Socioeconomic patterns and their dynamics in social networks and language

Márton Karsai

Department of Network and Data Science
CENTRAL EUROPEAN UNIVERSITY

Rényi Institute of Mathematics

Socioeconomic inequalities

Uneven distributions of social and economic positions, which characterise any modern society

Individual level:
- Education
- Occupation
- Mobility
- Social relationships
- Political opinion

Population level:
- Segregation in housing, education, race…
- Limited social mobility
- Poverty and immigration
- Global health issues

Photo by Tuca Vieira

Status homophily

Individuals with similar **social status characteristics** are more likely to associate with each other than by chance *(Lazarsfeld&Merton 1954)*

Socioeconomic status

vs.

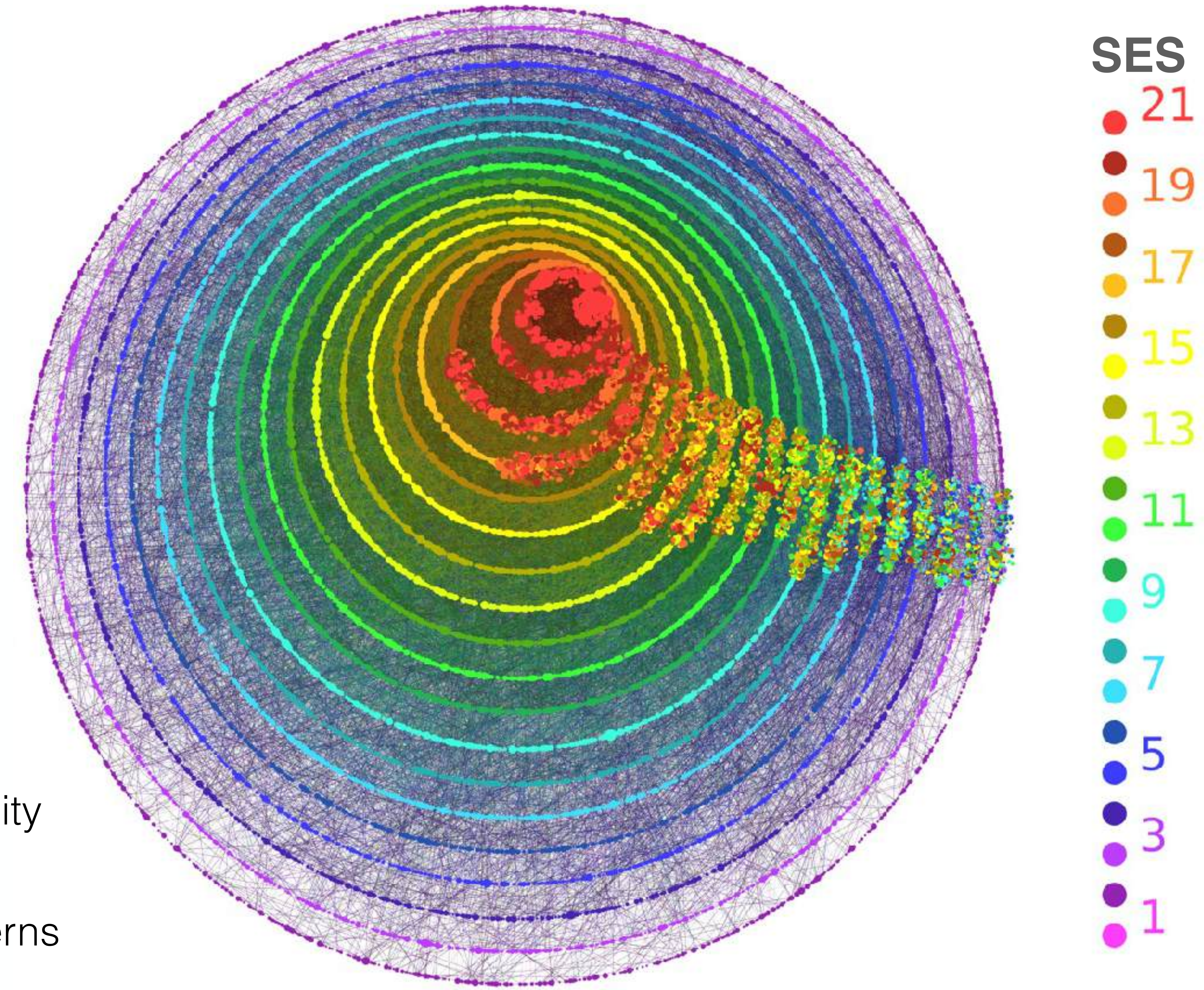
Network formation

How much individual SES and status homophily determines:

- Urban formation?
- Social tie formation?
- Mobility mixing patterns?

Do they lead to global phenomena like

- Segregation patterns in social network and mobility mixing?
- Dynamically changing network segregation patterns due to external shocks?



Socioeconomic core de-composition of a mobil phone call network by Leo, Hamelin and Karsai

Socioeconomic inequalities + Status homophily = Network segregation

To study we need …

DATA

- Individual's socioeconomic status
- Social network structure
- Mobility patterns

… of millions of individuals

Socioeconomic status inference

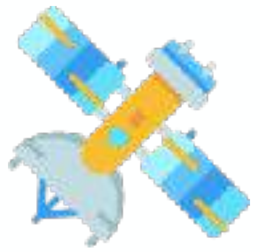
Segregation patterns of socioeconomic and mobility networks and their dynamics during external shocks

Socioeconomic patterns in standard language use

Socioeconomic Status Inference

“The heavy way”

SES inference from multimodal data



Daylight
satellite image

Google Maps

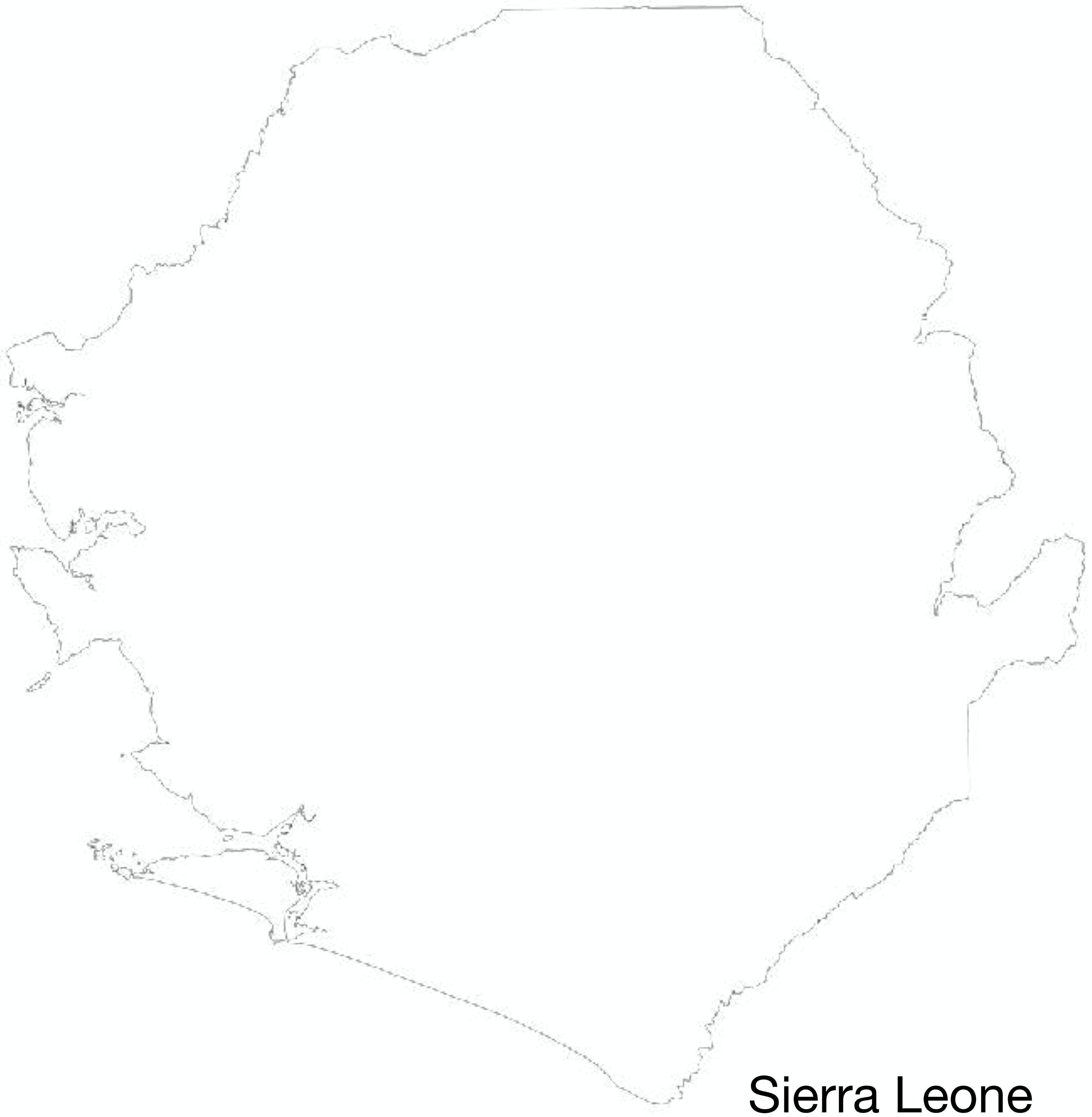
$200 free monthly usage

SES inference from multimodal data

Data	Source	Cost
Daylight satellite image	Google Maps	$200 free monthly usage
Nightlight intensity	Google Earth Engine	Free
Connectivity	unwiredlabs OpenCelliD	Free
Infrastructure	OpenStreetMap	Free
Target audience	facebook Ads	Free
Population	FACEBOOK Data for Good High Resolution Population Density Maps	Free (upon request)
Mobility network	FACEBOOK Data for Good Movement Between Tiles (before corona crisis)	Free (upon request)
GROUND TRUTH DATA: Survey data	The DHS Program Demographic and Health Surveys	Free (upon request)

Sierra Leone - Inferring poverty maps with multimodal data

A supervised learning approach

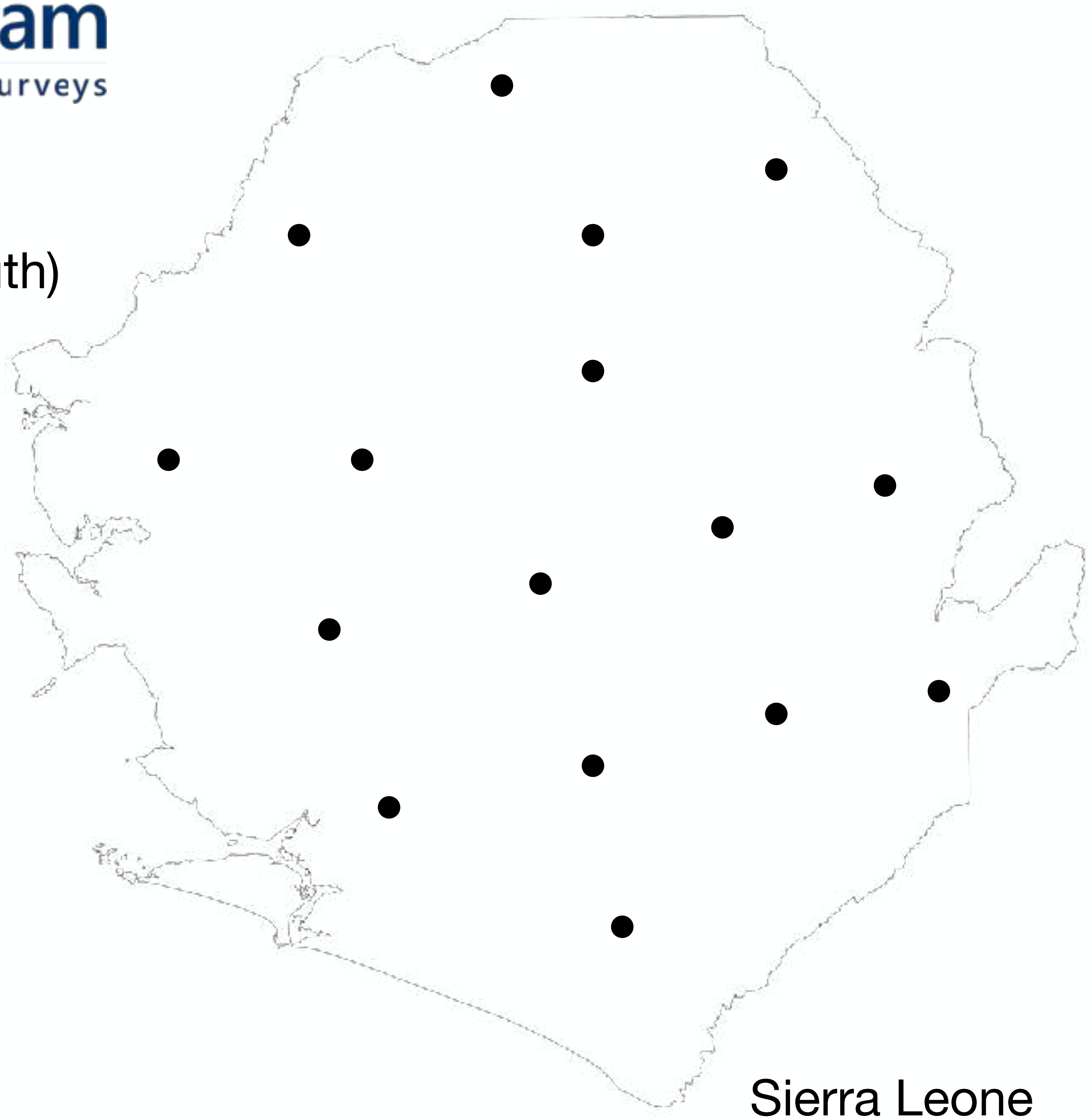


Sierra Leone

Sierra Leone - Inferring poverty maps with multimodal data

A supervised learning approach

1. Build a model using the DHS clusters (survey data)



Sierra Leone - Inferring poverty maps with multimodal data

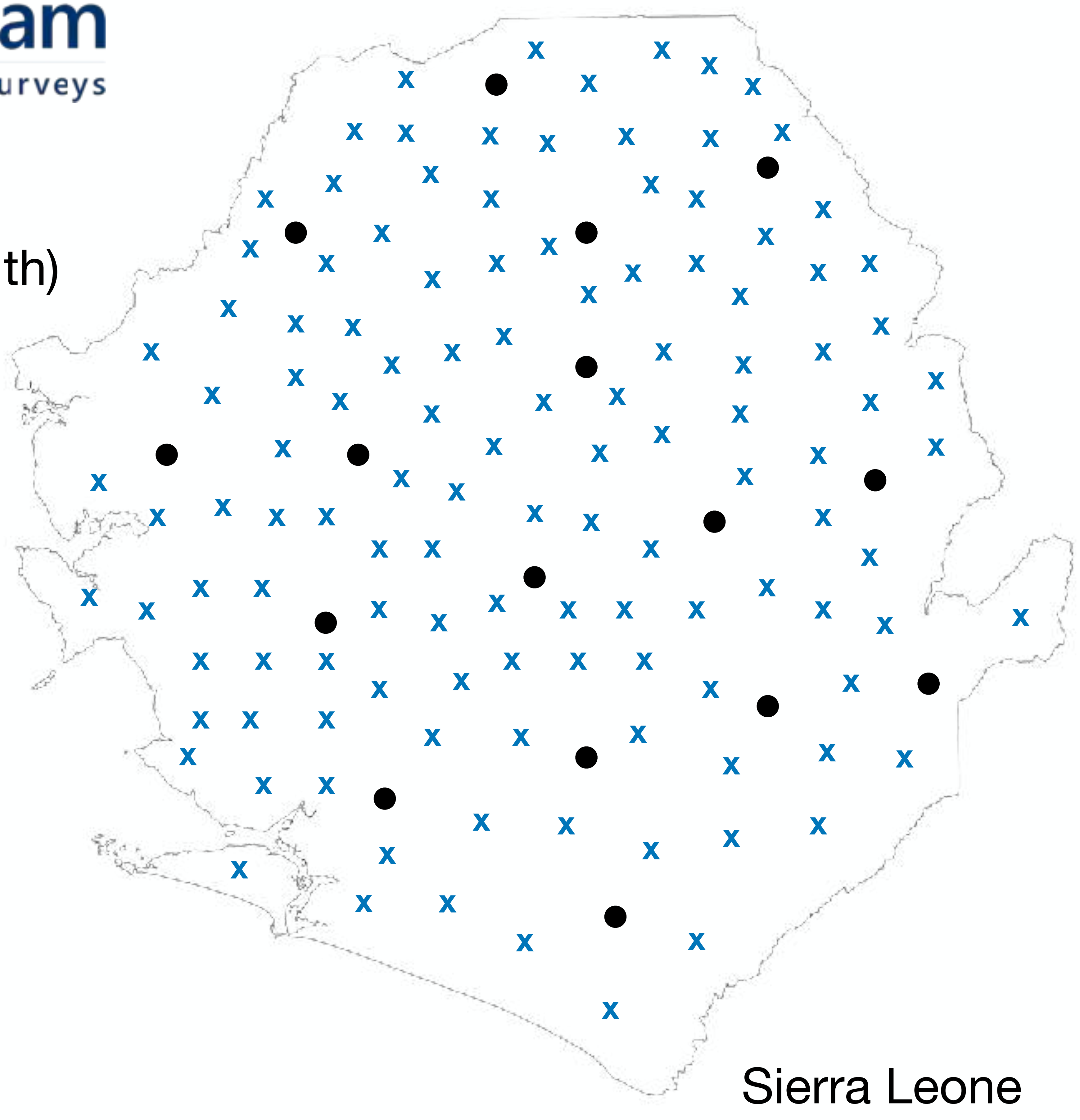
A supervised learning approach

1. Build a model using the DHS clusters (survey data)

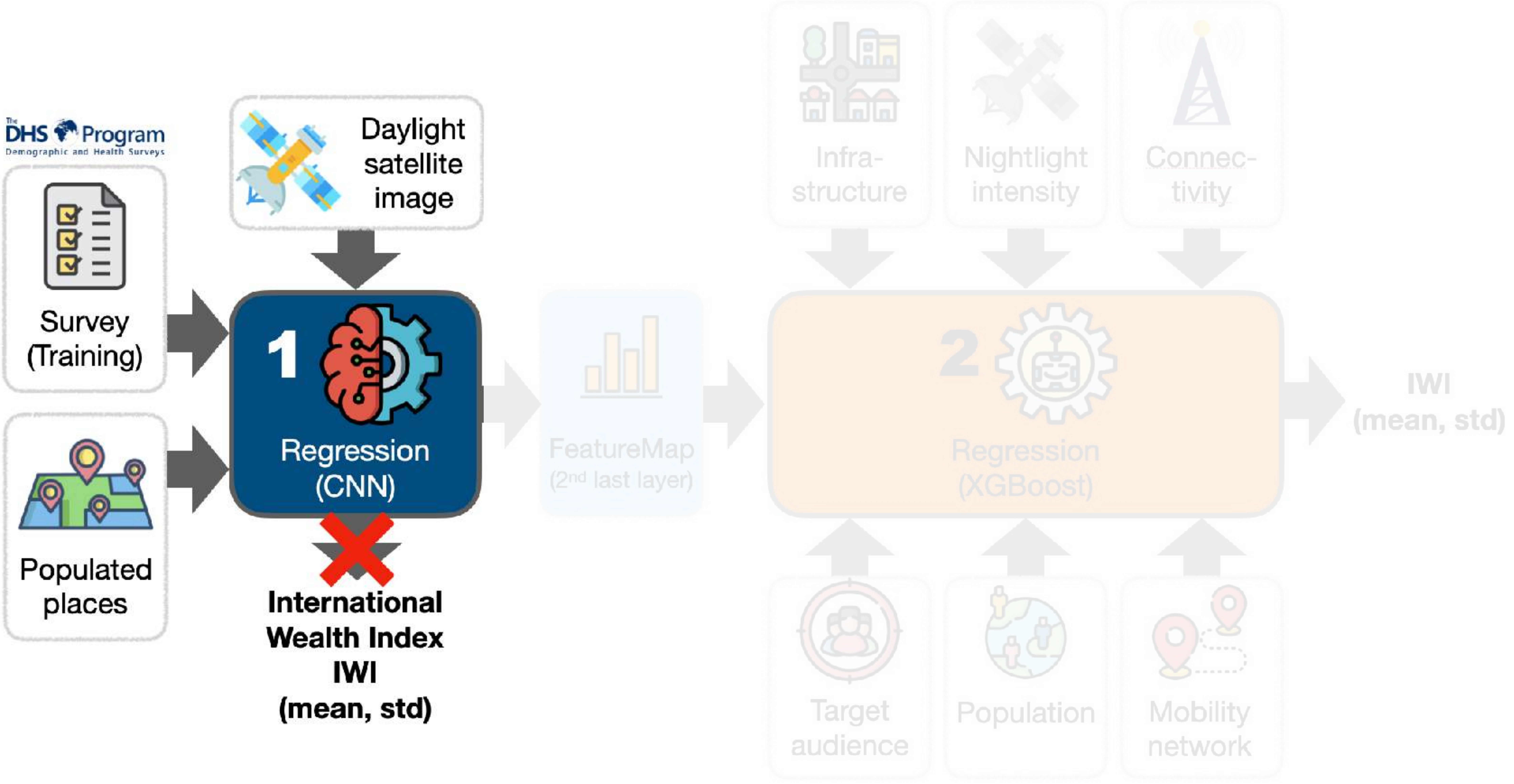
2. Apply that model on the populated places

