecture of the CNN model consists of eight hidden layers (three convolutional layers, three pooling layers, and two fully-connected layers). The input image is 9 (time slot) × 21 (row) × 42 (column) dimensions after filtering out three meaningless time slots (00:00–06:00 a.m.). Each convolutional layer adopted ReLU non-linear activation function and 3 × 3 kernels. The three convolutional layers respectively produced 64, 128, and 256-channel feature maps. In the first two pooling layers, the feature maps were subsampled by a scaling factor of 2, and in the third pooling layer, we adopted an adaptive pooling function. The resulting 256-channel feature map was flattened and fed into the fully connected layers (256->64->16) with the ReLU activation function. A linear function was used in the last layer to output the regression value. The GCRN model uses graph CNN for graph-structured data to identify meaningful spatial structures and recurrent neural networks (RNN) to capture dynamic patterns. The resulting embedding was then fed into the fully connected layers with the same architecture as the CNN model.

The regression outcomes for different models are presented in

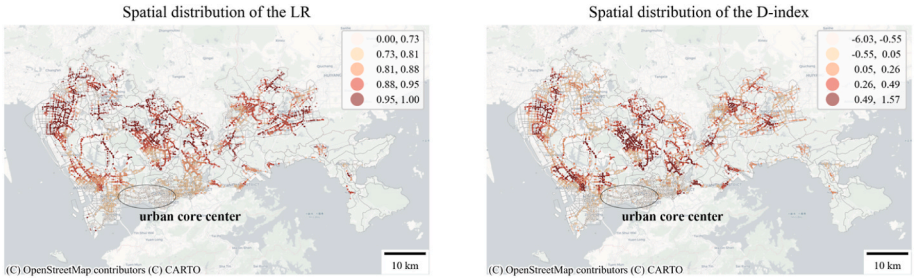



**Fig. 7.** Spatial distribution of the low-income ratio (left) and dissimilarity index (right) at stations.

Table 4. The baseline linear model exhibited the lowest $R^2$ values of 0.302 for the LR and 0.319 for the D-index. The adoption of XGB models significantly enhanced the predictive capabilities, resulting in higher $R^2$ values of 0.730 for the LR and 0.845 for the D-index. These improvements suggest the presence of complex interactions that extend beyond linear relationships between various mobility indicators and income status indices. In terms of analysing time-series activity footprints, the CNN model achieved $R^2$ values of 0.799 for the LR and 0.897 for the D-index, surpassing the performance of the linear and XGB models. To further investigate the correlation between mobility patterns and income status, the GCRN model was utilized. The GCRN model was implemented using the PyTorch geometric temporal package. Notably, the GCRN model demonstrated the highest predictive power among all the models, yielding an $R^2$ of 0.910 for the LR and 0.976 for the D-index.

Based on well-defined structured features, the XGB model, which is a tree-based ensemble model, exhibited considerable predictive capabilities by effectively handling complex interactions between variables. Remarkably, without relying on domain knowledge or feature engineering, dynamic activity footprints proved to be more effective in capturing the connection between spatiotemporal patterns and income status. However, compared to discrete activity footprints, the graph-based deep learning model displayed the highest potential in prediction tasks by capturing spatial interactions between activity places. These results underscore the efficacy of advanced ML techniques and human mobility characteristics in mapping income distribution. Furthermore, they highlight the notion that the retention of more information while characterizing movement patterns leads to increased accuracy in predicting income attributes.

### 3.4. Feature importance

Although income status and mobility were found to be highly correlated in this study, which extracted mobility indicators are strongly associated with income status remain under-explored. The notions of explainable AI provide us visibility into feature importance and how a regression model determines decisions and predictions. The Shapley Additive Explanations (SHAP) value statistic is the average marginal contribution of a feature to the model output or prediction with a different magnitude and signs across all possible coalitions (sets) of the features (Lundberg & Lee, 2017). Accordingly, the magnitude of SHAP values represents the estimation of feature importance, and the sign represents impact direction. Features with positive signs contribute to the prediction.

As shown in Fig. 8, for both measures of income status, the ranking of the importance of characteristics is generally consistent. The most important features are the spatial location of the Top 1 frequently visited station, which basically reflects the choice of residence place. The positive impact of latitude and the negative impact of longitude echo Shenzhen's urban structure, see Fig. 7. Different from some cities in Western countries, wealthy people in the biggest cities in China prefer living in the central places due to affluent urban opportunities and facilities. In general, there is an overall trend where areas in closer proximity to the city centre exhibit lower proportions of low-income transit users.

