How activity pattern associates with income status? Evidence from transit smart card data and AI methods

Qi-Li Gao[1,2]; Chen Zhong[3]; Yang Yue[1,4]

[1,2] Shenzhen Key Laboratory of Spatial Smart Sensing, Shenzhen University
Shenzhen, 518060, China
Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University
Hong Kong
qili.gao@outlook.com (correspondent author)

[3] The Bartlett Centre for Advanced Spatial Analysis (CASA), University College London (UCL)
Gower Street, London WC1E 6BT, UK
c.zhong@ucl.ac.uk

[4] Department of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University
Shenzhen, 518060, China
yeuyang@szu.edu.cn

Keywords: social inequality, socioeconomic status, activity pattern, smart card data, AI.

1. Introduction

Obtaining socioeconomic status (SES) is of great significance in social research, commercial intelligence, urban policy as well as transportation management (Ding, Huang, Zhao, & Fu, 2019). In particular, the disparities in activity patterns among socioeconomic groups have been considered as an important aspect of social inequality, such as social exclusion (Schönfelder & Axhausen, 2003). Many studies have documented that human activity patterns are highly associated with personal socioeconomic factors (Goulet-Langlois, Koutsopoulos, & Zhao, 2016; Lu & Pas, 1999; Zhang, Sari Aslam, Lai, & Cheng, 2020). Despite the great contributions of these studies, the evidence is still insufficient in several aspects. One of the main reasons is the difficulty of simultaneously obtaining activity data and socioeconomic data of a large number of people. Some works relied on detailed survey-based activity data, the samples were usually too small to be representative enough (Xie, Xiong, & Li, 2016). The availability of human mobility data makes it possible to track large-scale individuals' activity behaviours, the lack of socioeconomic information in such datasets poses challenges in building the relationship between SES and activity patterns (Ghosh & Ghosh, 2017). Although there has been an increase in combining human trajectory data and travel surveys, these studies mainly focused on demographic attributes, such as gender and age, the economic features especially income are insufficient due to personal sensitivity about economic level (Wu et al., 2019). Another disadvantage is the loss of interpret-ability of the prediction results derived from AI-based methods

(Zhang et al., 2020). Therefore, how and to what extent income status influences human activity behaviours have not been well addressed.

To tackle the above mentioned challenges, this research combines smart card data and travel survey data to build the link between income level and human activities. This study involves 1) the extraction of multi-dimensional activity features to capture the spatio-temporal patterns of individuals. 2) the estimation of socioeconomic attributes to represent the contexts of activity places. 3) the prediction of machine learning models to reveal the association between income level and human activity behaviours. The methods of explainable AI are used for the interpretation of features. The framework is applied to a case study of public transit users in Shenzhen, China to validate its effectiveness.

2. Dataset

2.1 Smart card data

The primary dataset in this study is six days of smart card records of public transit users observed in November 2016. The data consists of all trips by bus and subway. Subway records contain anonymous user ID, the origin and destination station, as well the tap-in and tap-out time of each trip. Bus records contain boarding information, including boarding time, bus number and the bus line information. OD trajectories were inferred from smart card records based on spatio-temporal regularities.

2.2 Travel survey data

The Shenzhen travel survey was conducted in the same period with the smart card data. A total of 68,029 households were interviewed face to face about the characteristics of households and travel-related questions. 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: RMB).

3. Methodology

3.1 Activity pattern characterization

The overview of the methodology framework is illustrated in Figure 1. We concluded our features of characterizing human activity patterns into six categories: activity intensity, activity extensity, activity diversity, travel efficiency, spatial location and temporal rhythm. These features quantify characteristics of travel trajectories from different dimensions and capture different perspectives of activity patterns. All the activity features were aggregated at the station scale by residential place.

3.2 Socioeconomic context estimation

We first defined 1 km buffer as the catchment area of a transit station, and proposed a population-weighted approach to calculate income levels at transit stations. Land use diversity, population density and transit accessibility, donated by density of bus stops and availability of subway in the station catchment area were calculated as the socioeconomic attributes of stations.

3.3 Machine learning models for regression

This study used linear regression (LR) model as baseline, and chose three widely-used models to perform regression task, including support vector machine (SVM), random forests (RF) and XGBoost. Three kinds of methods were adopted to measure feature importance: mean decrease impurity, permutation importance and Shapley Additive exPlanations (SHAP) value.