● DHS clusters (as ground truth)

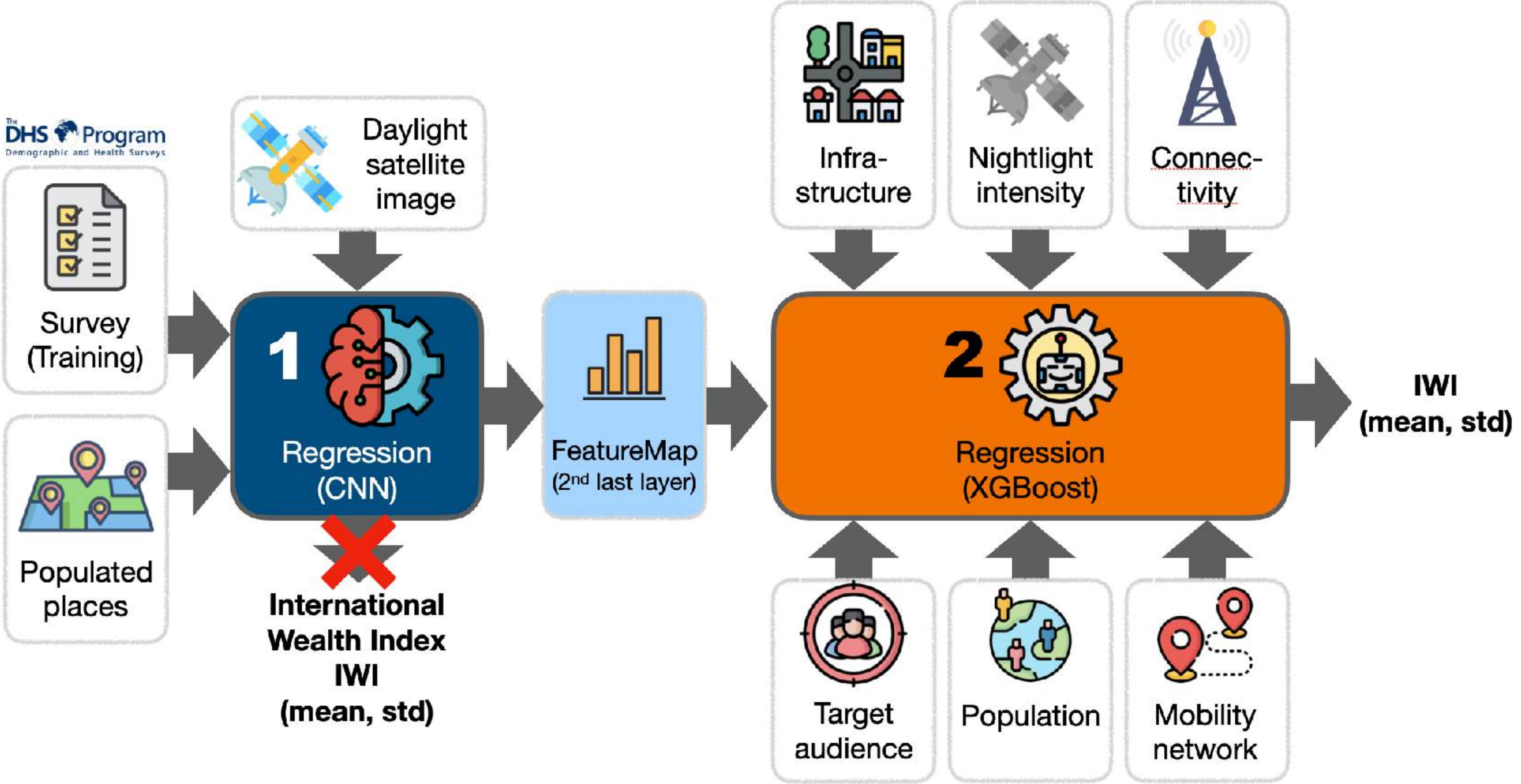
x Populated places



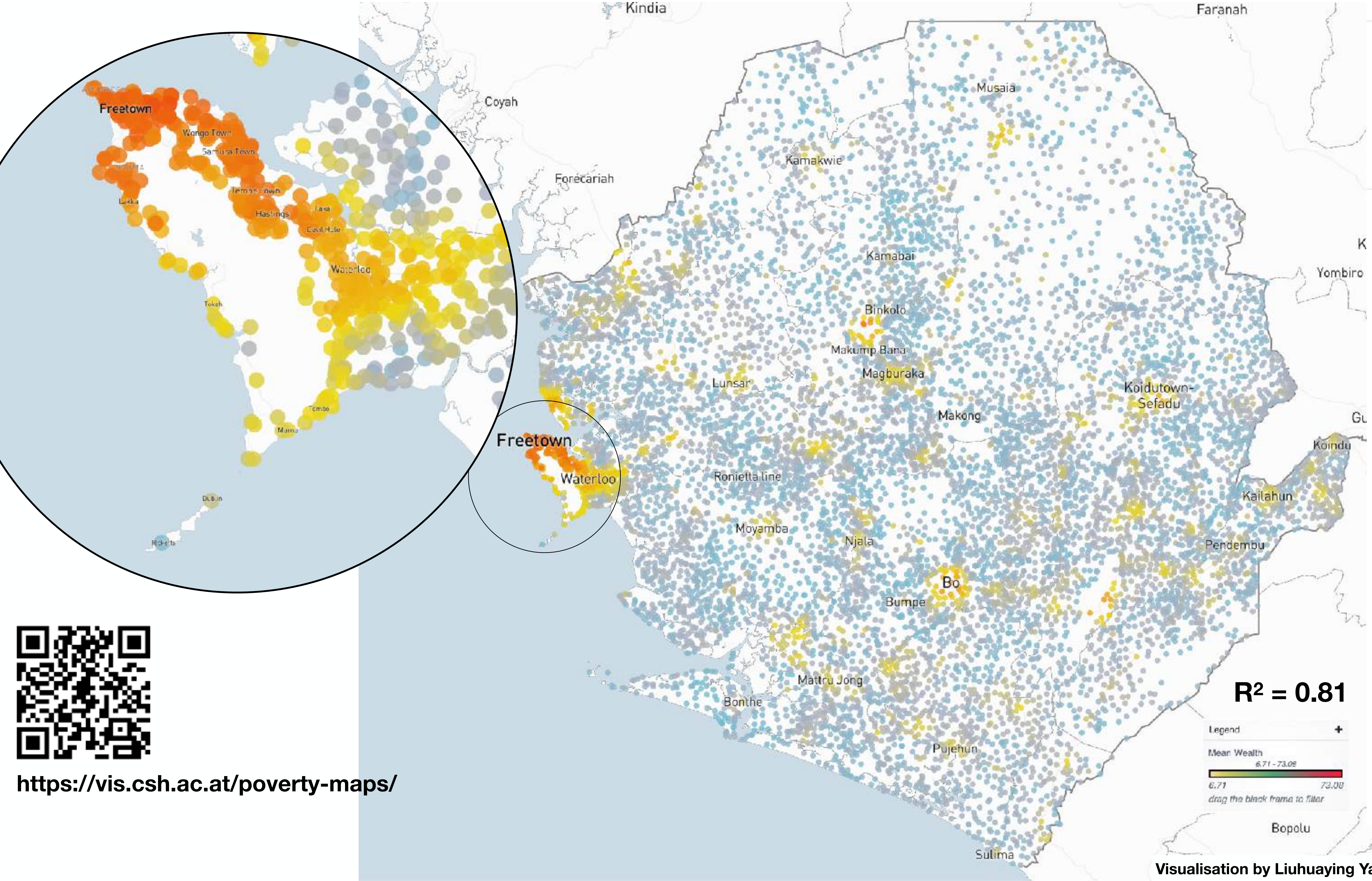
Sierra Leone - Socioeconomic data



Sierra Leone - Socioeconomic data



Sierra Leone - Socioeconomic map



Visualisation by Liuhuaying Yang

https://vis.csh.ac.at/poverty-maps/

Town

Wongo Town

Wealth Distribution mean 53.61 std 14.83

Mobility	
Number of visitors (average per day)	5431
Number of outgoers (average per day)	5367
PageRank of place	0.00488
Population	
Population density (in 1.61km)	31505.25
Demographics	
Number of tablet owners	500
Number of people with a Master's degree	1000
Number of people with a high school degree	500
Infrastructure	
Number of buildings	1870
Distance to closest road	0 m
Type of settlement	Urban
Nightlights	
Minimum luminosity (in 1.61km)	1.24
Mean luminosity (in 1.61km)	3.03
Maximum luminosity (in 1.61km)	10.73
Connectivity	
Distance to closest antenna	1047.98 m
Number of antennas (in 1.61km)	0

RESEARCH-ARTICLE OPEN ACCESS

Interpreting wealth distribution via poverty map inference using multimodal data

Authors: Lisette Espín-Noboa, János Kertész, Márton Karsai Authors Info & Claims

WWW '23: Proceedings of the ACM Web Conference 2023 • April 2023 • Pages 4029–4040

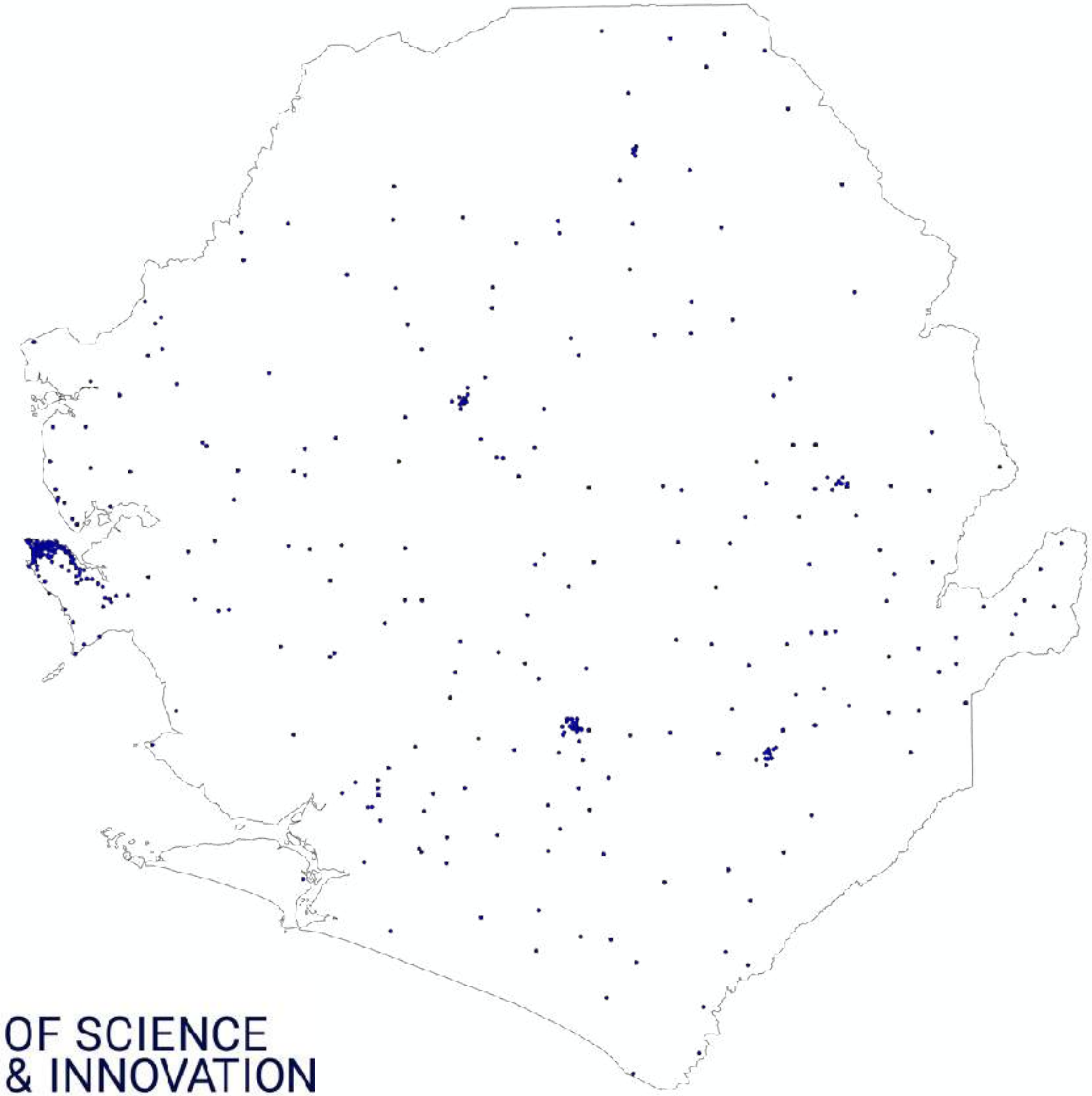
• https://doi.org/10.1145/3543507.3583862

Socioeconomic segregation patterns in social and mobility networks

Sierra Leone - Social network and mobility

Geo-localised mobile communication data

- Company with the 2nd largest market share
- 1.3M anonymised mobile-phone users
- CDR about calls and sms
- Location of company users at the time of communications
- Inferred home location from spatio-temporal trajectories of active users



unicef INNOVATION

DIRECTORATE OF SCIENCE TECHNOLOGY & INNOVATION Sierra Leone

Sierra Leone - Data combination

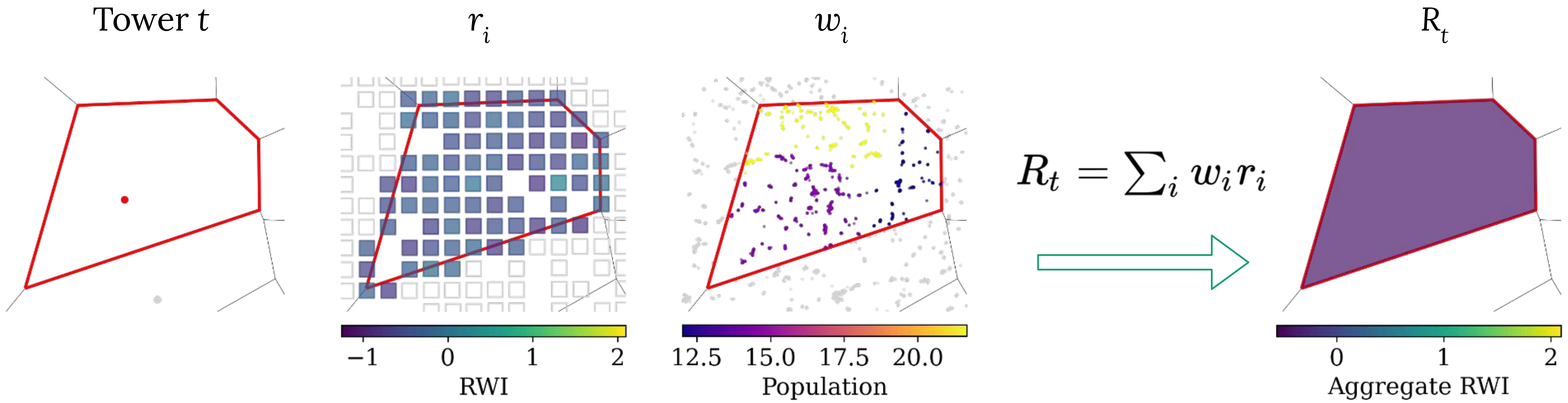
Geo-localised mobile communication data

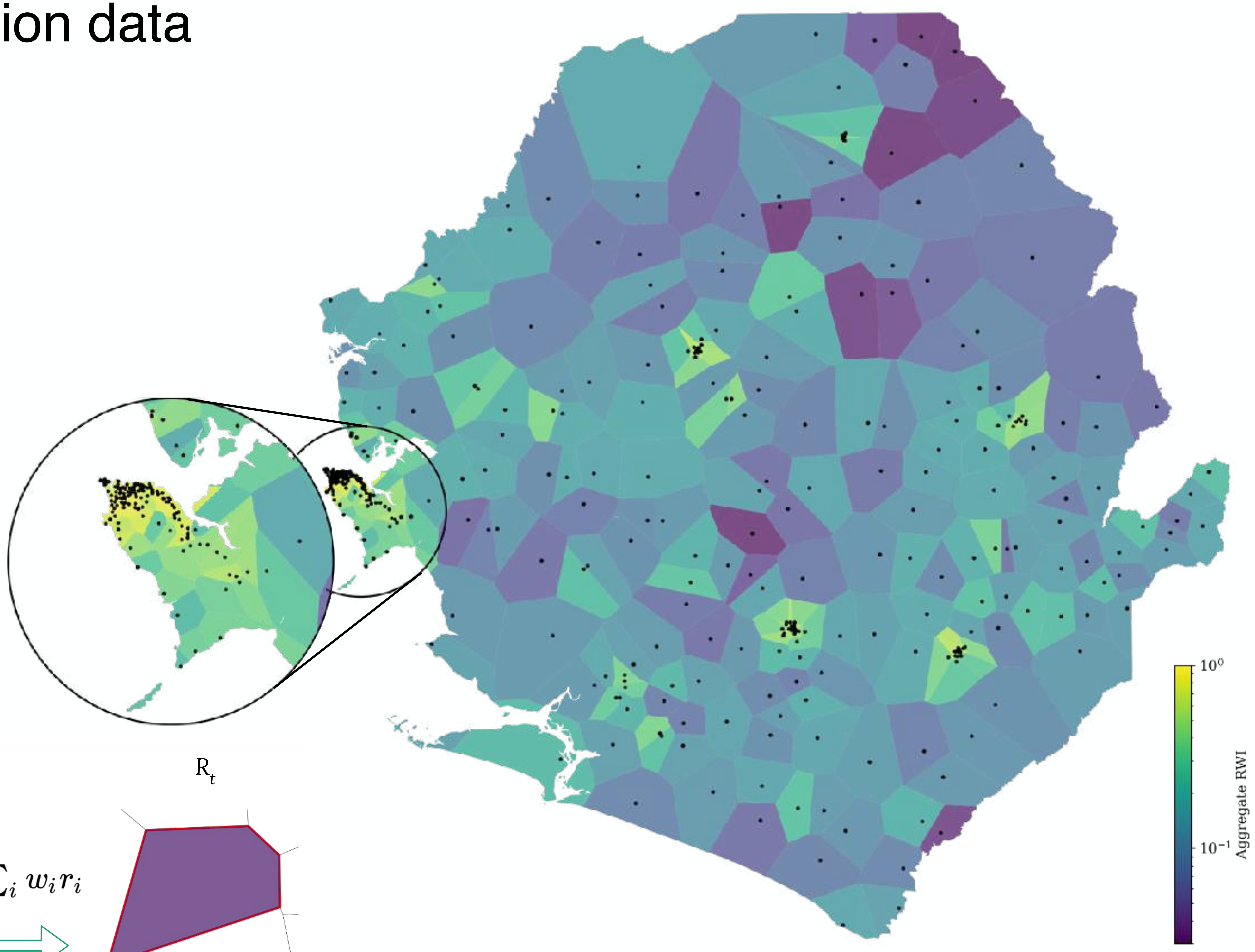
- Company with the 2nd largest market share
- 1.3M anonymised mobile-phone users
- CDR about calls and sms
- Location of company users at the time of communications
- Inferred home location from spatio-temporal trajectories of active users

