Slides

LIM Jing Zhi   Lucas SHEN

Outline 1/n

- *<Title of study>*:
  - Mobile phone data in Singapore (small densely urban city)
  - Residential vs experienced segregation
  - What is the difference?
  - What explains the difference?
- Data
  - Mobile phone GPS ping records
  - Jan–Mar 2020 (3 months/91 days)
  - Neighborhood-level census demographics & characteristics
- Segregation measures
  - Existing measures exists (Massey and Denton 1988)
  - **Exposure measure**: probability a type B individual will encounter a type W individual based on shared residential areas
  - Low exposure $\rightarrow$ high segregation

Outline 2/n

- Why it matters (aka motivation)
  - Labor market opportunities (Banerjee and Ingram 2018; Hensvik and Skans 2016)
  - Segregation worsens socioeconomic markers such as schooling, employment, and marriage (Chay et al. 2014; Chetty et al. 2016)
  - Exposure reduces discrimination (Rao 2019)
  - Exposure increases knowledge spillovers (Atkin et al. 2022)
- Revisiting segregation:
  - Segregation is usually based on census/residency (Hutchens 2001; Jones and Pebley 2014; Krivo et al. 2013; Palmer et al. 2013; Sin 2002)
  - Social mixing in a physical space
  - Experienced segregation—where I go—is different to residential segregation—where I live
  - Unsupervised learning algorithm to retrieve social clusters
  - Adapted experienced segregation measure
- Our findings:
  - Measures based on residence/census overstates experienced segregation
  - Key neighborhood amenities and travel accessibility affect segregation

Outline 3/n

- Segregation w/o accounting for day-to-day mixing is overstated:
  - Exposure based on real-time movement $>$ exposure based on residence
  - Experienced segregation $<$ wealth segregation
  - When we do not account for the choice in where people go in their day-to-day lives, segregation is overstated
- Segregation is linked to neighborhood amenities and travel access:
  - Access to transit stations linked to higher exposure of poor to wealthy
  - Travel inequity linked to higher experienced segregation
  - Parks linked to higher experienced segregation
- Asymmetries in experienced segregation:
  - Exposure of wealthy to poor much less linked to neighborhood factors than exposure of poor to wealthy

Data 1/4

- **Backbone of Data**: CITYDATA.ai GPS pings
  - Neighborhood-daily level
  - Jan–Mar 2020 (3 months, 91 days)
  - GPS pings $\longrightarrow$ O–D flows (Lee, Lim, and Shen 2021; Lim and Shen 2022)
  - Sparsity: Trim daily O–D flows + tests of representativeness (Lim and Shen 2022)
- Type: by income/wealth
  - Residence type: 1–2 room public flats vs private houses
  - House micro-transactions
- Other data: Neighborhood-level (200) characteristics
  - Age, gender, ethnic (census)
  - Neighborhood amenities (official shape files)
  - Crowd-sourced F&B POIs (OSM)
  - Travel access (from $\sim$2.5m H3-9 pairs)

Data 2/4: CITYDATA.ai GPS ping records

```
/.
├── 2020-01-01
│   ├── admiralty_subzone_sg
│   │   └── admiralty_subzone_sg_2020_01_01_deviceList.csv.gz
│   └── airport_road_subzone_sg
│       └── airport_road_subzone_sg_2020_01_01_deviceList.csv.gz
├── 2020-01-02
│   └── ...
├── ...
└── 2020-03-31
    ├── admiralty_subzone_sg
    │   └── admiralty_subzone_sg_2020_03_31_deviceList.csv.gz
    ├── airport_road_subzone_sg
    │   └── airport_road_subzone_sg_2020_03_31_deviceList.csv.gz
    └── ...
```

Example flat file:

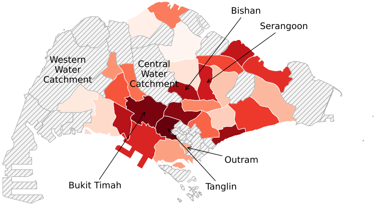
admiralty_subzone_sg_2020_01_01_deviceList.csv.gz

Hash	OS
8E545E1C31F91F777C894B3BD2C2E7D7044CC9DD	Android
40BD001563085FC35165329EA1FF5C5ECBDBBEEF	iOS
...	...
...	...
1D372C3AA0A28A5B7418A01405C621CCD523F73E	Android