Following spatial location features, all the attributes for characterising the contexts of residential places exhibit great significance in contributing to the model output. The availability of subways, land use diversity, and the total population of transit users are negatively correlated to the values of both income indicators of station catchment areas. Conversely, the number of bus stations exhibits a positive influence. These findings imply that areas characterised by high proportions of high-income users tend to enjoy enhanced transport accessibility, lower population density, and mixed land use. In contrast, transit users residing in areas dominated by low-income populations are more likely to rely on bus transportation due to the absence of subway systems, thereby indicating income inequality in terms of access to urban facilities and opportunities.

In terms of the dimension of activity extensity, an overall trend is observed whereby areas with a higher proportion of low-income transit users exhibit a larger average travel distance. This finding aligns with the residential pattern where individuals with higher wealth tend to reside in proximity to city centres and areas with improved transport accessibility. Consequently, when utilising public transit, these individuals travel shorter distances for their daily activities. The negative relationship between travel efficiency (*speed*) and income indicators corresponds to the availability of better transport accessibility, such as the presence of subways, which is more common among high-income transit users.

Concerning the dimension of temporal rhythm, the travel frequency during specific time slots, namely 4 (6:00–8:00), 5 (6:00–8:00), and 9 (16:00–18:00), plays a significant role in predicting income status. On average, transit users residing in lower-income areas tend to start their travels earlier in the morning and return later in the evening compared to individuals residing in areas with a lower proportion of low-income users.

Regarding activity intensity, transit users in lower-income areas exhibit a higher tendency towards engaging in regular activities while participating in fewer random activities. The latter category of activities is more likely to be associated with non-mandatory activities such as leisure pursuits. The indicators of measuring activity diversity are the least important, which is consistent with other studies in Singapore and Boston, which suggest that certain mobility indicators (e.g., travel diversity) are not effective factors in measuring mobility differences by income status (Xu, Belyi, Bojic, & Ratti, 2018). The results indicate that income status plays a more important role in determining residential choice, including the location of residence and its socioeconomic contexts, than other listed mobility indicators. From the perspective of social inequality, this implies that residential differentiation might be more significant than mobility-based disparities.

**Table 4**
Regression models and results.

| Output (Income status) | Input (Human mobility pattern) | Model | $R^2$ | MSE |
|---|---|---|---|---|
| Low-income ratio (LR) | Mobility indicators | Linear (baseline) | 0.302 | 0.010 |
| | Mobility indicators | XGBoost | 0.730 | 0.004 |
| | Activity footprints | CNN | 0.799 | 0.003 |
| | Travel graphs | GCRN | 0.910 | 0.001 |
| Dissimilarity (D-index) | Mobility indicators | Linear (baseline) | 0.319 | 0.720 |
| | Mobility indicators | XGBoost | 0.845 | 0.164 |
| | Activity footprints | CNN | 0.897 | 0.095 |
| | Travel graphs | GCRN | 0.976 | 0.022 |

## 4. Discussion

### 4.1. The limitations of this study

Despite the abovementioned achievements and potentials, there exist some limitations in this study. First, this study focused on certain social groups (i.e., public transit users) and lacked concern for other social groups. In fact, our derived evidence may not be applicable to other social groups. However, the proposed framework and methods are generalised, and when data that covers all social groups (i.e., mobile phone data) becomes available, we will verify the relationships revealed in this study. The second limitation is the long-standing inadequate interpretation of deep learning models in terms of their parameters generated from various forms of learning. This study has only explained the extracted mobility indicators in the tree-based machine learning models. Although it is verified that deep learning models have advantages in handling human mobility patterns and exhibit strong powers in