Socioeconomic map

- Population map (Facebook D4G)
- **Relative Wealth Index** (RWI) poverty map (Espín, Kertész and Karsai, to be published (2022); Chi et al. PNAS (2022))

Combined SES map

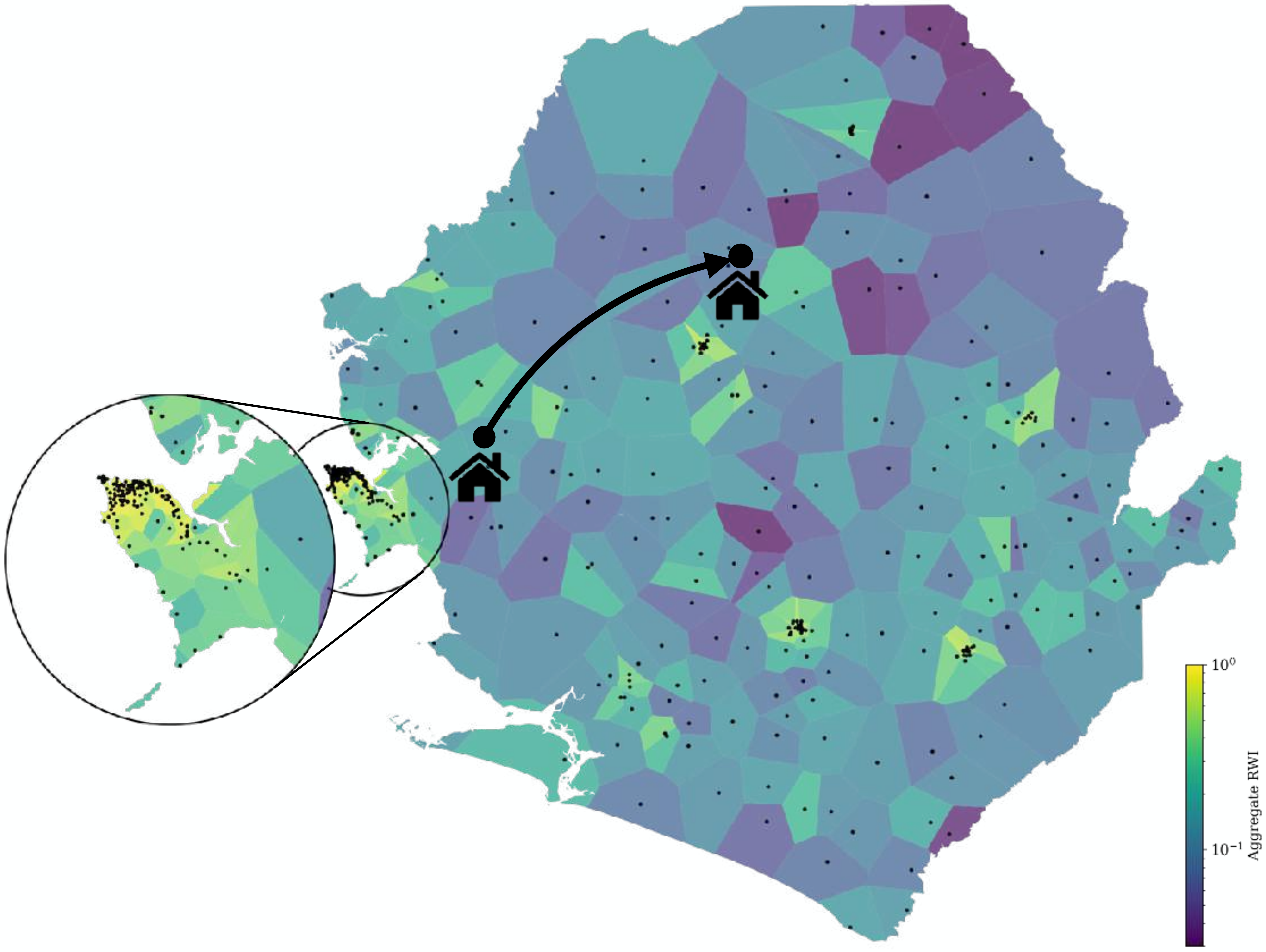




Sierra Leone - Socioeconomic networks

Social communication network

- **Nodes**: people
- **SES indicator**: RWI at cell tower of home location
- **Links**: calls/sms between people
- **Link weights**: number of communications



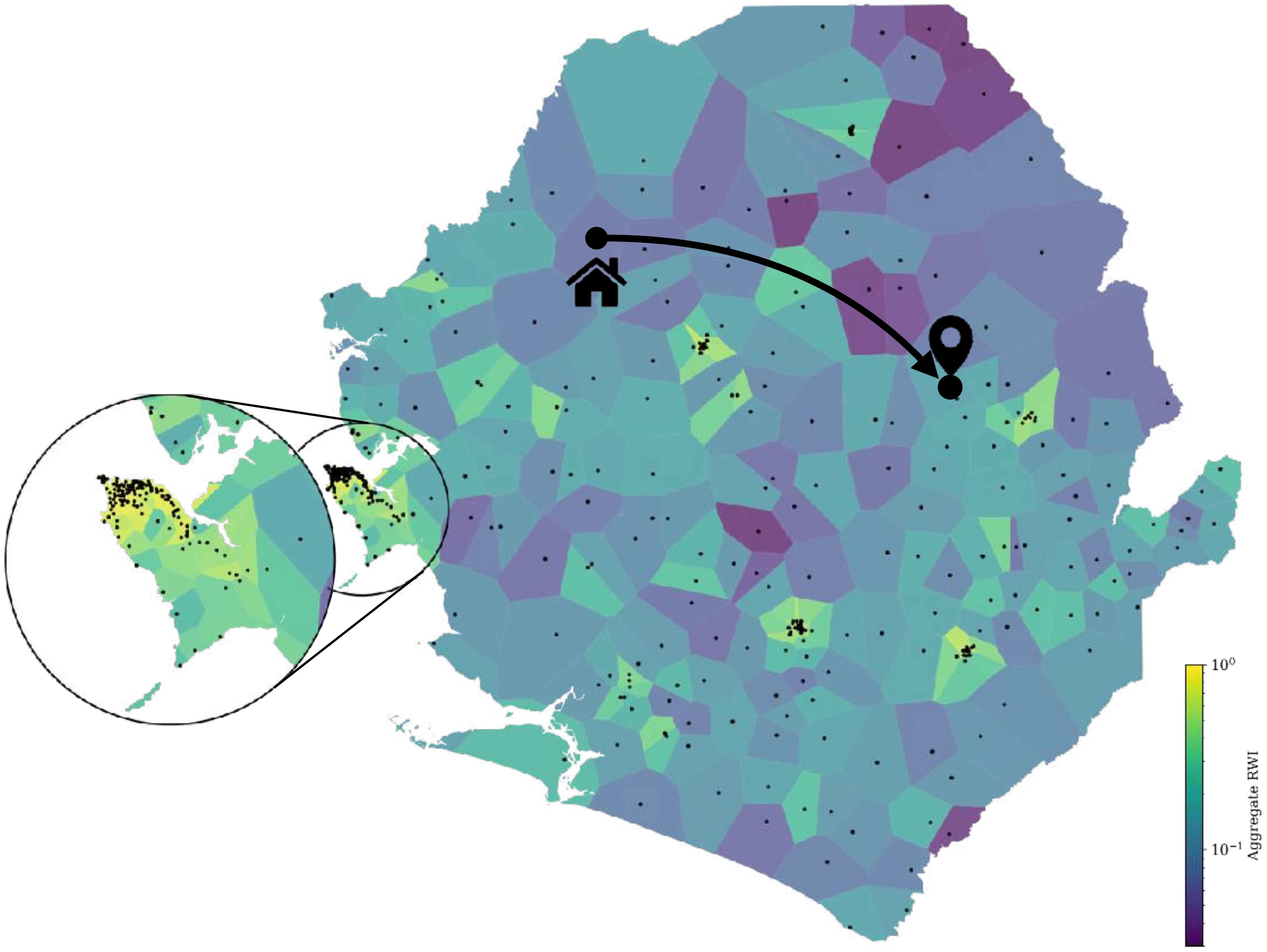
Sierra Leone - Socioeconomic networks

Social communication network

- **Nodes**: people
- **SES indicator**: RWI at cell tower of home location
- **Links**: calls/sms between people
- **Link weights**: number of communications

Mobility network

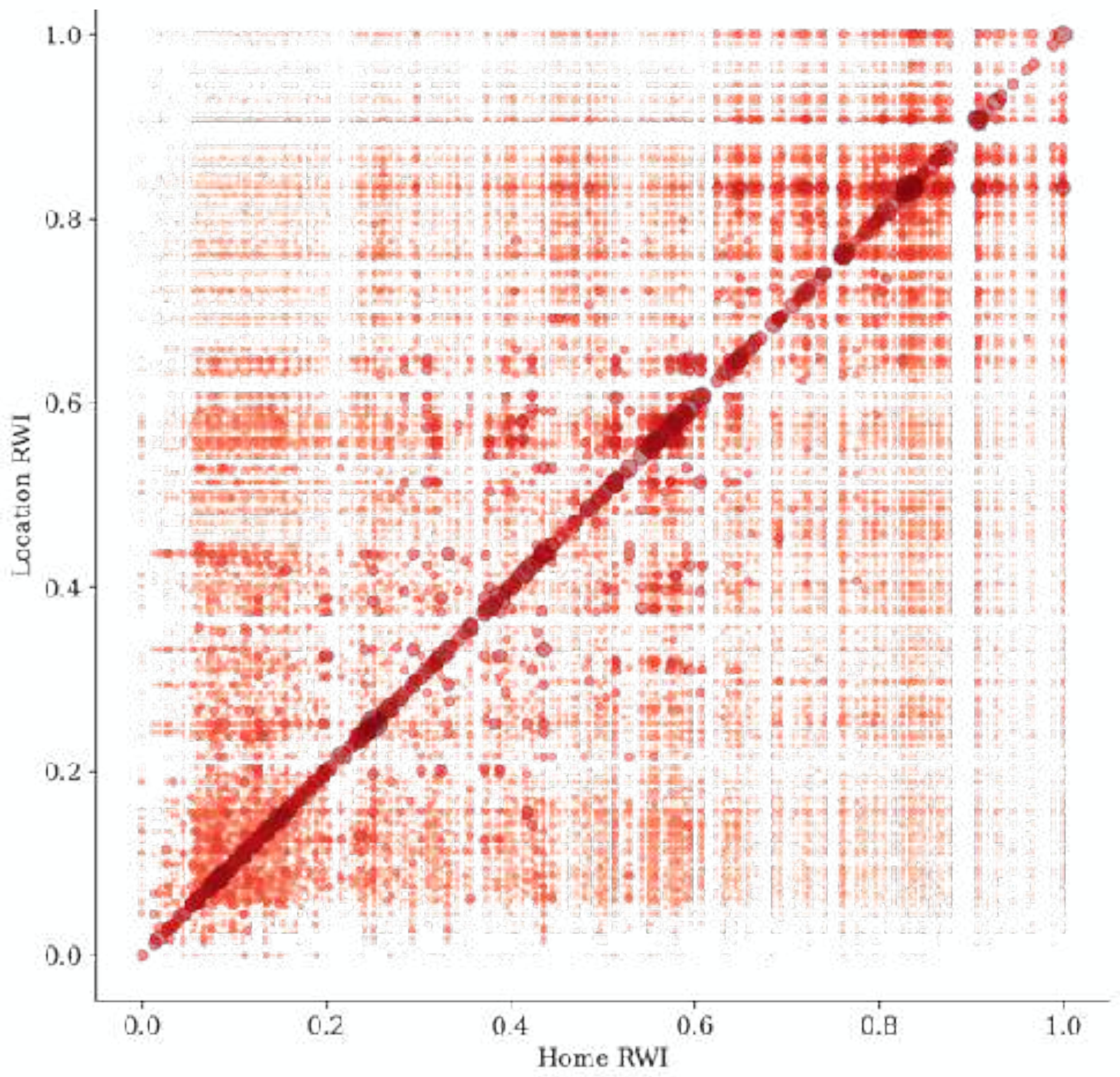
- **Nodes**: home locations and visited places
- **SES indicator**: RWI at cell towers at home or visited locations
- **Links**: connections between home and visited locations
- **Link weights**: number of visits