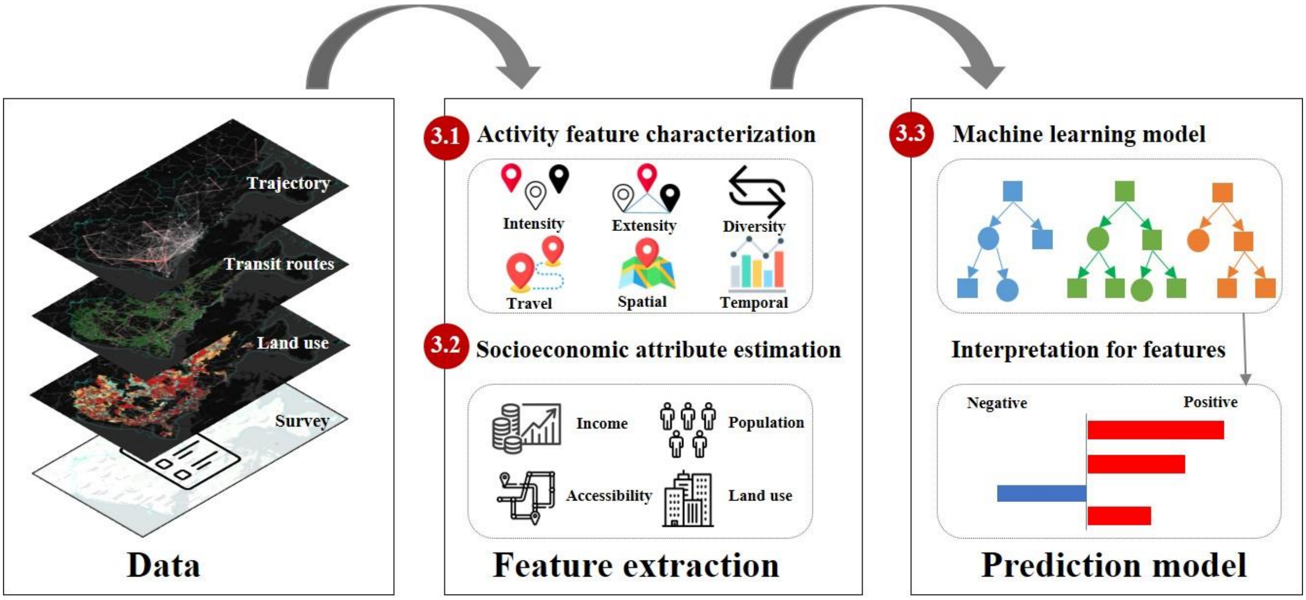

Figure 1: A overview of the methodology framework

4. Results and Discussion

4.1 Prediction results

The regression results of various combinations of features and models are illustrated in Table 1. When only activity features were considered, we yielded the lowest R^2 of 0.413 by LR model and the highest R^2 of 0.677 by RF. When we explored the association between income level and socioeconomic contexts at residence, all the models yielded better performances compared with activity features. The results indicate that income level is more related to residential socioeconomic status than other activity patterns. From the perspective of social segregation, residential segregation is more obvious than activity-related differentiation. When taking both activity features and socioeconomic features, we achieved the highest R^2 of 0.871 and the lowest MSE of 0.015 by RF. The improvement of model performance suggests that although income level primarily influence the choice of residential place, there is some impacts on other activity patterns, which should not be ignored.

Table 1: Regression model and results

Feature dimension	Model	R^2	MSE
Activity features	LR	0.413	0.069
	SVM	0.408	0.070
	RF	0.677	0.038
	XGBoost	0.671	0.039
Socioeconomic features	LR	0.263	0.095
	SVM	0.259	0.095
	RF	0.840	0.021
	XGBoost	0.759	0.031
Activity features + Socioeconomic features	LR	0.475	0.061
	SVM	0.474	0.061
	RF	0.871	0.015
	XGBoost	0.839	0.019

4.2 Feature importance

RF model performed the best in regression results, and thus was selected to interpret feature importance and the association between income variable and activity patterns. As Figure 2 shows, although feature importance varies slightly between different evaluation methods, the general trend of importance rank is: spatial location > socioeconomic context > activity extensity > activity intensity > travel efficiency > activity diversity > temporal rhythm. The results confirm that income status has the greatest impact on spatial locations and residential socioeconomic contexts. The SHAP value shows that wealthy people prefer the locations with diverse urban functions, better transport accessibility and low population density.

5. Conclusion

This study explored the connection between income status and various activity patterns by combining human mobility data and travel survey data. The results imply that income level is highly associated with human activity behaviours. However, spatial location and socioeconomic contexts (population density, land use diversity and transport accessibility) play the most important role in determining the prediction performance, suggesting that residential segregation by income is more significant than other activity-related differentiation.