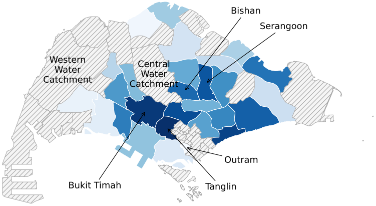
O-D panel

- Infer home using most frequent appearance
- Inter-neighborhood presence $\longrightarrow$ O-D flows

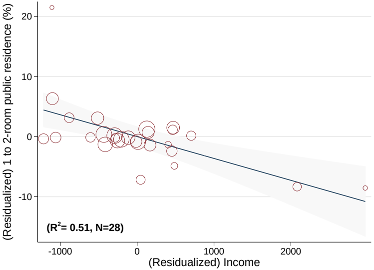
Residence type 1/2

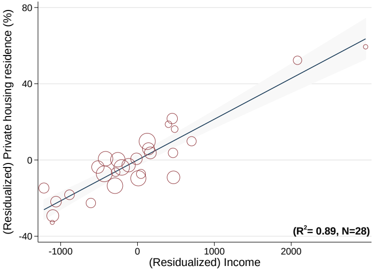


(a) Income

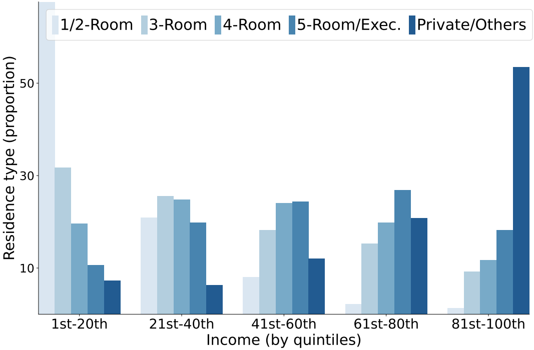


(b) Private residence





Residence type 2/2



Proxying income with residence type

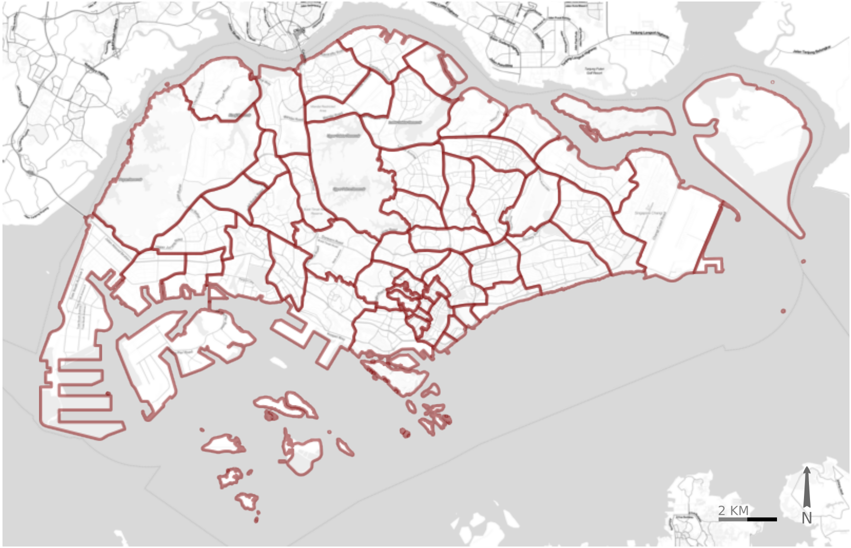
Residence type

- Proportion of residents living in 1–2-room public highrise flats (HDB)
- Residence data available at $\sim$200 neighborhood level
- L = public 1–2 rm residents
- H = private housing residents

Geography

- **5 regions**
- 55 census areas
- $\approx$ 300 subzones
- $\approx$ 200 neighbourhoods ($\approx 1 km^2$)

Geography



- 5 regions
- **55 census areas**
- $\approx$ 300 subzones
- $\approx$ 200 neighbourhoods ($\approx 1 km^2$)

Geography

- 5 regions
- 55 census areas
- $\approx$ **300 subzones**
- $\approx$ 200 neighbourhoods ($\approx 1 km^2$