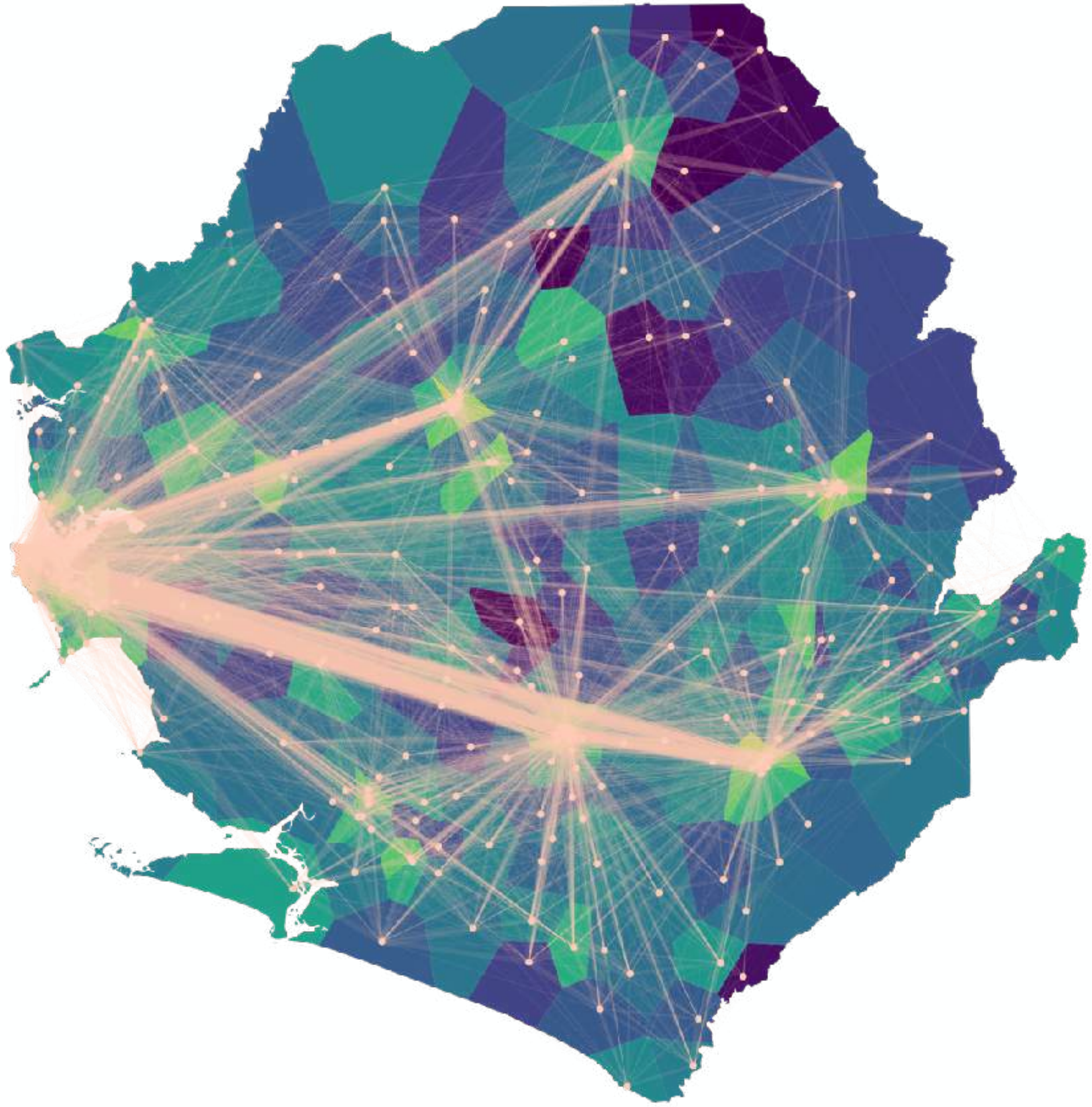


Sierra Leone - Socioeconomic network segregations

Network assortativity matrix

Mobility network

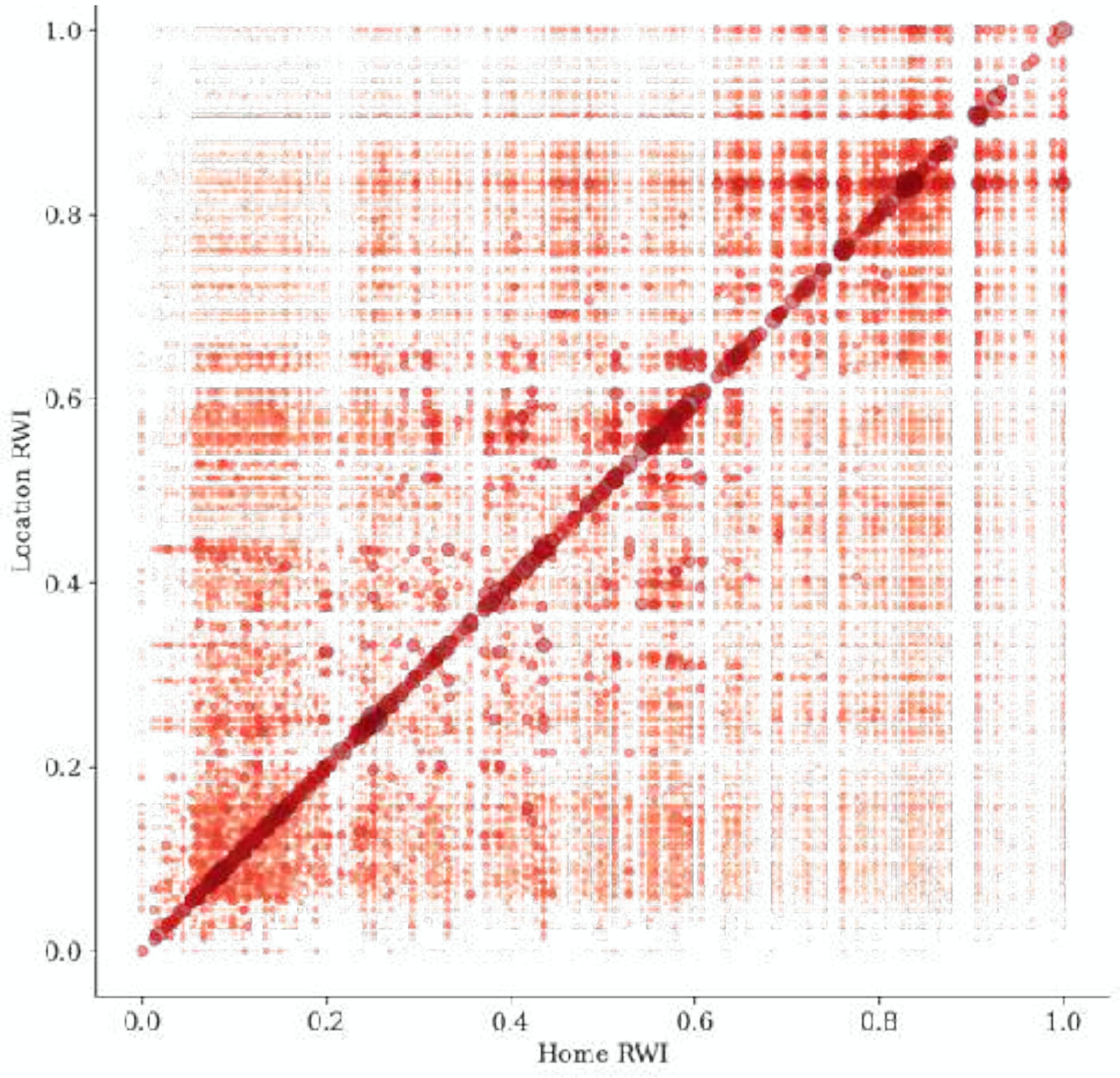




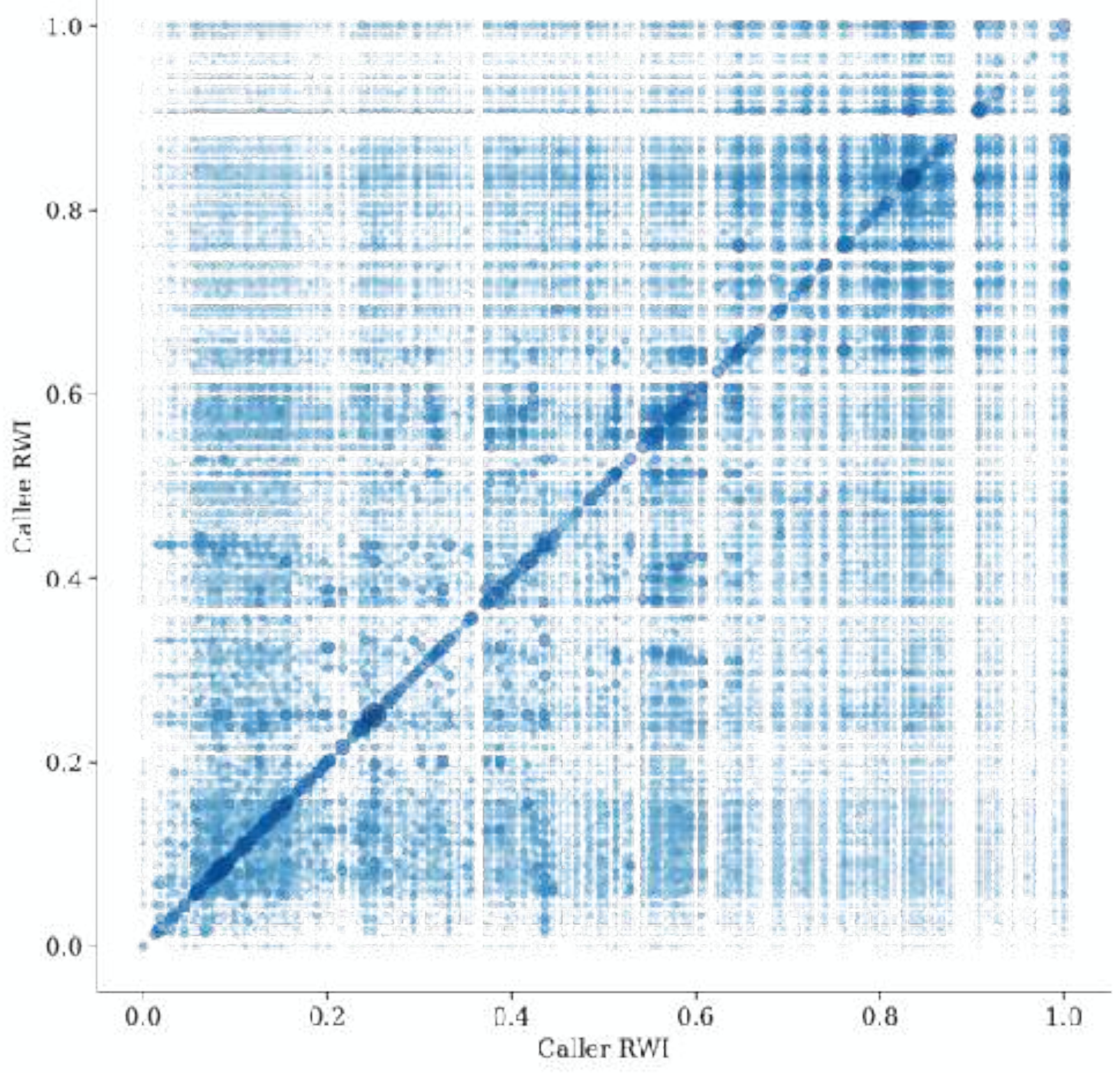
Sierra Leone - Socioeconomic network segregations

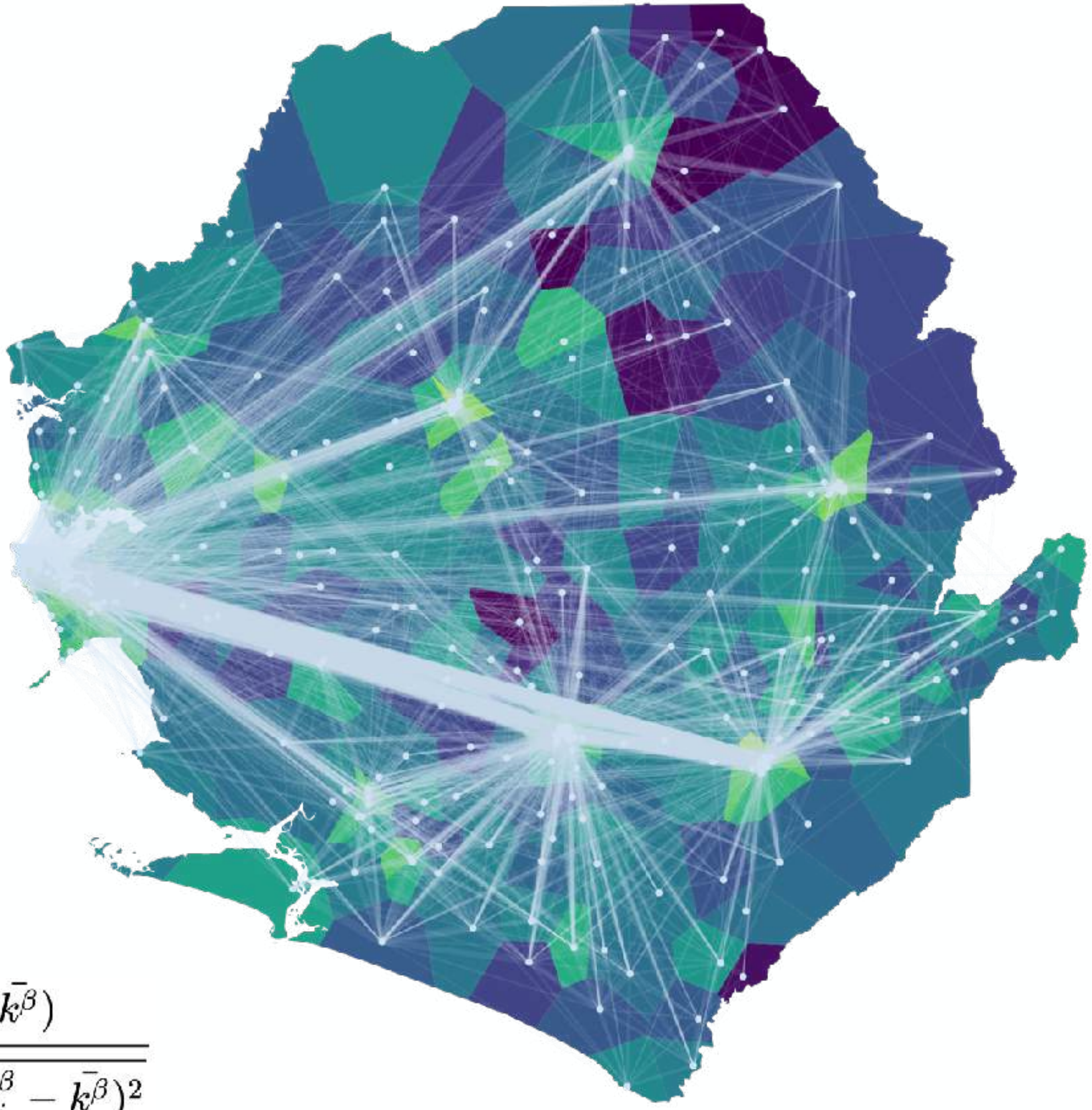
Network assortativity matrix

Mobility network **r = 0.66**



Social network **r = 0.51**



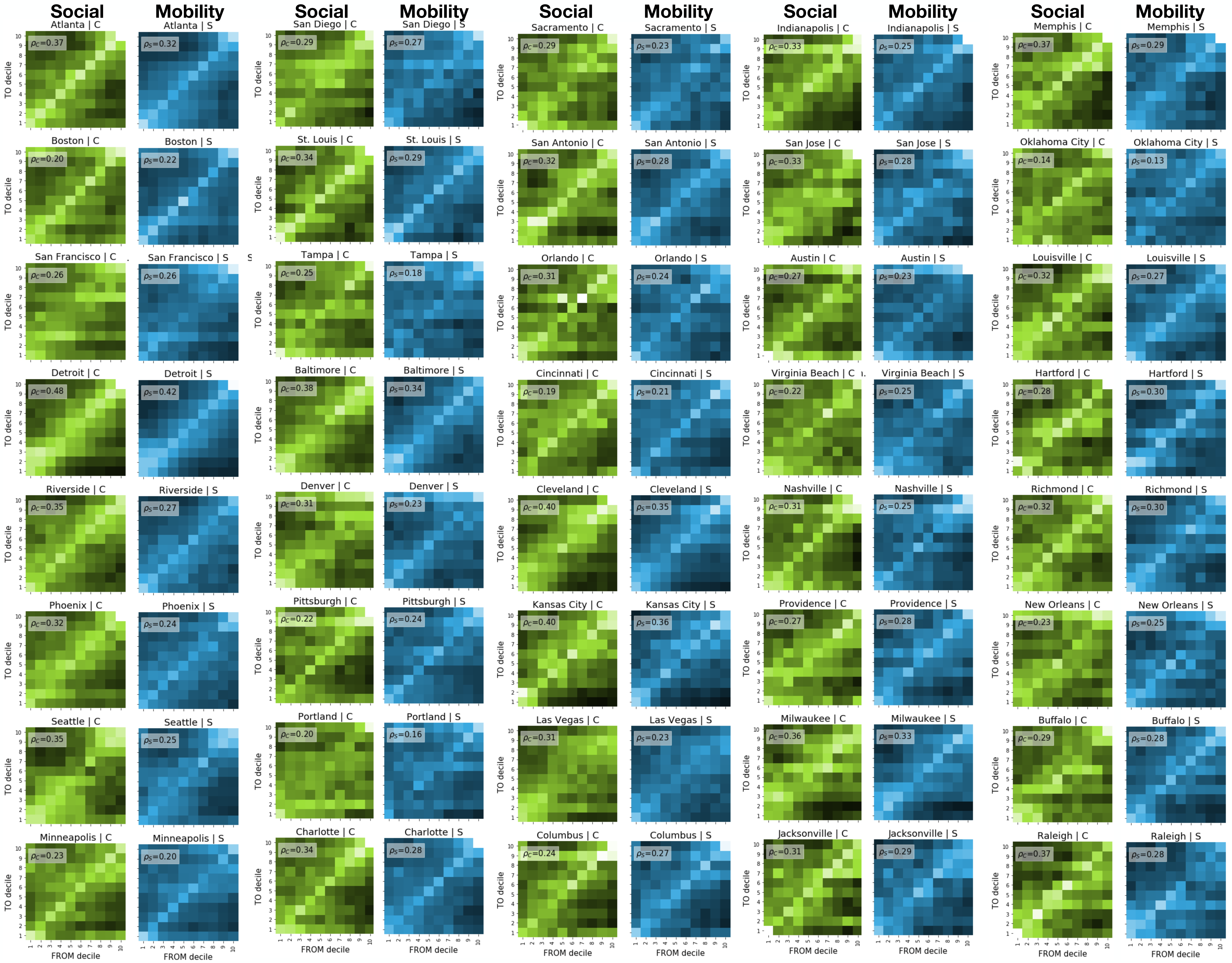


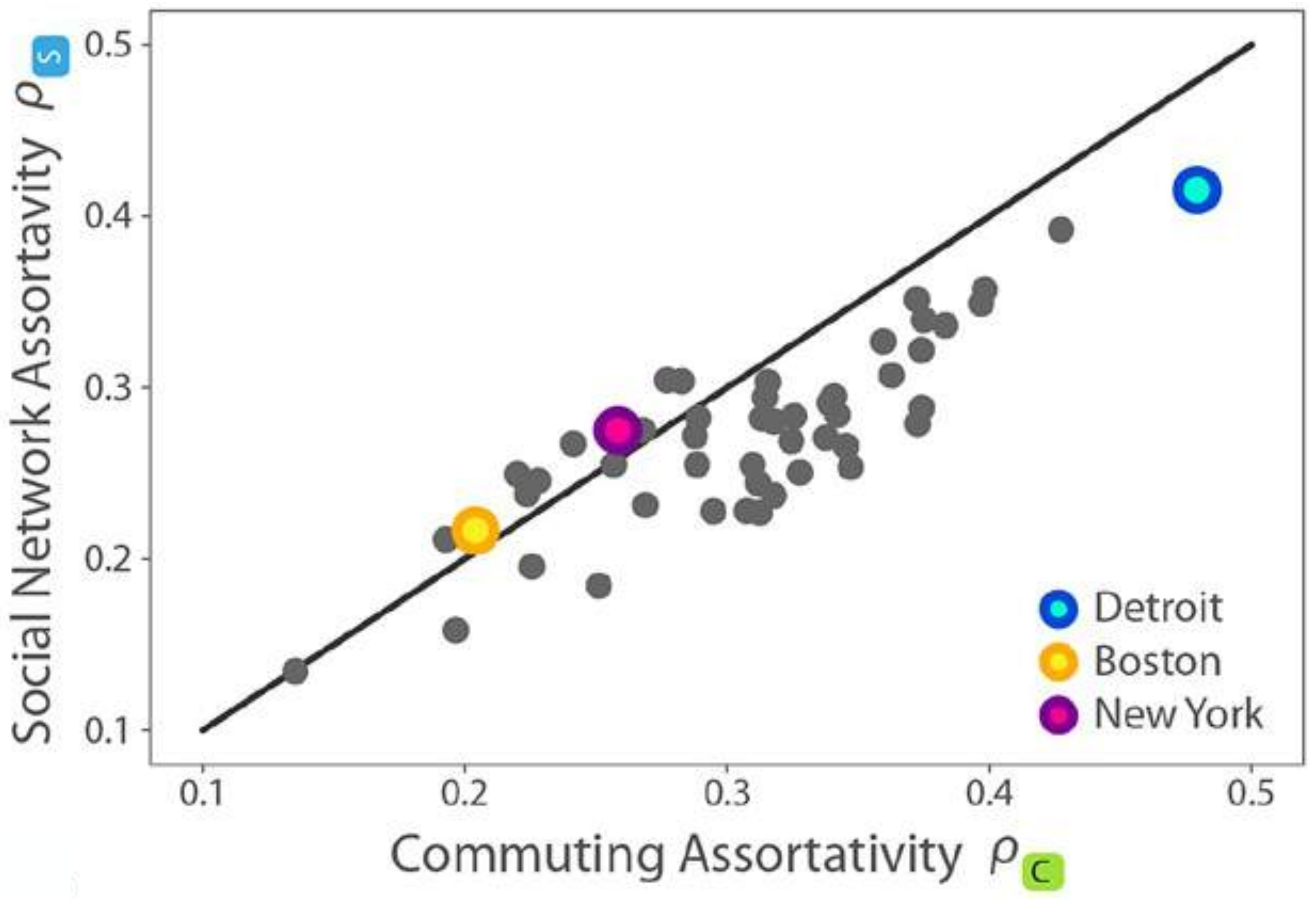
Global network segregation index:

Newman's assortativity coefficient

$$r(\alpha,\beta) = \frac{\sum_i (j_i^\alpha - \bar{j^\alpha})(k_i^\beta - \bar{k^\beta})}{\sqrt{\sum_i (j_i^\alpha - \bar{j^\alpha})^2}\sqrt{\sum_i (k_i^\beta - \bar{k^\beta})^2}}$$

US - Socioeconomic network segregations





Strong correlations between social and mobility assortativity in the 50 largest US cities

Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, M. Karsai
J. R. Soc. Interface **13** 125 (2016)