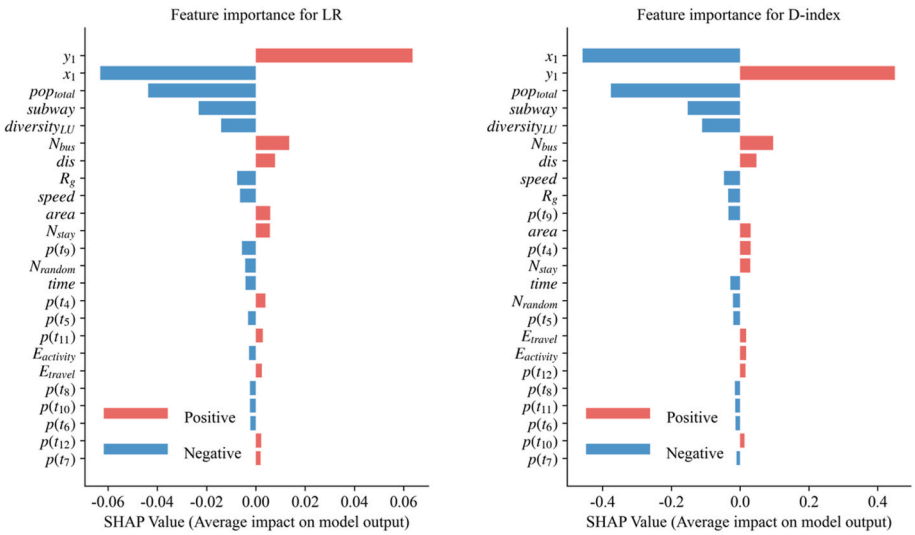



**Fig. 8.** Feature importance for low-income ratio (left) and dissimilarity index (right).

capturing complex relationships, the interaction mechanism remains unclear. This difficulty is expected to be overcome in future studies where we should focus on the further advancement of explainable AI techniques. Lastly, aside from the methods and models utilized in this study, we strongly encourage exploring alternative methods and machine learning models to assess the insights gained from the results.

### 4.2. The implications for social-spatial inequality

SES plays a vital role in shaping human activity and consequential mobility patterns. From the social perspective of sustainability, the difference in activity and human mobility patterns generates situations of isolation and social exclusion since daily activity pattern reflects the ease of access to urban opportunities (Hu et al., 2022; Núñez, Albornoz, León, & Zumelzu, 2022). For example, disadvantaged groups (e.g., low-income groups and migrants) have been found to lack access to urban facilities, employment opportunities, and essential resources (Li, Yue, Gao, Zhong, & Barros, 2022; Ta, Kwan, Lin, & Zhu, 2021). Moreover, differences in engaging in daily activities have been observed between local residents and migrants, men and women, car and transit users, and majority and racial minorities (Gao et al., 2022; Järv, Müürisepp, Ahas, Derudder, & Witlox, 2015; Lu & Pas, 1999; Wu et al., 2019). For instance, existing research suggests that low-income people often encounter significant constraints that lead to small activity spaces and fewer activity opportunities (Tao, He, Kwan, & Luo, 2020). The knowledge of differences in the daily experience of moving around the city across social groups enables improved decision-making and inclusion strategies for the allocation of urban resources (Wang, Kwan, & Hu, 2020). Therefore, understanding activity and mobility differentiation among various social groups has become a crucial aspect of achieving urban sustainability by creating a more equitable society (Gao et al., 2021; Hedman, Kadarik, Andersson, & Östh, 2021).

Besides the methodological advances in income estimating, the case study contributes to the wider debate about socio-spatial inequalities. Our observations suggest the presence of income inequality in terms of access to urban facilities and opportunities, whereby individuals in low-income areas face limited access to convenient transportation options such as subways and diverse land use, further exacerbating disparities in urban accessibility. Moreover, we found an average larger travel distance and lower travel efficiency among these areas with more low-income transit users. On average, transit users in lower-income-level areas tend to travel earlier, get off work later, and engage in fewer non-mandatory activities than those living in areas with fewer low-income users. This finding facilitates policymaking for addressing mobility-related urban inequalities (Gao et al., 2022). Besides, when considering both residential contexts and mobility characteristics, we found that residential differentiation by income is significantly greater than mobility disparities. This reminds us to reconsider urban segregation research despite the current debate about "looking beyond residential space" (Park & Kwan, 2018). As observed, higher-income areas enjoy a significant geographic advantage, which may reinforce the inequalities in access to resources near residential spaces.

## 5. Conclusion

Smart city development that leverages large spatiotemporal datasets and digital analytic tools has been highlighted as a response to the challenges in achieving sustainable urban development (Blasi, Ganzaroli, & De Noni, 2022). Since the proposal of sustainable development goals, fighting poverty has become a main concern because of the lack of up-to-date and accurate socioeconomic data. The relationship between human movement and SES could be applied to fine-grained and timely socio-economic monitoring (e.g., income) associated with the advanced AI technology (Blasi et al., 2022; Pappalardo et al., 2015).

To enable comparison among different representations of human mobility, we proposed three different measures and introduced machine learning approaches to work with more complicated data structures. In line with previous studies, we first examined some important mobility indicators based on a combination of millions of real travel records across the whole. Additionally, we have devised an enhanced depiction of mobility patterns that minimises information loss. Through an empirical study conducted in a mega city, Shenzhen, we have validated the exceptional effectiveness of our proposed framework and models in achieving precise and detailed income estimation. This alternative solution offers a viable option for rapidly developing cities, wherein the conventional survey method, which is both time-consuming and labour-intensive, can be potentially replaced by machine learning-based approaches utilising extensive human mobility big data. In addition to contributing to the existing literature on methodology, this study sheds light on socio-spatial inequality through the analysis of feature importance using explainable AI techniques. This, in turn, offers insights that can further the attainment of urban sustainable development goals.