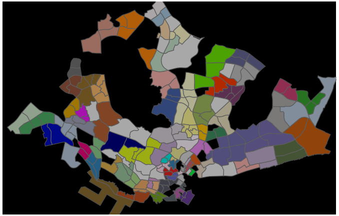
Geography

- 5 regions
- 55 census areas
- $\approx$ 300 subzones
- $\approx$ **200 neighbourhoods** ($\approx 1 km^2$)

Retrieving Social Clusters 1/4

Resident (Census planning area)

Real-time (social clusters)

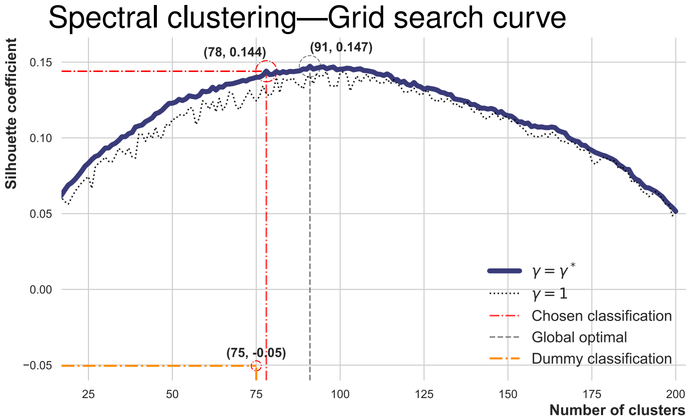


- Social clusters = clusters of neighborhoods where people are more likely to visit
- Clustering to study urban mobility & behavior (Di Clemente et al. 2018; Wang et al. 2018)

1. GPS pings $\longrightarrow$ O-D flows
2. O–D flows $\longrightarrow$ Distance matrix (Adachi et al. 2020; Tolbert and Sizer 1996)
3. Distance matrix $\longrightarrow$ Affinity matrix
4. Affinity matrix $\longrightarrow$ Spectral clustering (Pedregosa et al. 2011)

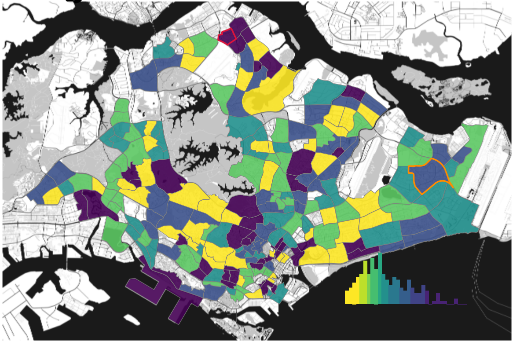
- Residential boundaries $\neq$ social boundaries

Retrieving Social Clusters 2/4



- Unsupervised
- Silhouette score
- Search over K = number of clusters
- Search over γ = coefficient in rbf kernel
- Preferred: K = 78

Retrieving Social Clusters 3/4



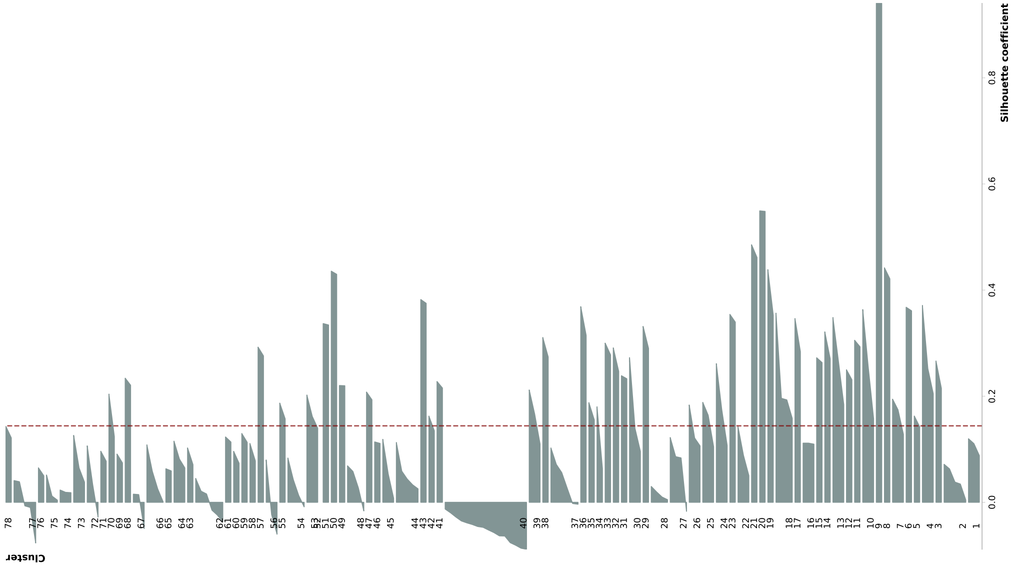
Evaluating clusters via silhouette score