E. Bokányi, S. Juhász, M. Karsai, B. Lengyel
Scientific Reports 11, 20829 (2021)

R. Millanida Hilman, G. Iñiguez, M. Karsai
EPJ Data Science 11, 32 (2022)

Socioeconomic network re-organization due to external shocks

Segregation patterns evolve slowly

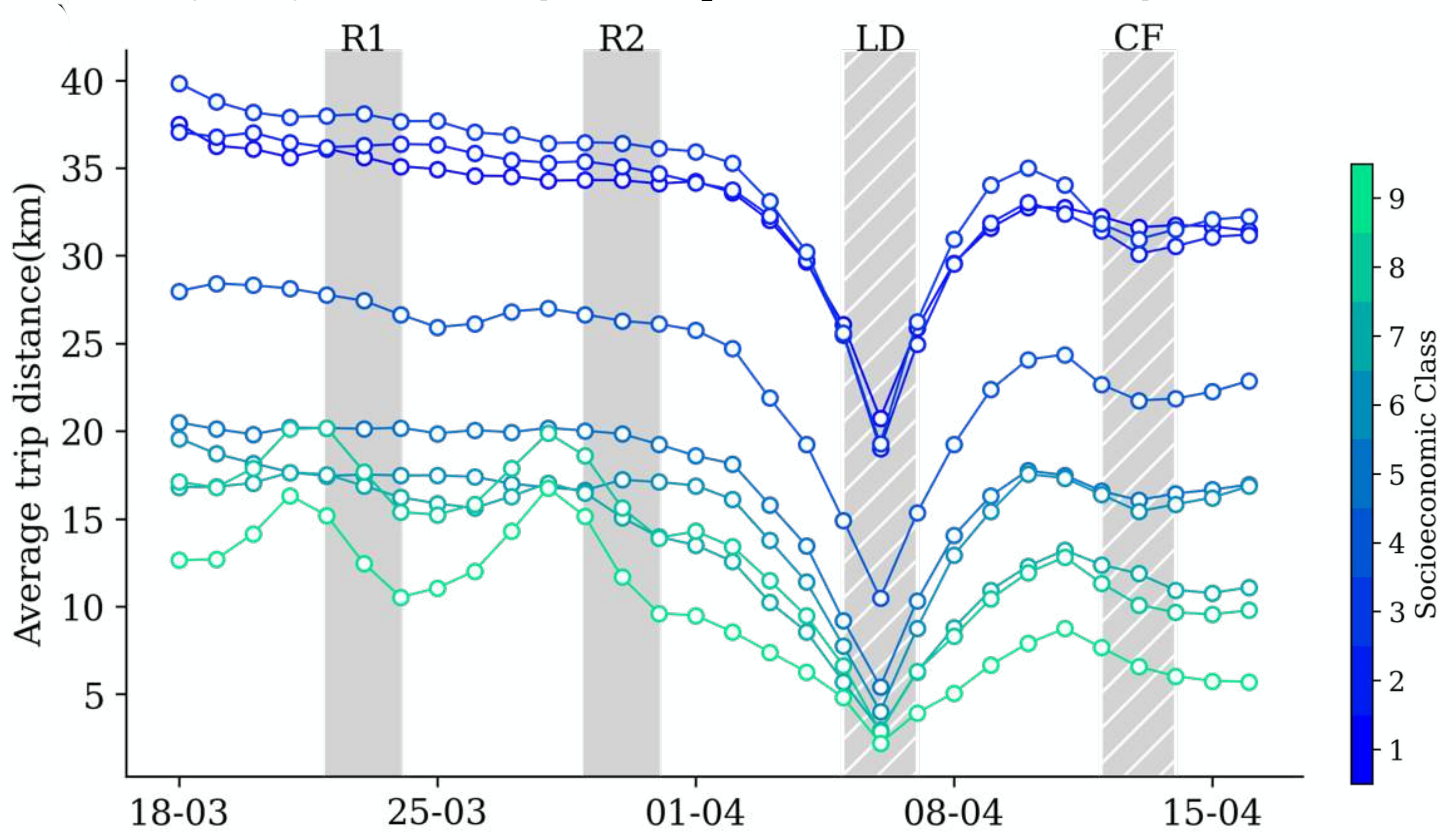
- Their change require the application of new policies
- New education strategies to induce social mobility
- Urban design projects to reduce residential segregation
- Development of transportation to help mobility mixing

How does segregation evolves due to external shocks like the COVID-19 pandemic?

- Do socioeconomic networks re-organise due to emergency interventions?

POLITICO

Mobility adjustment (Average travel distance)

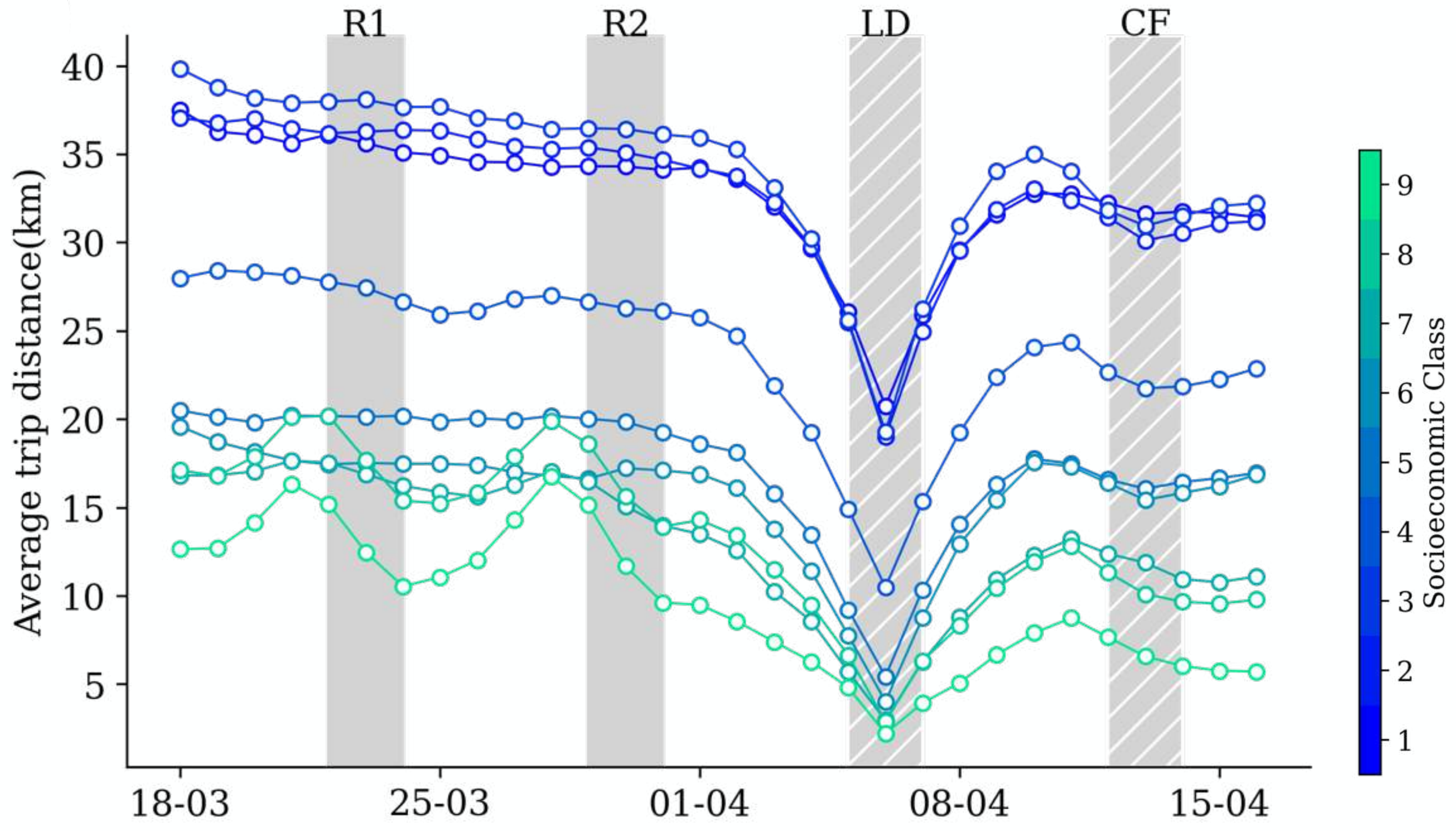


Sierra Leone

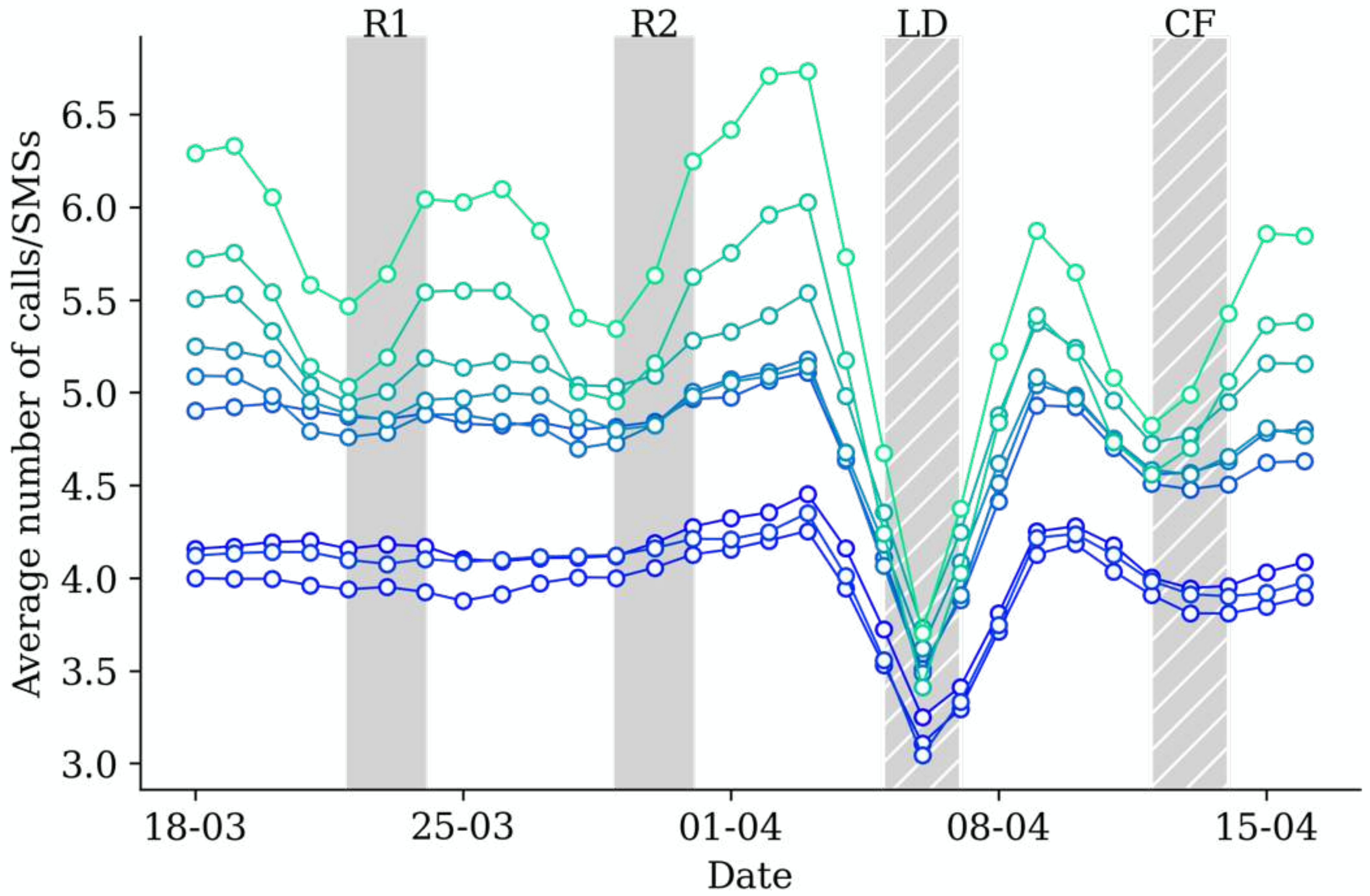
Mobility and social communication activities

- 4 weeks observations: 17/3-17/4 2020
- **R1**: 1st reference period (2 weeks before lockdown)
- **R2**: 2nd reference period (1 week before lockdown)
- **LD**: Full national lockdown period
- **CF**: National overnight curfew period (1 week after lockdown)

Mobility adjustment (Average travel distance)



Social-communication adjustment (Number of comm.)

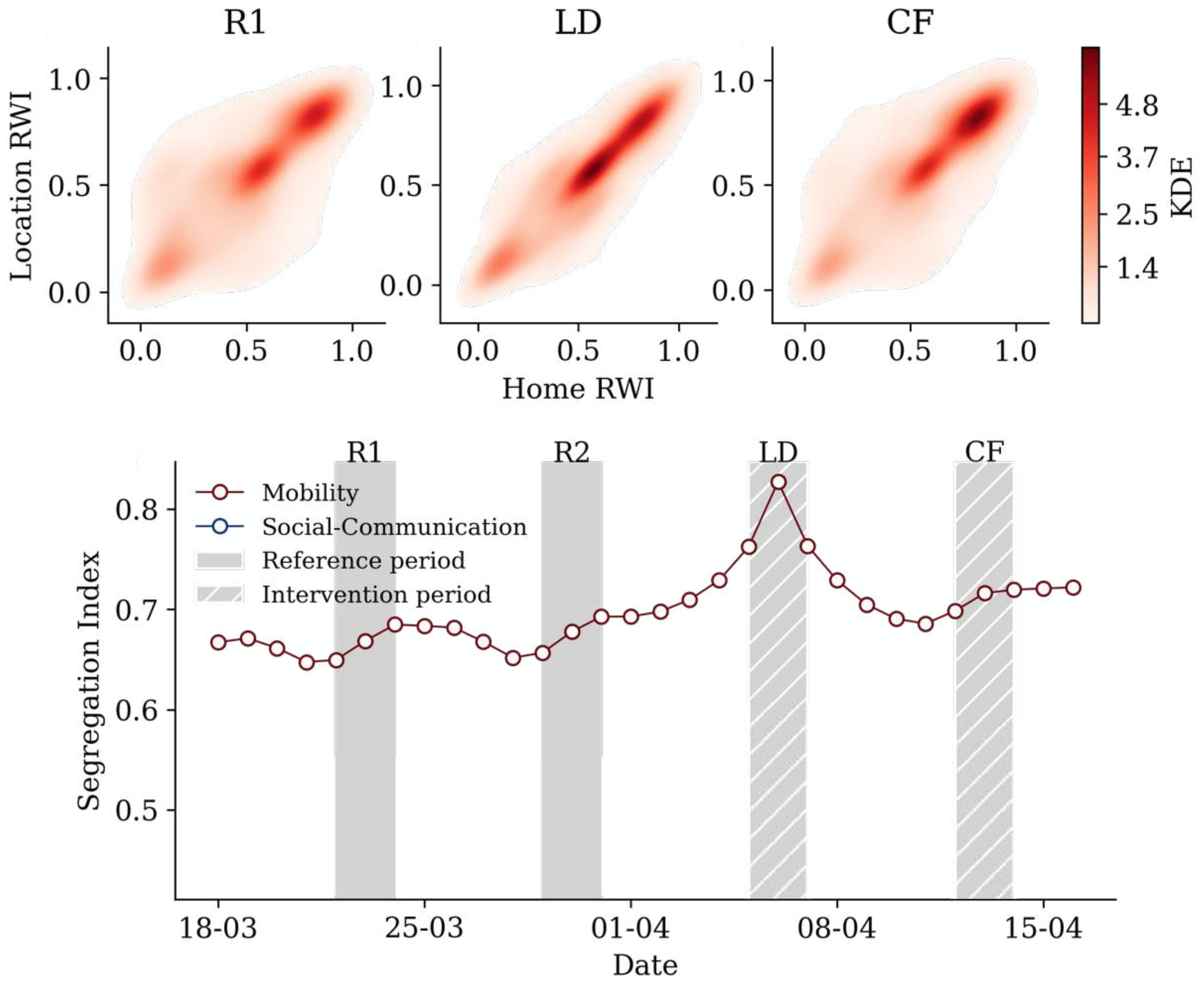


Sierra Leone

Mobility and social communication activities

- 4 weeks observations: 17/3-17/4 2022
- **R1**: 1st reference period (2 weeks before lockdown)
- **R2**: 2nd reference period (1 week before lockdown)
- **LD**: Full national lockdown period
- **CF**: National overnight curfew period (1 week after lockdown)

Different adjustment capacities appears for people with different socioeconomic background.