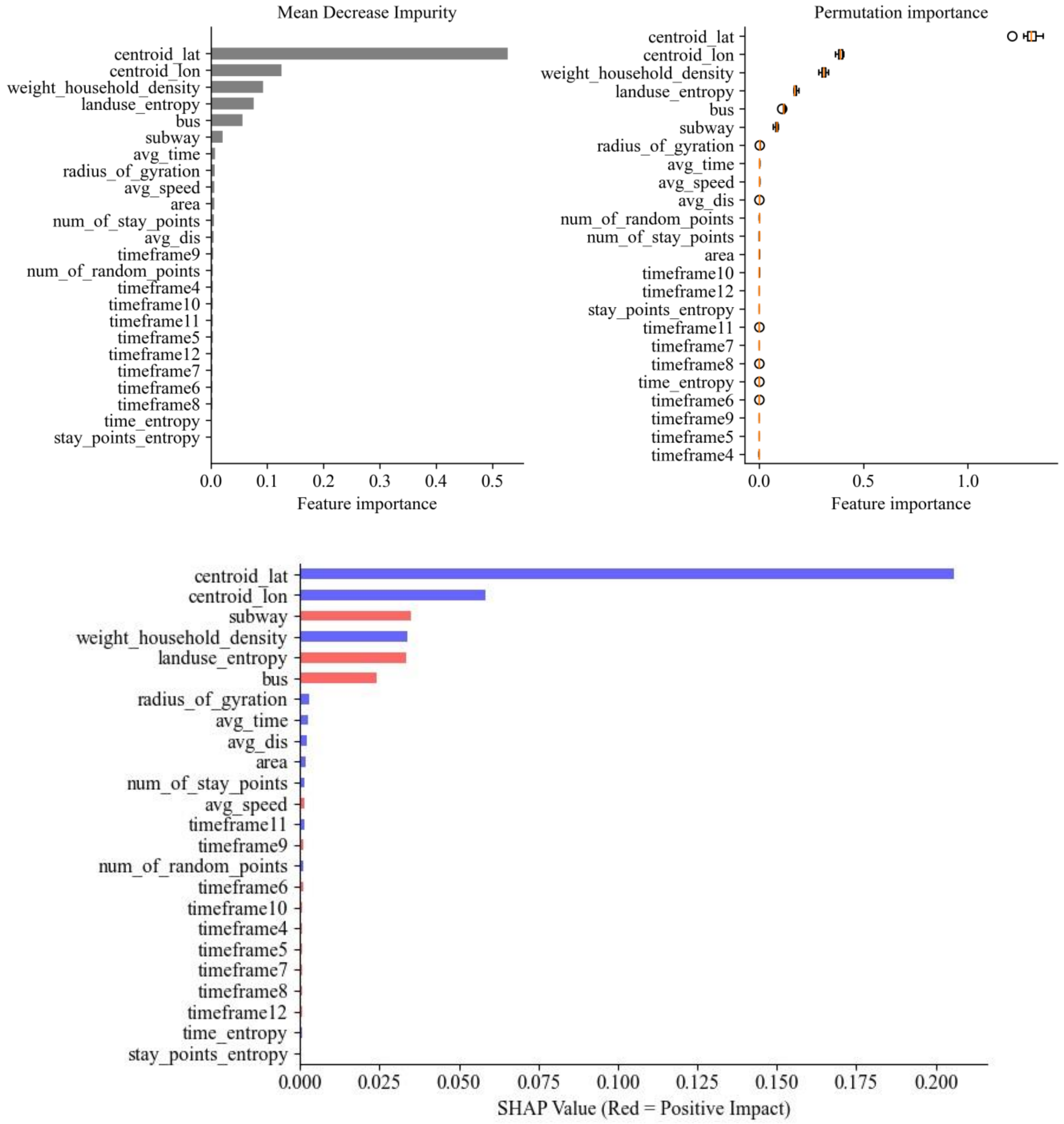

Figure 2. Feature importance

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 42001390, 71961137003) and the Postdoctoral Science Foundation of China (Grant No. 2021M692163) 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).

References

Ghosh, S., & Ghosh, S. K. (2017). Modeling of human movement behavioural knowledge from GPS traces for categorizing mobile users. Paper presented at the Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia.

Goulet-Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. Transportation research part C: emerging technologies, 64, 1-16.

Lu, X., & Pas, E. I. (1999). Socio-demographics, activity participation and travel behavior. Transportation Research Part A: Policy and Practice, 33(1), 1-18.

Schönfelder, S., & Axhausen, K. W. (2003). Activity spaces: measures of social exclusion? Transport Policy, 10(4), 273-286.

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., & Liu, Y. (2019). Inferring demographics from human trajectories and geographical context. Computers, Environment and Urban Systems, 77, 101368.

Xie, K., Xiong, H., & Li, C. (2016). The correlation between human mobility and socio-demographic in megacity. Paper presented at the 2016 IEEE International Smart Cities Conference (ISC2).

Zhang, Y., Sari Aslam, N., Lai, J., & Cheng, T. (2020). You are how you travel: A multi-task learning framework for Geodemographic inference using transit smart card data. Computers, Environment and Urban Systems, 83, 101517.