(Rousseeuw 1987)

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \in [-1, 1]$$

- b_i = mean intra-cluster distance
- a_i = mean inter-cluster distance
- Denominator normalizes score to [-1,1]
- 400% better than dummy classifications

Retrieving Social Clusters 4/4



Quantifying social segregation 1/2: Residential Exposure (L to H)

$$E_a^{LH} = \frac{1}{|\mathcal{L}_a|} \sum_{i \in \mathcal{L}_a} s_{c(i)}^H \qquad (1)$$

- Residential exposure of type L to type H (traditional measure)
  (Athey et al. 2020; Massey and Denton 1988)
- $|\cdot|$ = size
- $s_{c(i)}^H$ = share of H in individual i's census unit c (eg neighborhoods)
- E = prob. that L physically come into contact with H (Massey and Denton 1988)
- a = census planning area of neighborhoods

Quantifying social segregation 2/2: Real-time Exposure (L to H)

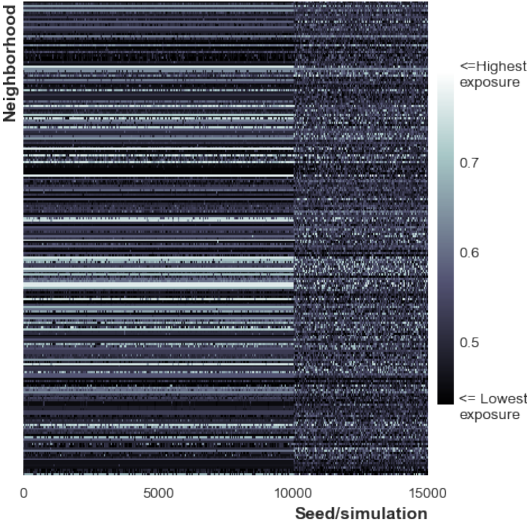
$$\tilde{E}_a^{LH} = \frac{1}{|\tilde{\mathcal{L}}_a|} \sum_{i \in \tilde{\mathcal{L}}_a} \tilde{s}_{c(i)}^{H} \tag{2}$$

- Real-time exposure is analogous— $\tilde{\cdot}$ denotes measures imputed using GPS pings
- Eg $\tilde{s}_{c(i)}^{H}$ imputed using flows from GPS pings:

$$\tilde{n}_d^L = \sum_d \underbrace{\pi_{od} \cdot \underbrace{\pi_o^L \cdot Devices_o}_{\text{share of devices imputed as type L}}}_{\text{share of imputed L-type that visit d}}$$

- a = social cluster of neighborhoods (from spectral clustering)

Real-time Exposure: Sensitivity of clustering



- Testing sensitivity of spectral clustering
- Spectral clustering uses K-means, which is sensitive to (pseudo-random) init.
- Iterations 0–99,999: Different init. for social clusters
- Iterations 10,000–15,000: Different init. for dummy clusters

Quantifying Experienced Segregation 1/2

Wealth segregation: $$s_d = \underbrace{\frac{2}{3}}_{\substack{\text{norm.}\\ \text{constant}}} \sum_q \underbrace{\left|\pi_{dq} - \frac{1}{4}\right|}_{\substack{\text{disparity =}\\ \text{deviation from}\\ \text{equal distr.}}} \in [0,1] \quad \text{(Morales et al. 2019)} \tag{3}$$

- Wealth quartile inferred by origin neighborhood & house microtransactions
- π_{dq} is proportion of visitors to neighborhood d from wealth quartile q:

$$\pi_{dq} = \frac{\overbrace{\sum_o \pi_{od} \cdot \mathbf{1}\{q_o = q\} \cdot Devices}^{\text{devices in wealth quartile q from o to d}}}{\sum_o \pi_{od} \cdot Devices}$$