Sierra Leone

Dynamics of network segregation

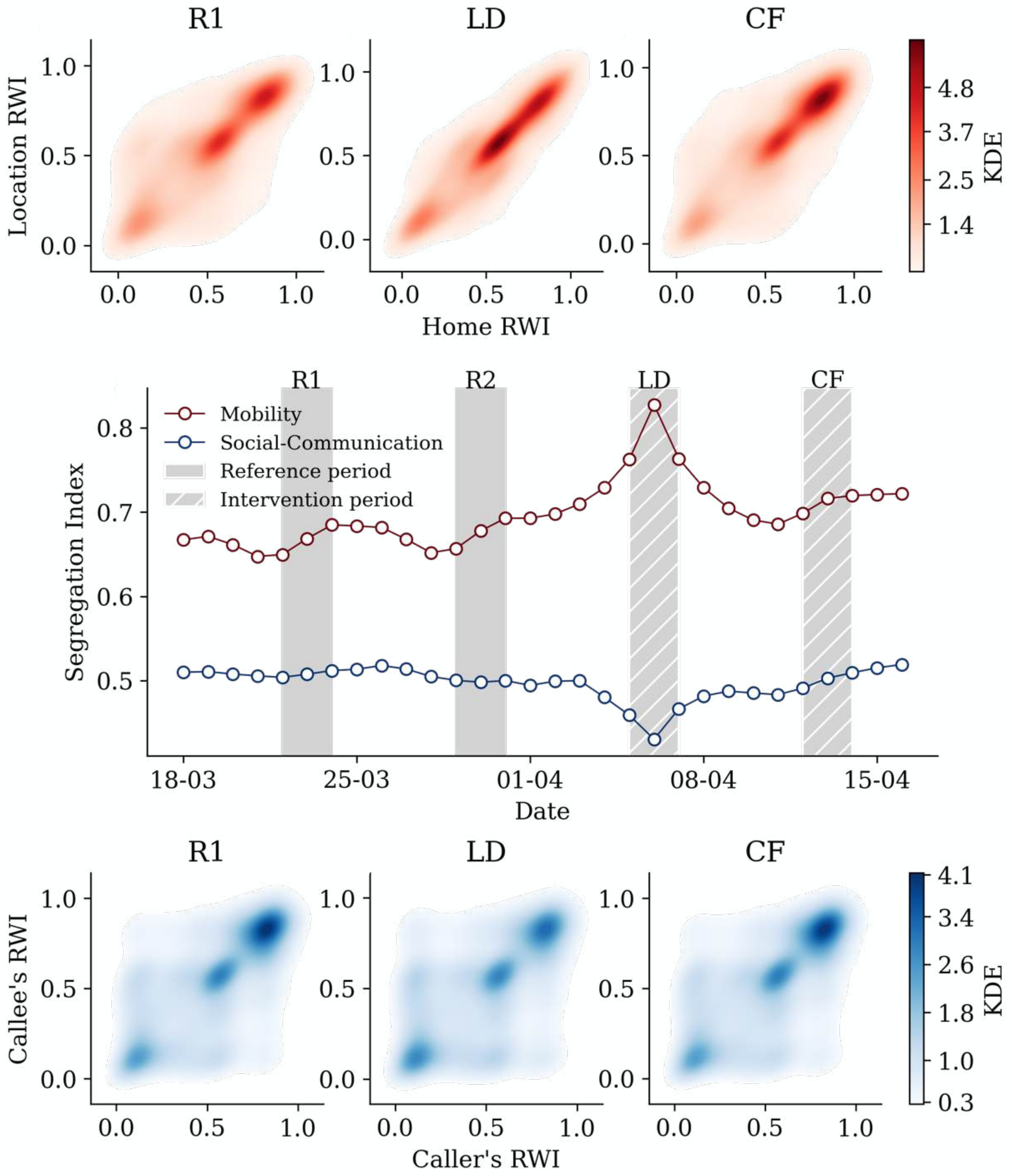
Mobility segregation increased during lockdown

Bonaccorsi, G., et al., PNAS 117.27 15530-15535 (2020)

Yabe, T., Bueno, B. G. B., Dong, X., Pentland, A., & Moro, E. *arXiv:2207.06895 (2022)*

Hilman, R., Sekara, V., Herranz, M. & Karsai, M., to be published (2023)

Napoli, L., Sekara, V., Herranz, M. & Karsai, M., to be published (2023)



Sierra Leone

Dynamics of network segregation

Mobility segregation increased during lockdown

Bonaccorsi, G., et al., PNAS 117.27 15530-15535 (2020)

Yabe, T., Bueno, B. G. B., Dong, X., Pentland, A., & Moro, E. *arXiv:2207.06895 (2022)*

Hilman, R., Sekara, V., Herranz, M. & Karsai, M., to be published (2023)

Napoli, L., Sekara, V., Herranz, M. & Karsai, M., to be published (2023)

Social network segregation decreased during lockdown

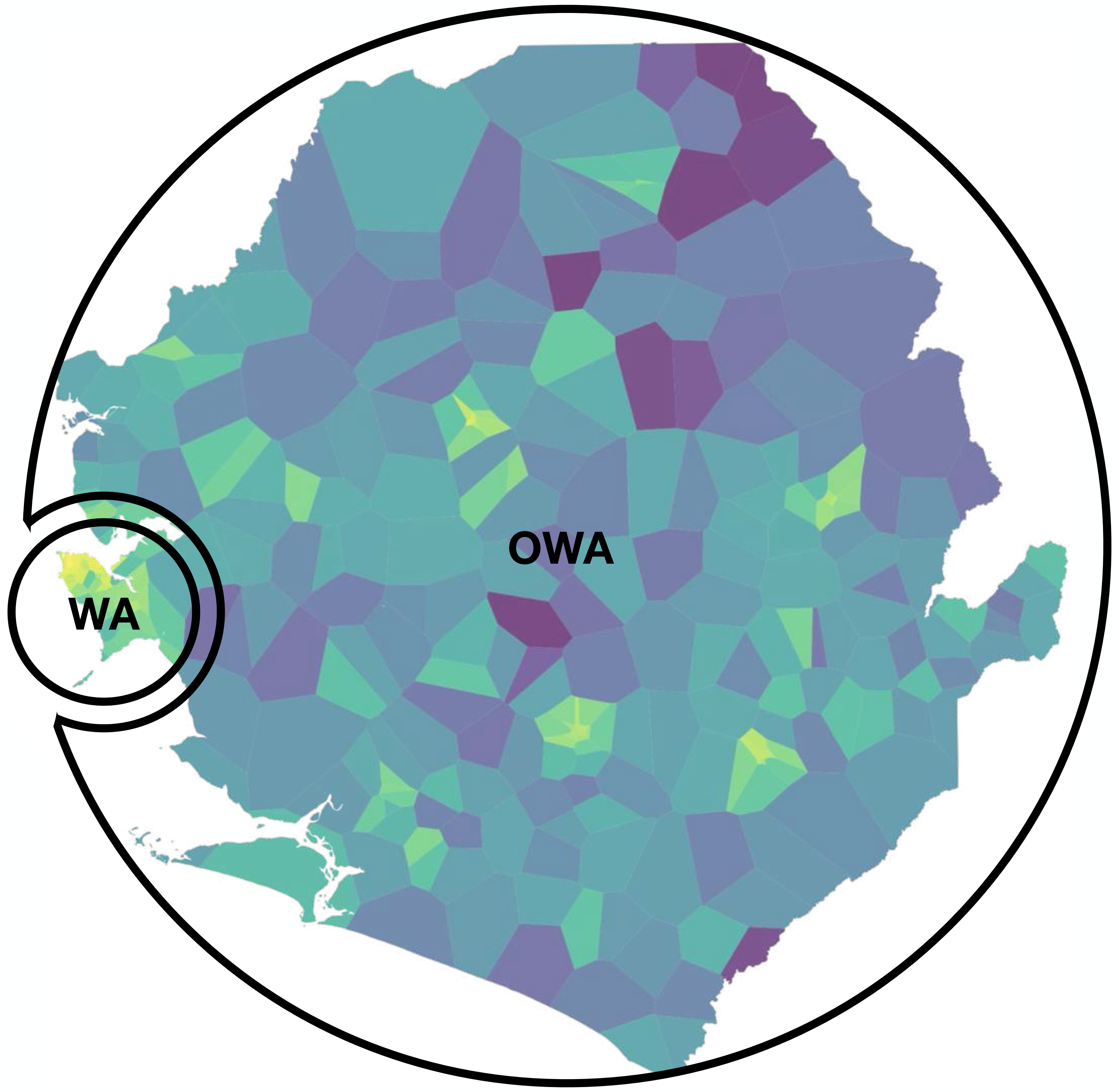
Sierra Leone - Spatial re-organisation of socioeconomic networks

Urban area

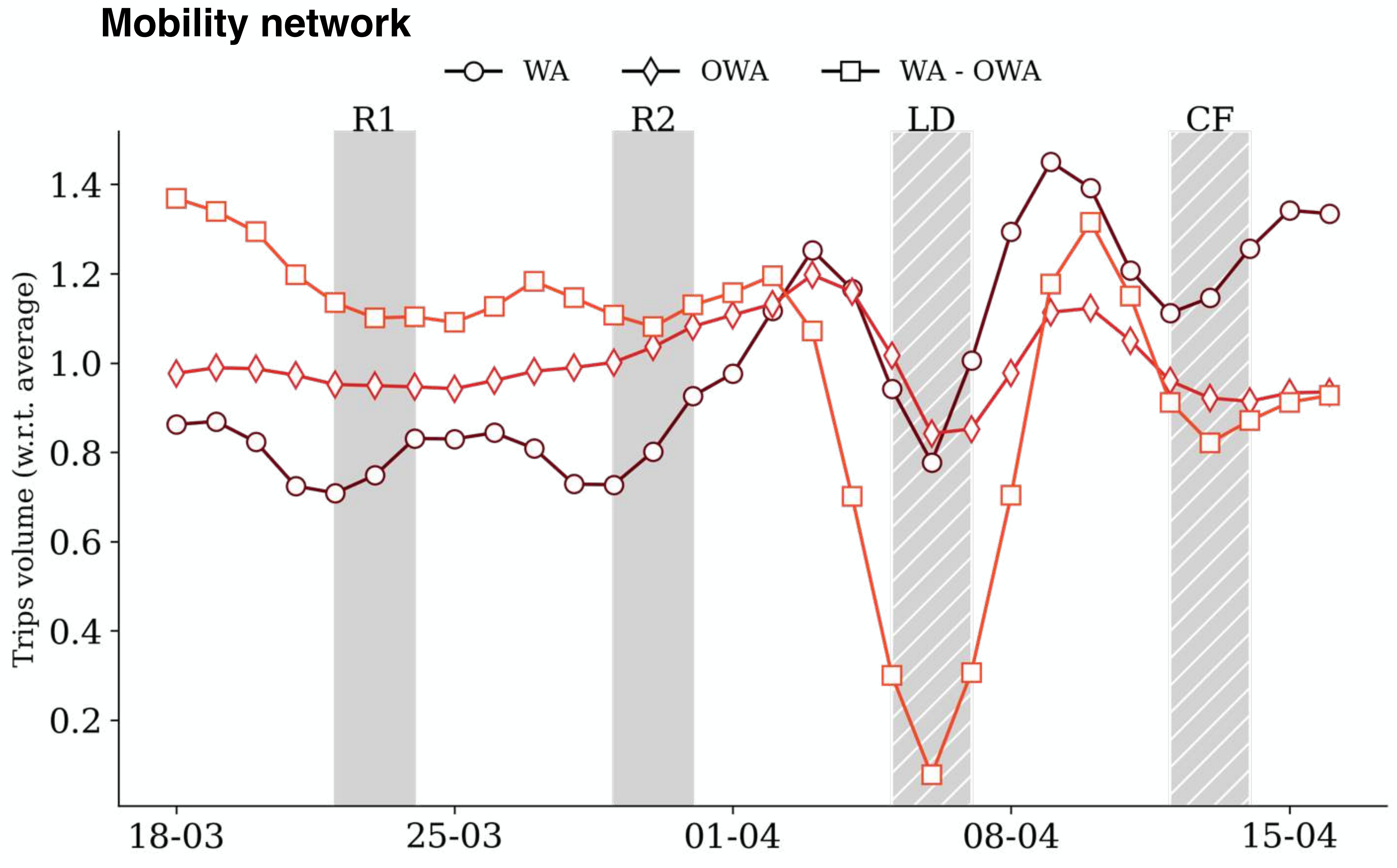
- Western Area
- Includes the capital district and the richest census tracts

Rural area

- Out of Western Area
- Everything but the capital district
- Contains mostly lower SE classes



Relatively the largest mobility reduction between WA and OWA



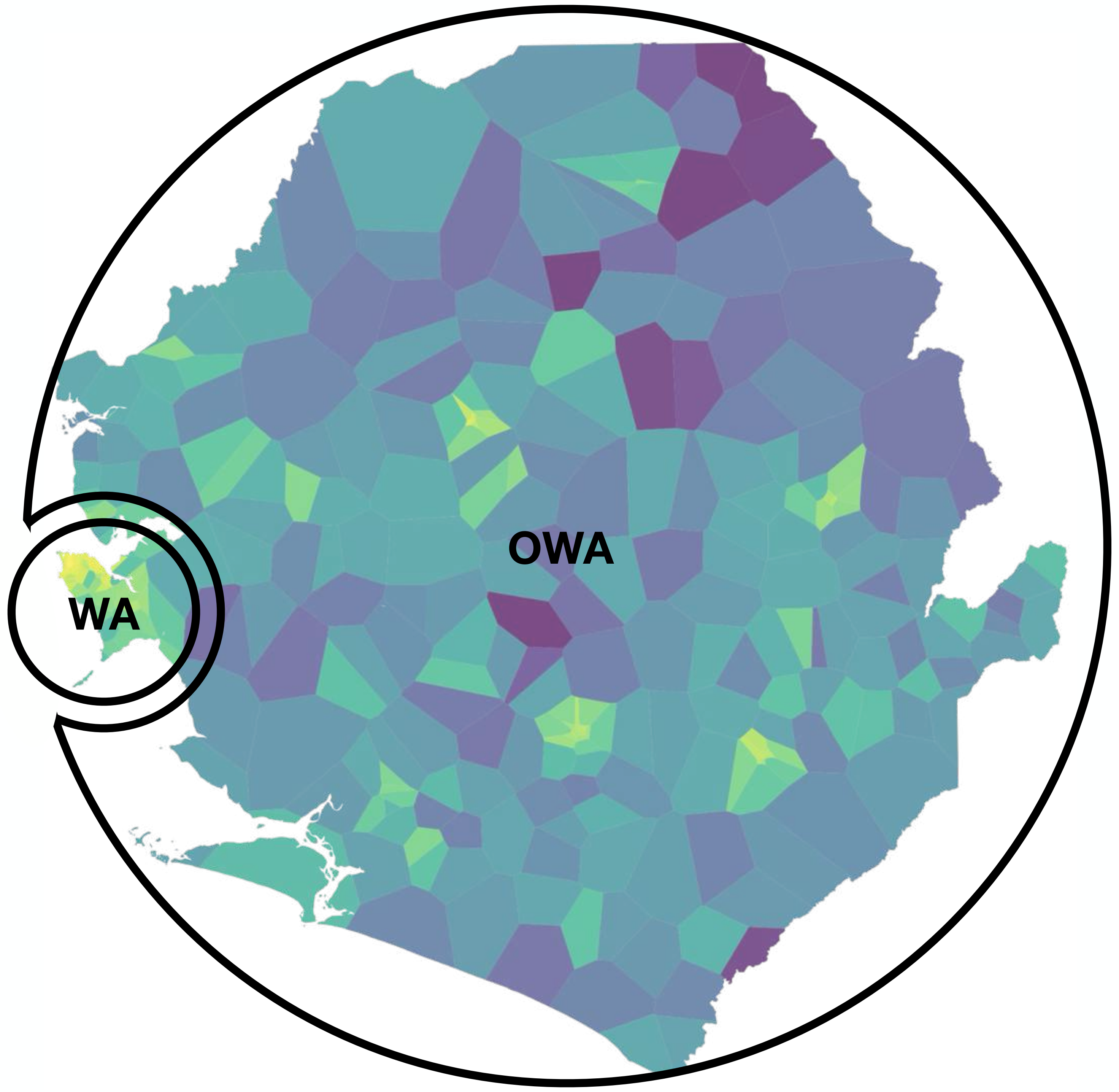
Sierra Leone - Spatial re-organisation of socioeconomic networks

Urban area

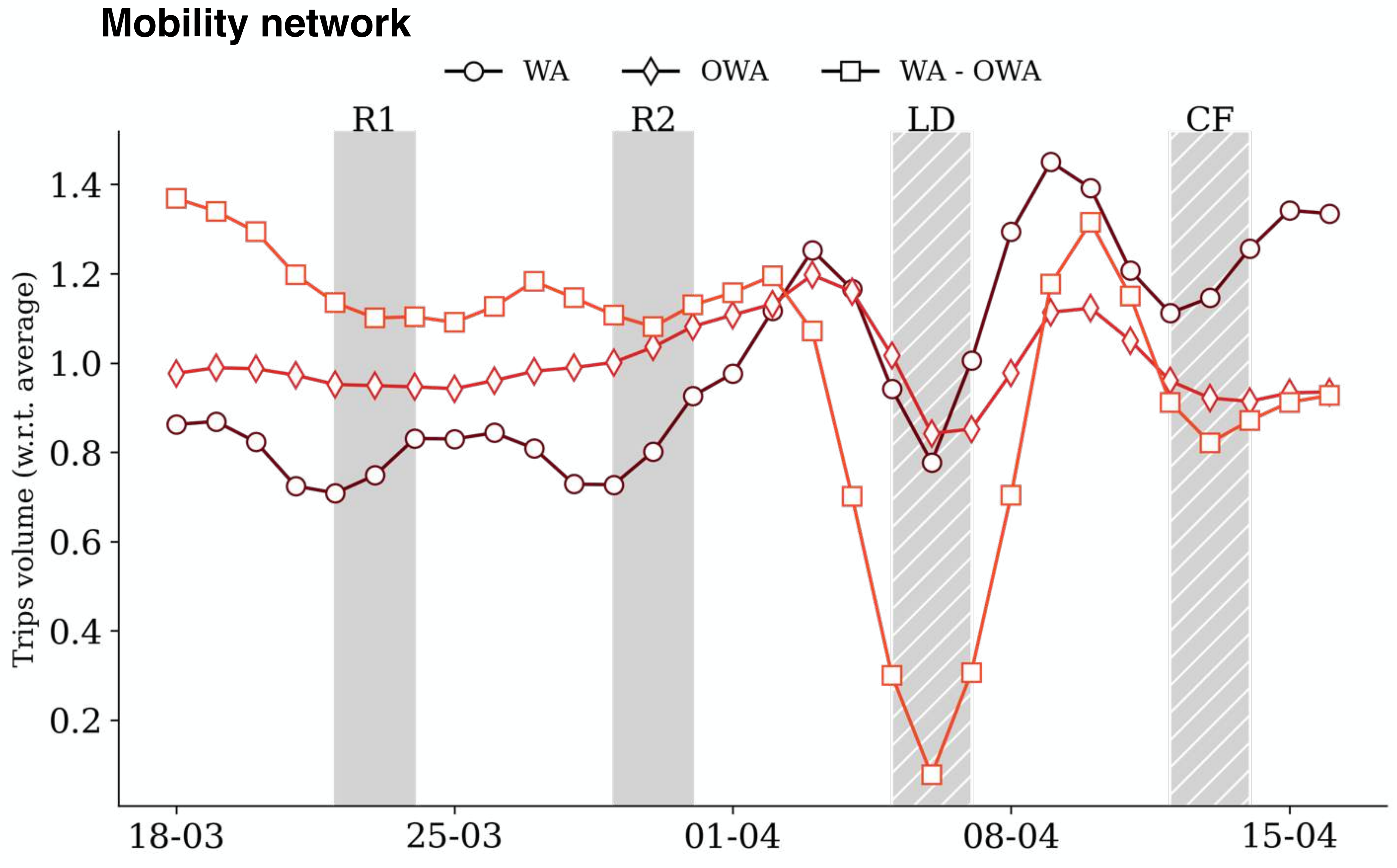
- Western Area
- Includes the capital district and the richest census tracts