## Declaration of competing interest

The authors have no conflict of interest.

## Acknowledgement

This research was supported by the National Natural Science Foundation of China (Grant No. 42001390) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 949670), and from ESRC under JPI Urban Europe/NSFC (Grant No. ES/T000287/1).

## Appendix

**Table A1**
The examples of transit OD trips

| Card ID | Boarding time | Boarding stop | Boarding location | Alighting time | Alighting stop | Alighting location | Distance (km) | Time (minute) | Type |
|---|---|---|---|---|---|---|---|---|---|
| 020****511 | 17:51:34 | Yannan | 114.***434<br>22.***223 | 18:21:49 | Xiameilin | 114.***677<br>22.***421 | 7.89 | 30.25 | subway |
| 020****776 | 08:08:21 | Huangbeiling | 114.***266<br>22.***805 | 08:47:47 | Shijiezhichuang | 114.***284<br>22.***844 | 17.13 | 39.43 | subway |
| 020****776 | 09:08:38 | BV10382704 | 113.***330<br>22.***939 | 09:52:48 | BV10244496 | 113.***218<br>22.***640 | 9.55 | 44.17 | bus |

**Table A2**
The examples of travel surveyed households

| Household ID | Number of family members | Income level | TAZID |
|---|---|---|---|
| 0 | 3 | 1 | 1002 |
| 1 | 1 | 1 | 1034 |
| 2 | 2 | 2 | 2005 |
| 3 | 2 | 3 | 5106 |

## References

Allahviranloo, M., & Recker, W. (2013). Daily activity pattern recognition by using support vector machines with multiple classes. *Transportation Research Part B: Methodological, 58*, 16–43. https://doi.org/10.1016/j.trb.2013.09.008

Blasi, S., Ganzaroli, A., & De Noni, I. (2022). Smartening sustainable development in cities: Strengthening the theoretical linkage between smart cities and SDGs. *Sustainable Cities and Society, 80*, Article 103793.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science, 353*(6301), 753–754. https://doi.org/10.1126/science.aah5217

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Casali, Y., Yonca, N. A., Comes, T., & Casali, Y. (2022). *Machine learning for spatial analyses in urban areas: A scoping review*. Sustainable Cities and Society, Article 104050.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences, 108*(21), 8589–8594.

Ding, S., Huang, H., Zhao, T., Fu, X., 2019. Estimating socioeconomic status via temporal-spatial mobility analysis - a case study of smart card data, In: 2019 28th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9.

El Mahrsi, M. K., Côme, E., Oukhellou, L., & Verleysen, M. (2017). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems, 18*(3), 712–728. https://doi.org/10.1109/TITS.2016.2600515

Gao, Q.-L., Yue, Y., Tu, W., Cao, J., & Li, Q.-Q. (2021). Segregation or integration? Exploring activity disparities between migrants and settled urban residents using human mobility data. *Transactions in GIS, 25*(6), 2791–2820. https://doi.org/10.1111/tgis.12760

Gao, Q.-L., Yue, Y., Zhong, C., Cao, J., Tu, W., & Li, Q.-Q. (2022). Revealing transport inequality from an activity space perspective: A study based on human mobility data. *Cities, 131*, Article 104036. https://doi.org/10.1016/j.cities.2022.104036

Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., … Feng, S. (2020). Mapping essential urban land use categories in China (EULUC-China): Preliminary results for 2018. *Science Bulletin, 65*(3), 182–187.

Hedman, L., Kadarik, K., Andersson, R., & Östh, J. (2021). Daily mobility patterns: Reducing or reproducing inequalities and segregation? *Social Inclusion, 9*(2), 208–221.

Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLoS One, 13*(11), Article e0207112. https://doi.org/10.1371/journal.pone.0207112

Hu, S., Xiong, C., Younes, H., Yang, M., Darzi, A., & Jin, Z. C. (2022). Examining spatiotemporal evolution of racial/ethnic disparities in human mobility and COVID-19 health outcomes: Evidence from the contiguous United States. *Sustainable Cities and Society, 76*, Article 103506. https://doi.org/10.1016/j.scs.2021.103506

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*, 2680–2698.

Ledesma, C., Garonita, O. L., Flores, L. J., Tingzon, I., & Dalisay, D. (2020). *Interpretable poverty mapping using social media data, satellite images, and geospatial information. arXiv preprint arXiv:2011.13563*.