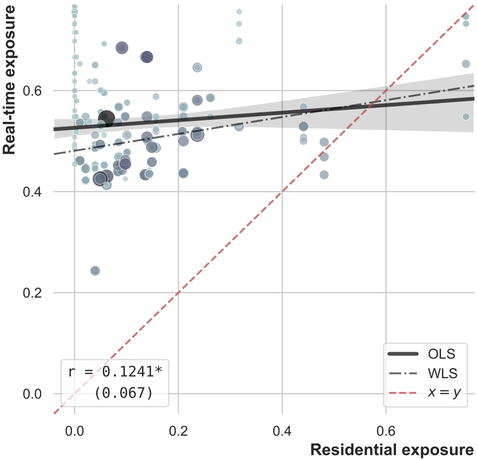
Quantifying Experienced Segregation 2/2

$$\textbf{Experienced segregation: } s_o = \underbrace{\frac{2}{3}}_{\text{norm. constant}} \sum_q \underbrace{\left| \pi_{oq} - \frac{1}{4} \right|}_{\text{disparity = deviation from equal distr.}} \in [0, 1] \quad \text{(Morales et al. 2019)} \tag{4}$$

- π_{oq} is the probability that someone from o will be exposed to someone from wealth quartile q
- Experienced segregation = where I go *And* where other people go

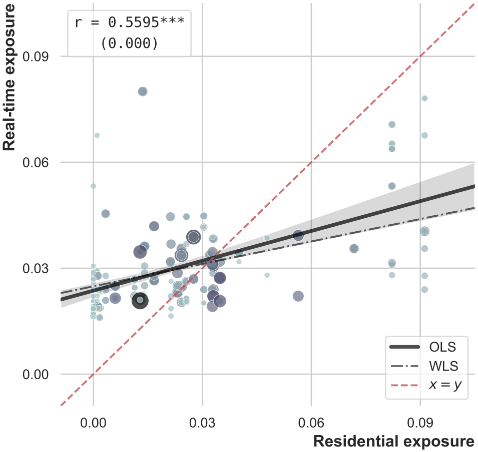
$$\pi_{oq} = \sum_d \underbrace{\pi_{od}}_{\text{prob. of flow from o to d}} \cdot \underbrace{\pi_{dq}}_{\text{prob. visitor from d is from wealth quartile q}}$$

Social segregation is overstated?



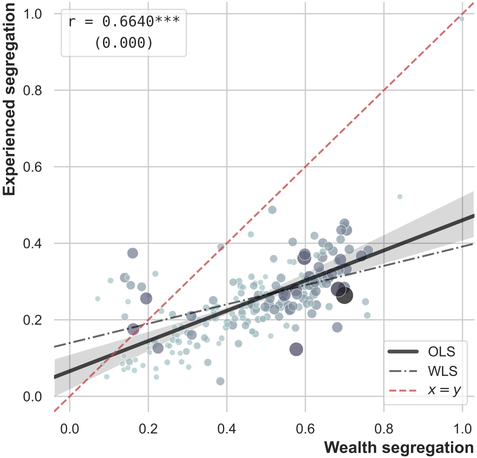
- Parity plot—Real-time vs residential
- **Poor to wealthy**
- Real-time exposure of poor to wealthy is higher than residential exposure of poor to wealthy

Social segregation is overstated?



- Parity plot—Real-time vs residential
- **Wealthy to poor**
- Real-time exposure of wealthy to poor is (mostly) higher than residential wealthy to poor
- Asymmetric—(im)parity less clear than poor to wealthy

Experienced segregation



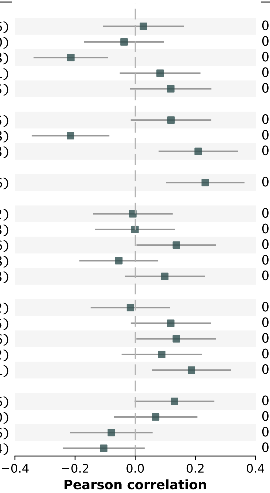
- Parity plot—Experienced vs wealth segregation
- Experienced segregation is lower than wealth segregation

Correlates of Residuals 1/n

- We describe the factors that correlate with lower segregation based on our measure of real-time exposure and experienced segregation
- Residualise real-time measures using their residential counterparts
- Reporting correlates of residuals with neighborhood factors
- Factors that stand out:
  - National parks
  - F&B
  - Travel access