Rural area

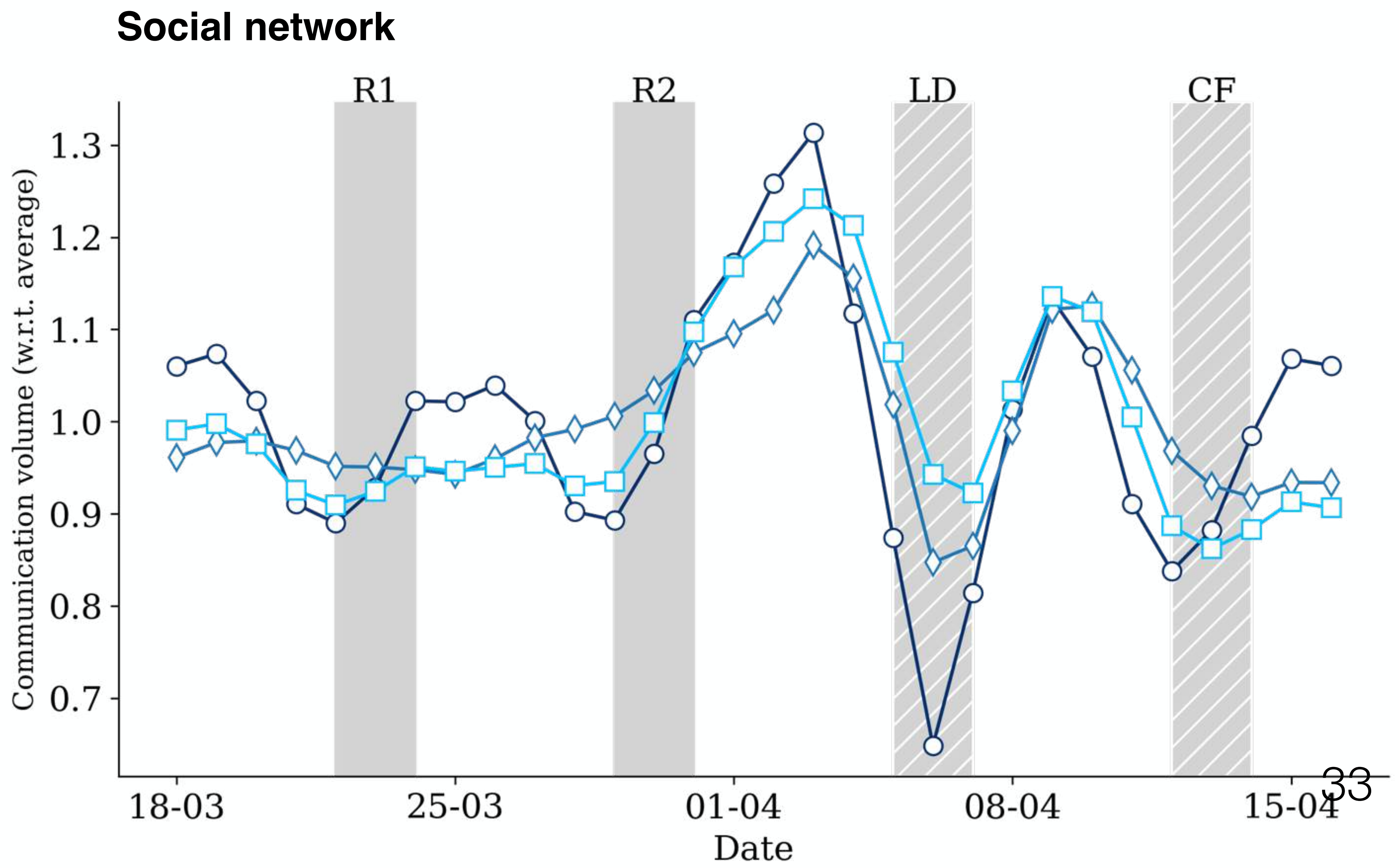
- Out of Western Area
- Everything but the capital district
- Contains mostly lower SE classes



Relatively the largest mobility reduction between WA and OWA



Relatively the smallest communication reduction between WA and OWA

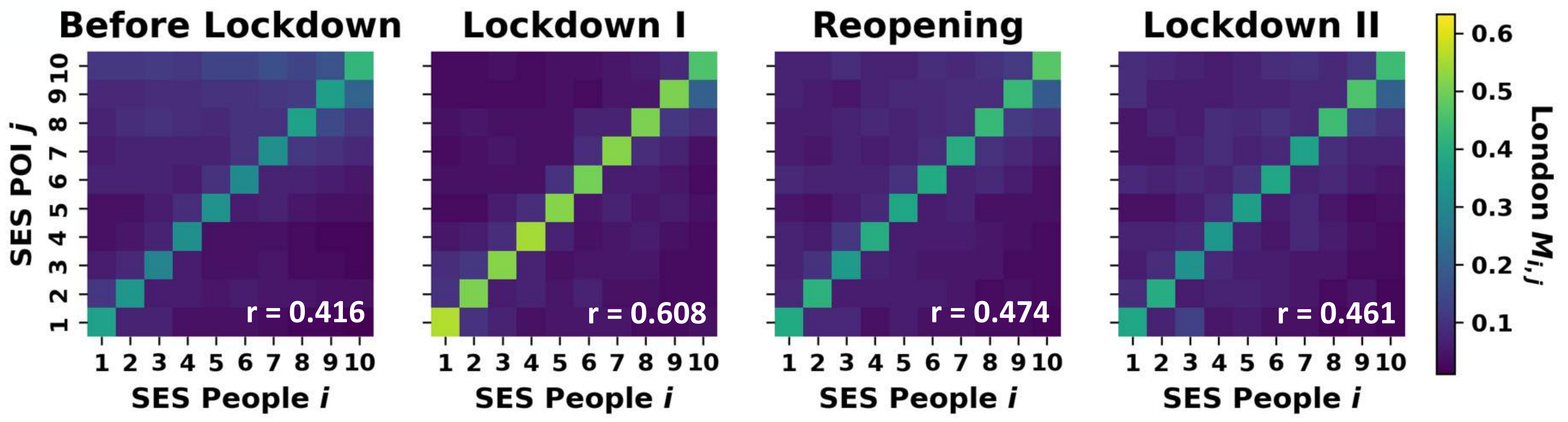


Residual mobility segregation after COVID-19 interventions

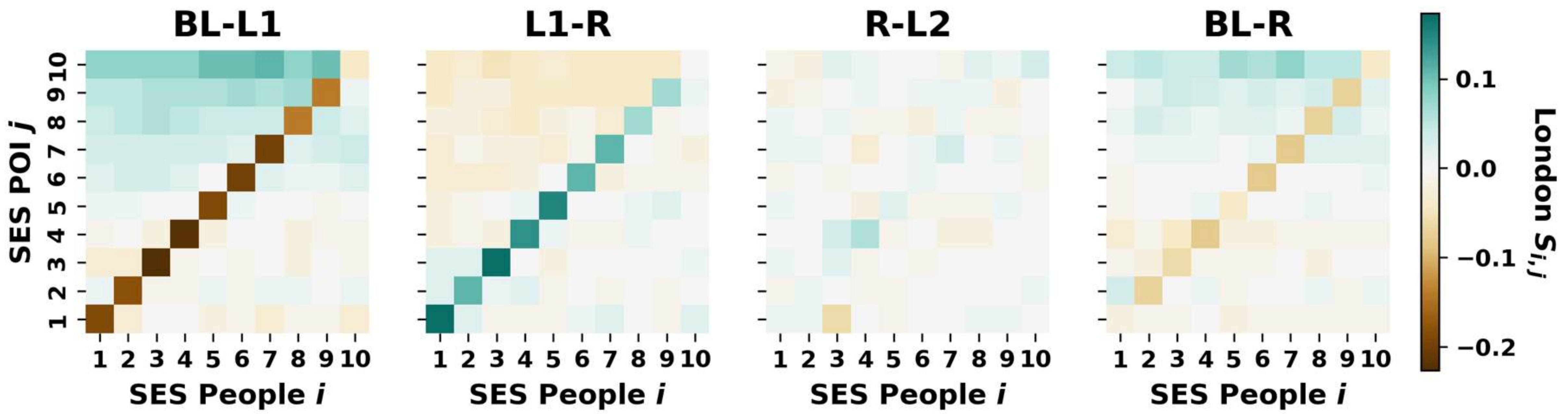
Mobility trajectories during COVID-19 intervention periods from several metropolitan areas

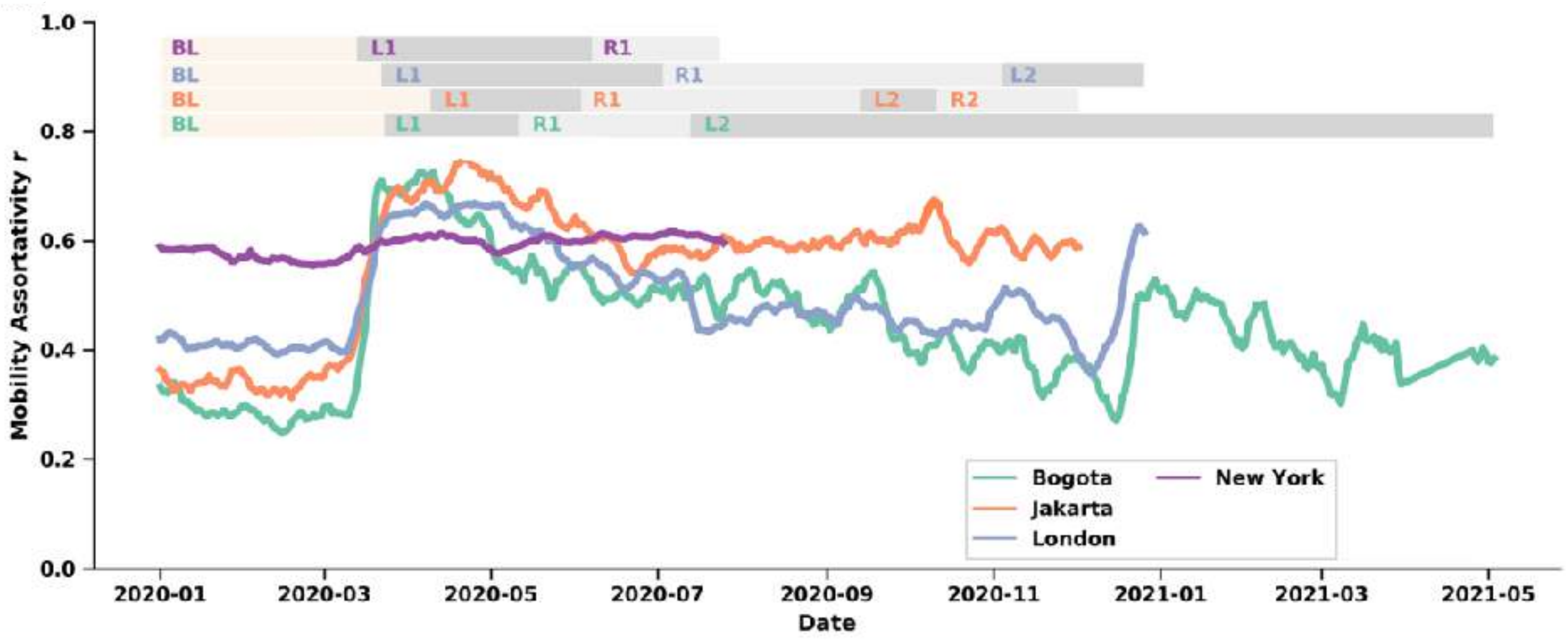
- **Bogota, Jakarta, London, New York**

Mobility assortativity matrices of London between 2020-2021



Difference matrices between different intervention periods





Hilman, R., Sekara, V., Herranz, M. & Karsai, M., ArXiv:2310.03557 (2023)

PNAS

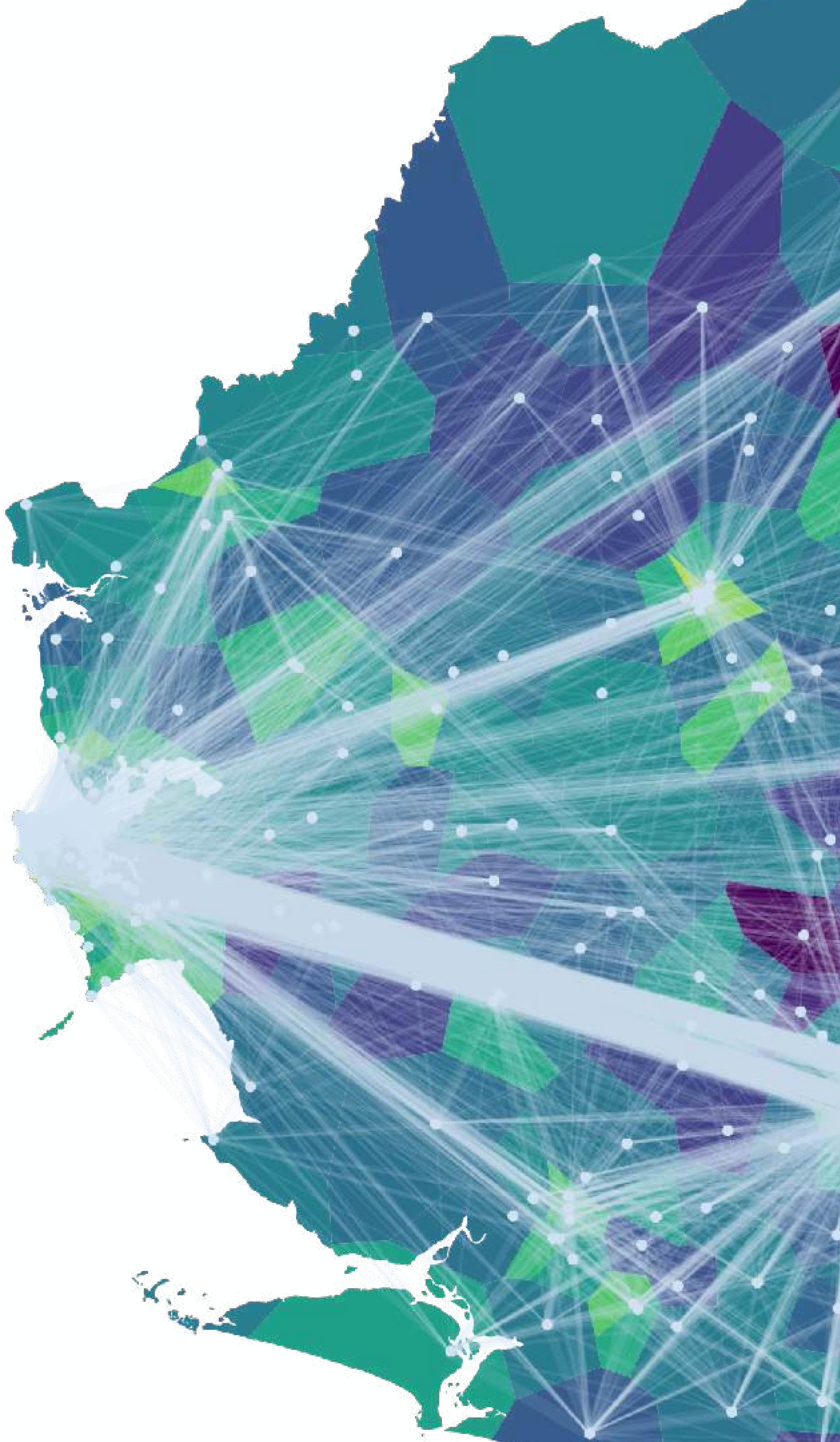
RESEARCH ARTICLE | COMPUTER SCIENCES |

Socioeconomic reorganization of communication and mobility networks in response to external shocks

Ludovico Napoli, Vedran Sekara, Manuel García-Herranz, and Márton Karsai Authors Info & Affiliations

Edited by Andrea Rinaldo, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland; received March 31, 2023; accepted October 23, 2023

December 7, 2023 | 120 (50) e2305285120 | https://doi.org/10.1073/pnas.2305285120



Language variance *vs* socioeconomic status

There is more than one way of saying the same thing, which may depend on SES, location, time, etc.

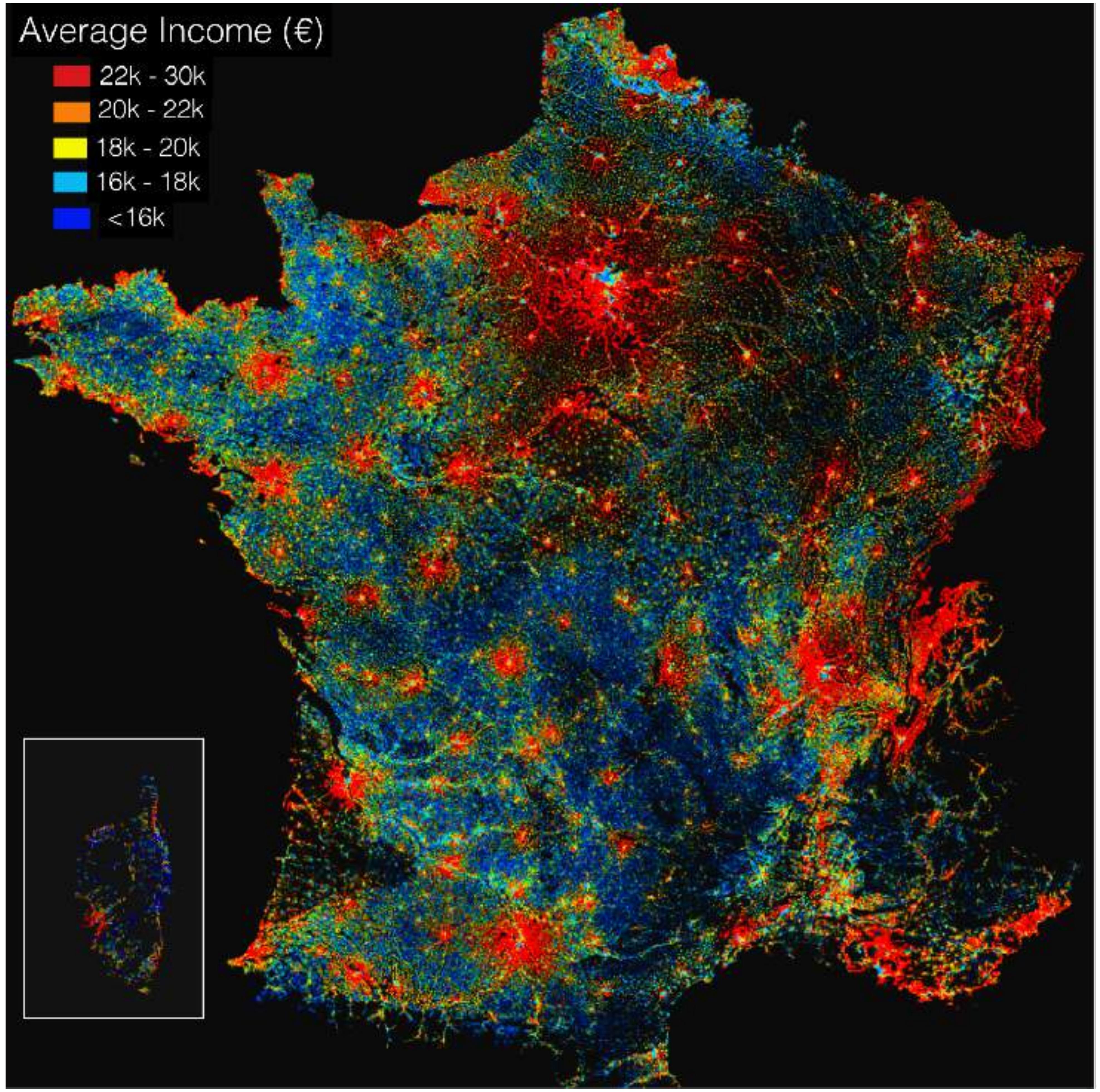
Datasets

Sociolinguistic features

- Twitter Decahose feed from 2 recording periods: July 2014 - June 2015 and February 2016 - May 2017
  —> 15-25 % of all public communication in Twitter (France)
- **2.5M** users, **170M** messages, social network, profile meta-information
- Joint Filters:
  - Tweets generated in GMT - GMT+1 timezone
  - Written in French
- **Linguistic data**: Retweets, URLs, emoticons, mentions(@), hashtags (#), punctuation removed + detection linguistic markers
- **Network data**: Social ties between pairs of users identified by taking mentions as proxies of social interactions
  —> 509K users and 4M links
- **Geolocated data**: 2% of dataset: self-provided position or the place from which the tweet was posted.