Li, C., Bai, L., Liu, W., Yao, L., & Travis Waller, S. (2021). Urban mobility analytics: A deep spatial–temporal product neural network for traveler attributes inference. *Transportation Research Part C: Emerging Technologies, 124*, Article 102921. https://doi.org/10.1016/j.trc.2020.102921

Li, Q.-Q., Yue, Y., Gao, Q.-L., Zhong, C., & Barros, J. (2022). Towards a new paradigm for segregation measurement in an age of big data. *Urban Informatics, 1*(1), 5. https://doi.org/10.1007/s44212-022-00003-3

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*.

Lu, X., & Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. *Transportation Research Part A: Policy and Practice, 33*(1), 1–18. https://doi.org/10.1016/S0965-8564(98)00020-2

Núñez, F., Albornoz, E., León, J., & Zumelzu, A. (2022). Socially sustainable mobility: Strategic analysis to identify accessibility barriers. *Sustainable Cities and Society, 76*, Article 103420. https://doi.org/10.1016/j.scs.2021.103420

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabási, A.-L. (2015). Returners and explorers dichotomy in human mobility. *Nature Communications, 6*(8166). https://www.nature.com/articles/ncomms9166#supplementary-information.

Park, Y. M., & Kwan, M.-P. (2018). Beyond residential segregation: A spatiotemporal approach to examining multi-contextual segregation. *Computers, Environment and Urban Systems, 71*, 98–108. https://doi.org/10.1016/j.compenvurbsys.2018.05.001

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PLoS One, 10*(9), Article e0138717.

Seo, Y., Defferrard, M., Vandergheynst, P., & Bresson, X. (2018). Structured sequence modeling with graph convolutional recurrent networks. In *Paper presented at the international conference on neural information processing*.

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Paper presented at the proceedings of the SIGCHI conference on human factors in computing systems*. Toronto, Ontario: Canada. https://doi.org/10.1145/2556288.2557358.

Solomon, A., Livne, A., Katz, G., Shapira, B., & Rokach, L. (2021). Analyzing movement predictability using human attributes and behavioral patterns. *Computers, Environment and Urban Systems, 87*, Article 101596. https://doi.org/10.1016/j.compenvurbsys.2021.101596

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011), 2011//. *Prediction of socioeconomic levels using cell phone records*. Berlin, Heidelberg: Adaption and Personalization. Paper presented at the User Modeling.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., … Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface, 14*(127), Article 20160690. https://doi.org/10.1098/rsif.2016.0690

Ta, N., Kwan, M.-P., Lin, S., & Zhu, Q. (2021). The activity space-based segregation of migrants in suburban Shanghai. *Applied Geography, 133*, Article 102499. https://doi.org/10.1016/j.apgeog.2021.102499

Tao, S., He, S. Y., Kwan, M.-P., & Luo, S. (2020). Does low income translate into lower mobility? An investigation of activity space in Hong Kong between 2002 and 2011. *Journal of Transport Geography, 82*, Article 102583. https://doi.org/10.1016/j.jtrangeo.2019.102583

Wang, H., Kwan, M.-P., & Hu, M. (2020). Social exclusion and accessibility among low- and non-low-income groups: A case study of nanjing, China. *Cities, 101*, Article 102684. https://doi.org/10.1016/j.cities.2020.102684

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., et al. (2019). Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems, 77*, Article 101368. https://doi.org/10.1016/j.compenvurbsys.2019.101368

Xie, K., Xiong, H., & Li, C. (2016). The correlation between human mobility and socio-demographic in megacity. In *Paper presented at the 2016 IEEE international smart cities conference (ISC2)*, 12-15 Sept. 2016.

Xu, Y., Belyi, A., Bojic, I., & Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems, 72*, 51–67. https://doi.org/10.1016/j.compenvurbsys.2018.04.001

Zhang, Y., & Cheng, T. (2020). A deep learning approach to infer employment status of passengers by using smart card data. *IEEE Transactions on Intelligent Transportation Systems, 21*(2), 617–629. https://doi.org/10.1109/TITS.2019.2896460

Zhang, Y., Cheng, T., & Aslam, N. S. (2019). *Exploring the relationship between travel pattern and social-demographics using smart card data and household survey*. Paper presented at the ISPRS Geospatial Week 2019.

Zhu, P., Zhao, S., Wang, L., & Al Yammahi, S. (2017). Residential segregation and commuting patterns of migrant workers in China. *Transportation Research Part D: Transport and Environment, 52*, 586–599. https://doi.org/10.1016/j.trd.2016.11.010. Part B).