Real-time exposure—Poor to Wealthy

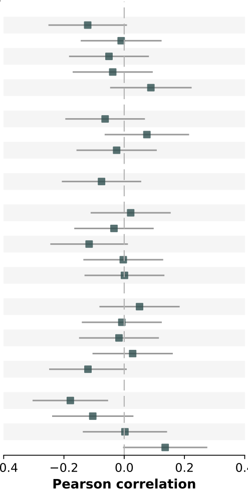
Variable	Est.(95% Conf. Int.)	P-value	N
Age group			
Age group 0–9	0.03(-0.11 to 0.16)	0.69	213
Age group 10–19	-0.04(-0.17 to 0.10)	0.59	213
Age group 20–39	-0.21(-0.34 to -0.08)	0.0***	215
Age group 40–64	0.08(-0.05 to 0.21)	0.22	215
Age group > 65	0.12(-0.02 to 0.25)	0.08*	212
Ethnic mix			
% majority ethnic	0.12(-0.02 to 0.25)	0.08*	215
% minority ethnic x	-0.22(-0.34 to -0.08)	0.0***	198
% minority ethnic y	0.21(0.08 to 0.33)	0.0***	215
Gender mix			
% female	0.23(0.10 to 0.36)	0.0***	215
Amenities			
Libraries	-0.01(-0.14 to 0.12)	0.91	219
Supermarkets	-0.00(-0.13 to 0.13)	0.99	219
National parks	0.14(0.00 to 0.26)	0.04**	219
Sport facilities	-0.05(-0.19 to 0.08)	0.42	219
Tourist attractions	0.10(-0.03 to 0.23)	0.14	219
Food & beverage			
Hawkers	-0.02(-0.15 to 0.12)	0.81	219
Restaurants	0.12(-0.01 to 0.25)	0.08*	218
Cafes	0.14(0.00 to 0.26)	0.04**	218
Bars & pubs	0.09(-0.05 to 0.22)	0.19	218
Fast-foods	0.19(0.06 to 0.31)	0.0***	218
Travel access			
Transit stations	0.13(-0.00 to 0.26)	0.05**	219
Transit time	0.07(-0.07 to 0.20)	0.33	201
Driving (4w) time	-0.08(-0.22 to 0.06)	0.26	197
Travel inequity	-0.10(-0.24 to 0.04)	0.14	197



- Neighborhoods w/ more “young people” have lower exposure
- Ethnic mix matters
- Neighborhoods with more parks have more exposure

Real-time exposure—Wealthy to Poor

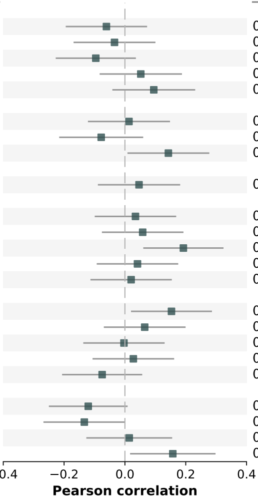
Variable	Est.(95% Conf. Int.)	P-value	N
Age group			
Age group 0–9	-0.12(-0.25 to 0.01)	0.07*	213
Age group 10–19	-0.01(-0.14 to 0.12)	0.88	213
Age group 20–39	-0.05(-0.18 to 0.08)	0.46	215
Age group 40–64	-0.04(-0.17 to 0.10)	0.58	215
Age group > 65	0.09(-0.05 to 0.22)	0.19	212
Ethnic mix			
% majority ethnic	-0.06(-0.20 to 0.07)	0.35	215
% minority ethnic x	0.08(-0.06 to 0.21)	0.29	198
% minority ethnic y	-0.03(-0.16 to 0.11)	0.71	215
Gender mix			
% female	-0.08(-0.21 to 0.06)	0.27	215
Amenities			
Libraries	0.02(-0.11 to 0.15)	0.75	219
Supermarkets	-0.03(-0.17 to 0.10)	0.62	219
National parks	-0.12(-0.25 to 0.02)	0.08*	219
Sport facilities	-0.00(-0.14 to 0.13)	0.96	219
Tourist attractions	0.00(-0.13 to 0.13)	0.99	219
Food & beverage			
Hawkers	0.05(-0.08 to 0.18)	0.45	219
Restaurants	-0.01(-0.14 to 0.13)	0.91	218
Cafes	-0.02(-0.15 to 0.12)	0.8	218
Bars & pubs	0.03(-0.11 to 0.16)	0.68	218
Fast-foods	-0.12(-0.25 to 0.01)	0.07*	218
Travel access			
Transit stations	-0.18(-0.30 to -0.05)	0.01***	219
Transit time	-0.10(-0.24 to 0.03)	0.14	201
Driving (4w) time	0.00(-0.14 to 0.14)	0.97	197
Travel inequity	0.14(-0.00 to 0.27)	0.05*	197



- More transit stations—lower exposure

Experienced segregation

Variable	Est.(95% Conf. Int.)	P-value	N
Age group			
Age group 0–9	-0.06(-0.19 to 0.08)	0.38	210
Age group 10–19	-0.03(-0.17 to 0.10)	0.62	210
Age group 20–39	-0.10(-0.23 to 0.04)	0.16	212
Age group 40–64	0.05(-0.08 to 0.19)	0.45	212
Age group > 65	0.09(-0.04 to 0.23)	0.17	209
Ethnic mix			
% majority ethnic	0.01(-0.12 to 0.15)	0.85	212
% minority ethnic x	-0.08(-0.22 to 0.06)	0.27	195
% minority ethnic y	0.14(0.01 to 0.27)	0.03**	212
Gender mix			
% female	0.05(-0.09 to 0.18)	0.5	212
Amenities			
Libraries	0.03(-0.10 to 0.17)	0.61	216
Supermarkets	0.06(-0.08 to 0.19)	0.39	216
National parks	0.19(0.06 to 0.32)	0.0***	216
Sport facilities	0.04(-0.09 to 0.17)	0.54	216
Tourist attractions	0.02(-0.11 to 0.15)	0.76	216
Food & beverage			
Hawkers	0.15(0.02 to 0.28)	0.02**	216
Restaurants	0.07(-0.07 to 0.20)	0.34	215
Cafes	-0.00(-0.14 to 0.13)	0.96	215
Bars & pubs	0.03(-0.11 to 0.16)	0.68	215
Fast-foods	-0.07(-0.21 to 0.06)	0.27	215
Travel access			
Transit stations	-0.12(-0.25 to 0.01)	0.07*	216
Transit time	-0.13(-0.27 to 0.01)	0.06*	198
Driving (4w) time	0.01(-0.13 to 0.15)	0.84	194
Travel inequity	0.16(0.02 to 0.29)	0.02**	194



- Parks
- Hawkers
- Travel inequity

Discussion

Are flows pro-social?

- Social exposure zones can capture daily trip hops
- Trip hops constitute an important form of interaction where individuals share physical activity space

  (Athey et al. 2020; Cagney et al. 2020)

Discussion

Experiential vs. physical segregation

- GPS and geolocation data allows us to incorporate the fact that segregation is also behavioural and mobility choices
- Real-time based measures of exposure and segregation based on experience are lower than their residential counterparts

  (Athey et al. 2020)
- Segregation is more than where we live
- What does this say about housing policies that uses quotas to socially integrate?

Dicussion

The bigger (areal) picture

- Why do environmental/areal factors matter?
- Where we live matters

  (hence all the lit. on neighborhood choices, e.g., Agostinelli et al. 2021; Ferreira and Wong 2022)
- Zip code > genetic code?
- Singapore: highly connected + densely urban + mix of urban planning + social policy = consciously created common spaces. What does this say about physical vs. experiential segregation and how it relates to other holistic health & population outcomes?

Adachi, Daisuke, Taiyo Fukai, Daiji Kawaguchi, and Yukiko Umeno Saito. 2020. "Commuting Zones in Japan." https://daisukeadachi.github.io/assets/papers/commuting_zone_latest.pdf.

Agostinelli, Francesco, Margaux Luflade, and Paolo Martellini. 2021. "On the Spatial Determinants of Educational Access." Working Papers 2021-042, Human Capital and Economic Opportunity Working Group, https://ideas.repec.org/p/hka/wpaper/2021-042.html.

Athey, Susan, Billy A Ferguson, Matthew Gentzkow, and Tobias Schmidt. 2020. "Experienced segregation." 10.3386/w27572.

Atkin, David, Keith Chen, and Anton Popov. 2022. "The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley." 10.3386/w30147.

Banerjee, Mitali, and Paul Ingram. 2018. "Fame as an illusion of creativity: Evidence from the pioneers of abstract art." https://dx.doi.org/10.2139/ssrn.3258318.

Cagney, Kathleen A., Erin York Cornwell, Alyssa W. Goldman, and Liang Cai. 2020. "Urban Mobility and Activity Space." *Annual Review of Sociology* 46 (1): 623–648. 10.1146/annurev-soc-121919-054848.

Chay, Kenneth Y, Jonathan Guryan, and Bhashkar Mazumder. 2014. "Early life environment and racial inequality in education and earnings in the United States." Working Paper 20539, National Bureau of Economic Research. 10.3386/w20539.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment." *American Economic Review* 106 (4): 855–902. 10.1257/aer.20150572.

Di Clemente, Riccardo, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C. González. 2018. "Sequences of purchases in credit card data reveal lifestyles in urban populations." *Nature Communications* 9 (1): 1–8. 10.1038/s41467-018-05690-8.

Ferreira, Fernando V, and Maisy Wong. 2022. "Neighborhood Choice After COVID: The Role of Rents, Amenities, and Work-From-Home." Working Paper 29960, National Bureau of Economic Research. 10.3386/w29960.

Hensvik, Lena, and Oskar Nordström Skans. 2016. "Social networks, employee selection, and labor market outcomes." *Journal of Labor Economics* 34 (4): 825–867. 10.1086/686253.

Hutchens, Robert. 2001. "Numerical measures of segregation: desirable properties and their implications." *Mathematical social sciences* 42 (1): 13–29. https://doi.org/10.1016/S0165-4896(00)00070-6.

Jones, Malia, and Anne R Pebley. 2014. "Redefining neighborhoods using common destinations: social characteristics of activity spaces and home census tracts compared." *Demography* 51 (3): 727–752. 10.1007/s13524-014-0283-z.

Krivo, Lauren J, Heather M Washington, Ruth D Peterson, Christopher R Browning, Catherine A Calder, and Mei-Po Kwan. 2013. "Social isolation of disadvantage and advantage: The reproduction of inequality in urban space." *Social Forces* 92 (1): 141–164. 10.1093/sf/sot043.

Lee, Shu En, Jing Zhi Lim, and Lucas Shen. 2021. "Segregation Across Neighborhoods in a Small City." *Asia Competitiveness Institute Research Paper Series* Research Paper #07-2021, https://lkyspp.nus.edu.sg/docs/default-source/aci/acirp202107.pdf?sfvrsn=a862240a_2.

Lim, Jing Zhi, and Lucas Shen. 2022. "Neighborhood Mismatch and Visits in a Small City."

Massey, Douglas S., and Nancy A. Denton. 1988. "The Dimensions of Residential Segregation." *Social Forces* 67 (2): 281–315. 10.1093/sf/67.2.281.

Morales, Alfredo J, Xiaowen Dong, Yaneer Bar-Yam, and Alex 'Sandy' Pentland. 2019. "Segregation and polarization in urban areas." *Royal Society Open Science* 6 (10): 190573. 10.1098/rsos.190573.

Palmer, John R B, Thomas J Espenshade, Frederic Bartumeus, Chang Y Chung, Necati Ercan Ozgencil, and Kathleen Li. 2013. "New approaches to human mobility: Using mobile phones for demographic research." *Demography* 50 (3): 1105–1128. 10.1007/s13524-012-0175-z.

Pedregosa, F., G. Varoquaux, A. Gramfort et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12 2825–2830. 10.5555/1953048.2078195.

Rao, Gautam. 2019. "Familiarity Does Not Breed Contempt: Generosity, Discrimination, and Diversity in Delhi Schools." *American Economic Review* 109 (3): 774–809. 10.1257/aer.20180044.

Rousseeuw, Peter J. 1987. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics* 20 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

Shen, Lucas. 2022. "Forestplot: A Python package to make publication-ready but customizable forest plots." 10.5281/zenodo.7101525.

Sin, Chih Hoong. 2002. "Segregation and marginalisation within public housing: The disadvantaged in Bedok New Town, Singapore." *Housing Studies* 17 (2): 267–288. https://doi.org/10.1080/02673030220123225.

Tolbert, Charles M., and Molly Sizer. 1996. "US Commuting Zones and Labor Market Areas."

Wang, Qi, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. 2018. "Urban mobility and neighborhood isolation in America's 50 largest cities." *Proceedings of the National Academy of Sciences* 115 (30): 7735 LP – 7740. 10.1073/pnas.1802537115.