Socioeconomic features

- Set of sociodemographic aggregated indicators for each 200m x 200m square patch across France



Indicators proxy of SES

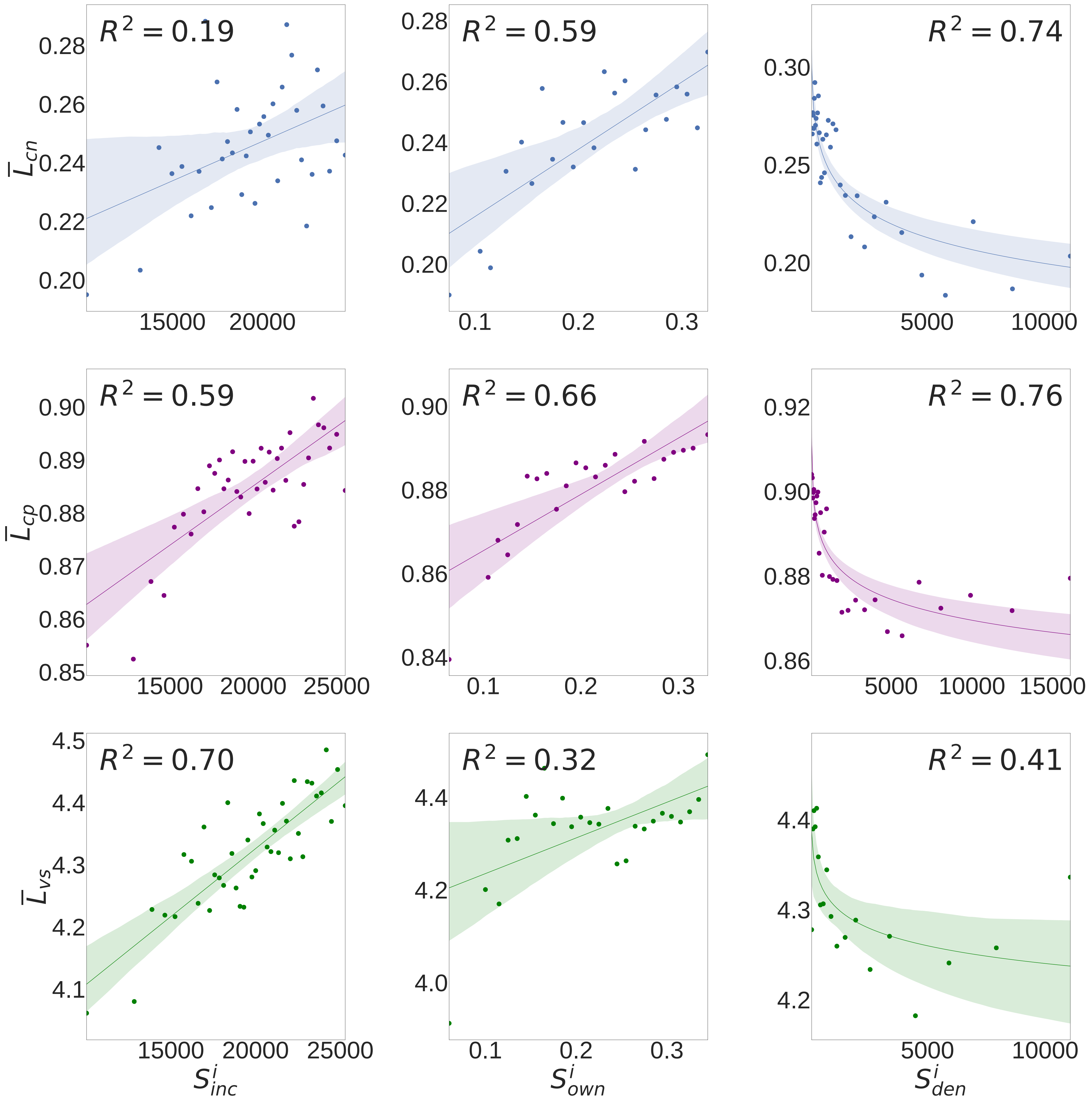
- Mean yearly income
- Fraction of owners
- Population density

Linguistic variables

Linguistic markers:

- **Frequently** used by the whole population
- Assign **evidently** the standard/non-standard usage of language

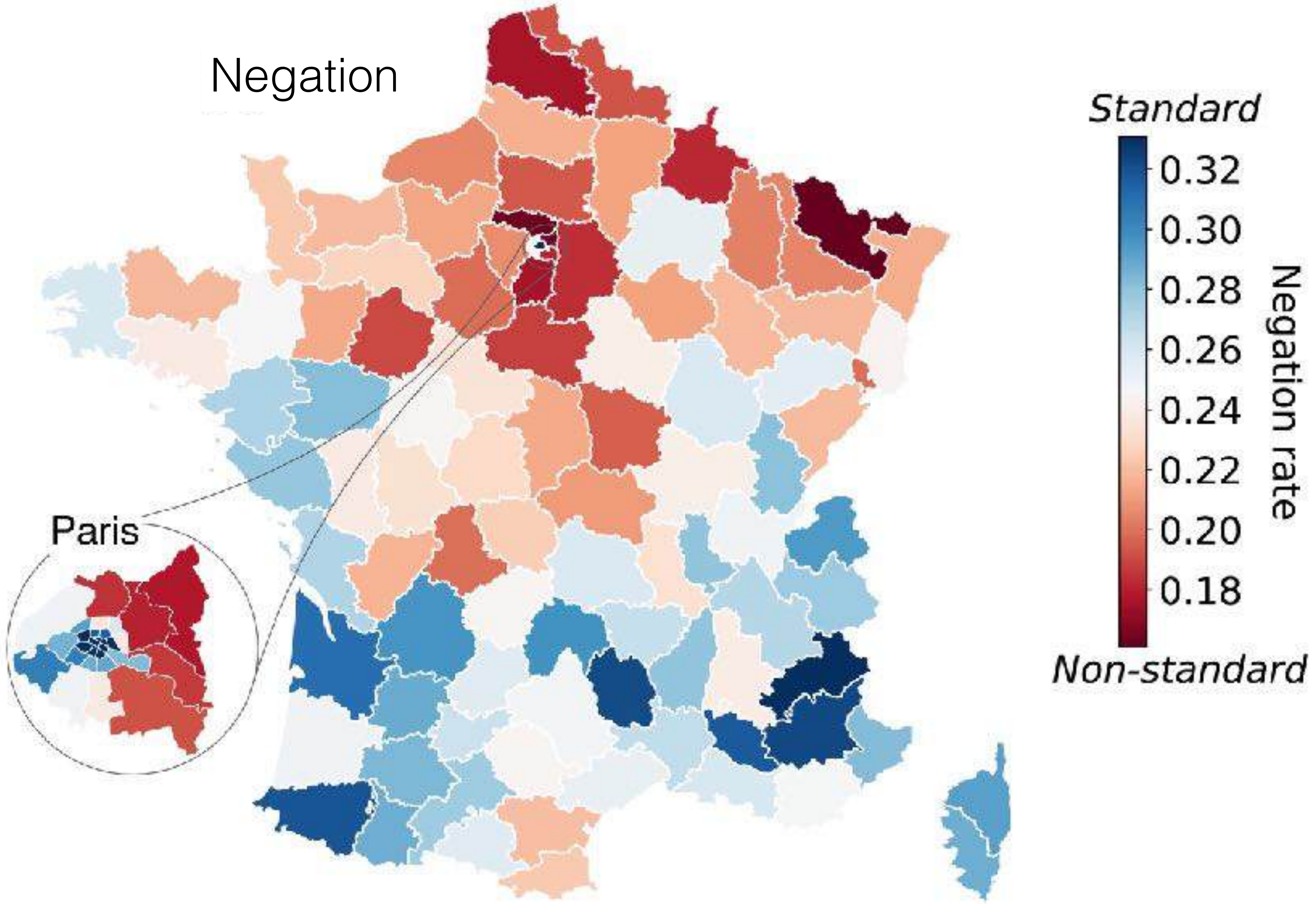
	Standard	Non-Standard	Metric
Negations	NE + verb + pas 'not' *je n ' ai pas*	NE deletion *je ' ai pas*	$L^u_{cn} = \frac{n^u_{cn}}{n^u_{cn} + n^u_{incn}}$
Plurals	Plurals with final '-s' / 'x': *des actes*	Final '-s' /'x' deleted *les policier*	$L^u_{cp} = \frac{n^u_{cp}}{n^u_{cp} + n^u_{incp}}$
Normalized vocabulary size			$L^u_{vs} = \frac{N^u_{vs}}{N^u_{lw}}$



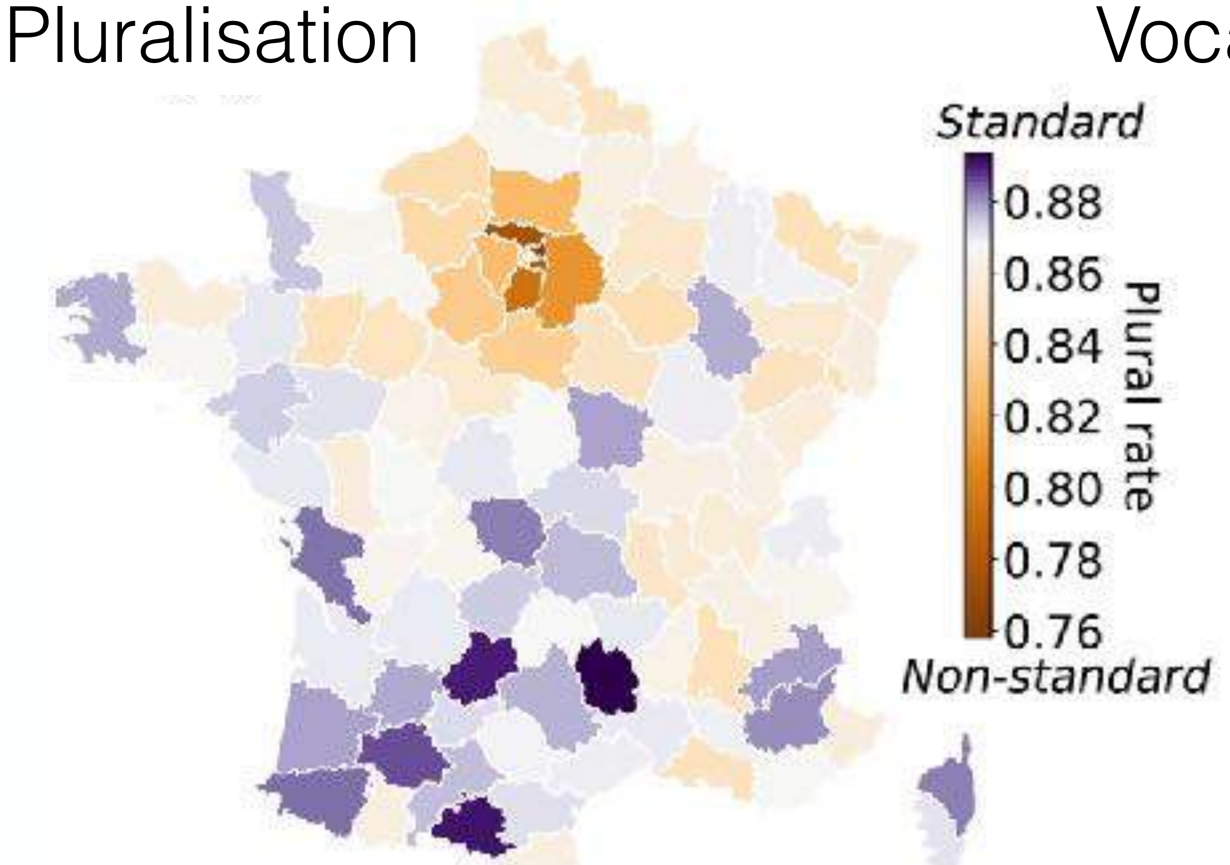
Lower SES users more prone to use **non-standard** expressions and have a **smaller vocabulary** set size than high SES users

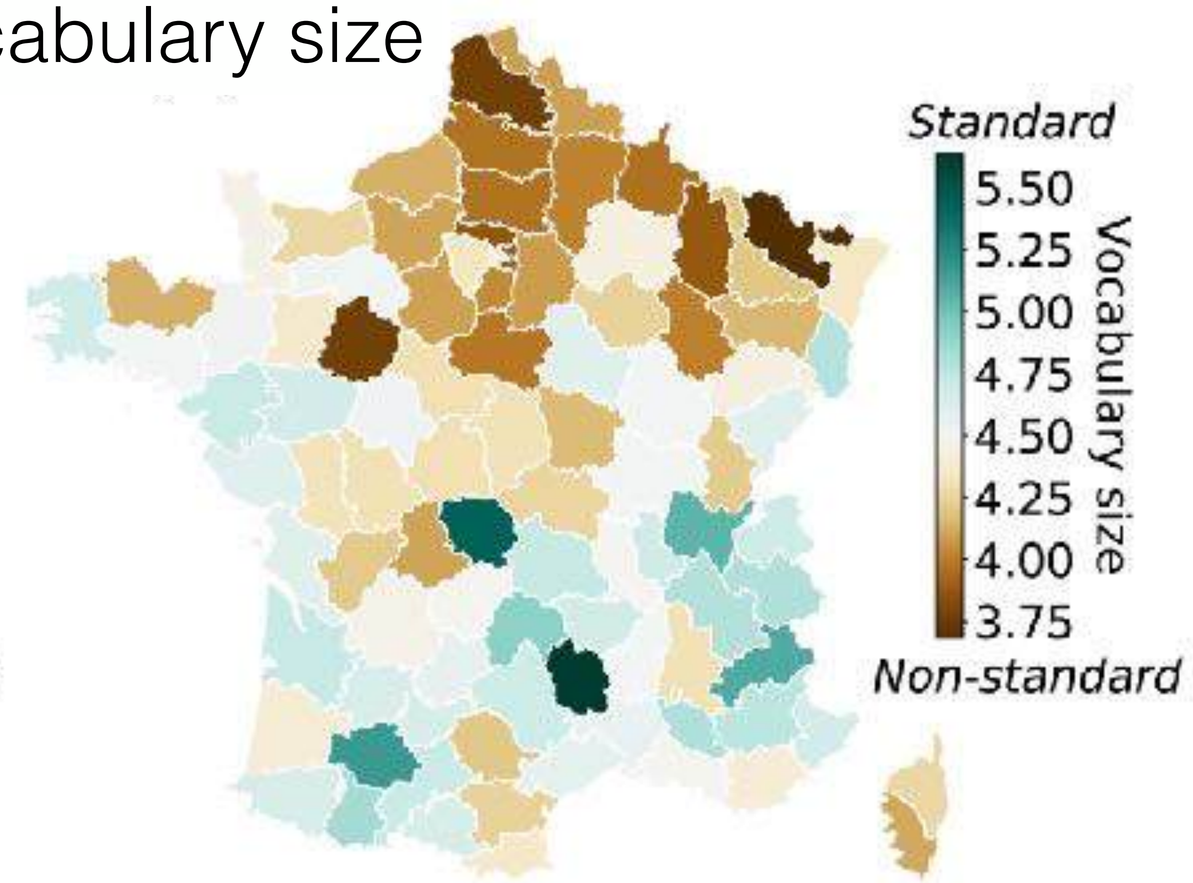
Spatial correlations

- **Study of the regional variability of selected linguistic markers**
- Assign each geolocated user to a department of France
- Compute average values of linguistic markers for each department i in the country



- **Large scale**: people living in Northern France use less standard language in comparison to Southern
- **Short scale**: standard language usage is determined by local variation of SES (as seen for Paris)
- **Location (globally)** and **socioeconomic variability (locally)** determines the language usage.



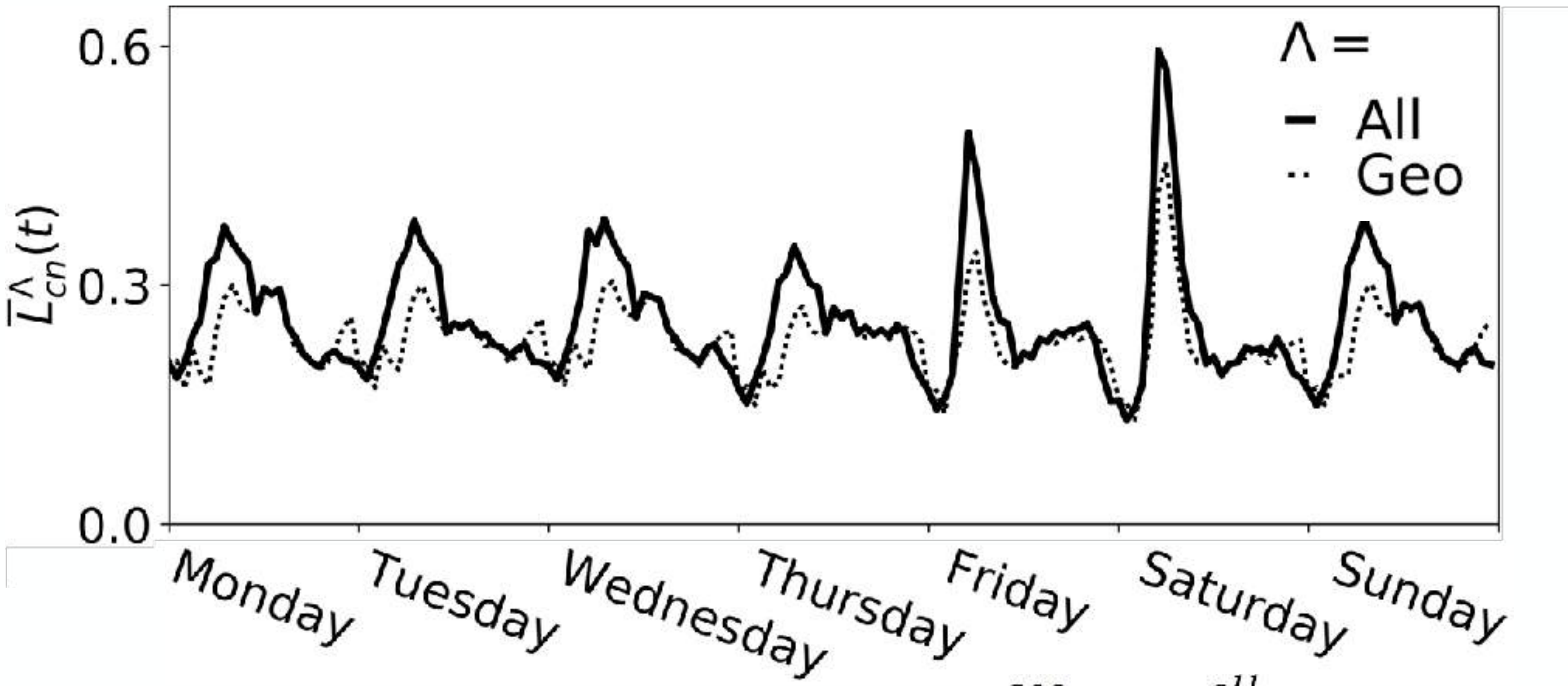


Temporal Correlations

Temporal user activity patterns vs. language variability

- Aggregate negations by hour over 1 week for each user
- Compute average value for each hour
- Significant correlation between 2 temporal series : Geolocated users representative of whole set of users with respect to temporal pattern.
- **Standard language is used with higher rate during the day than during the night**

Standard negation

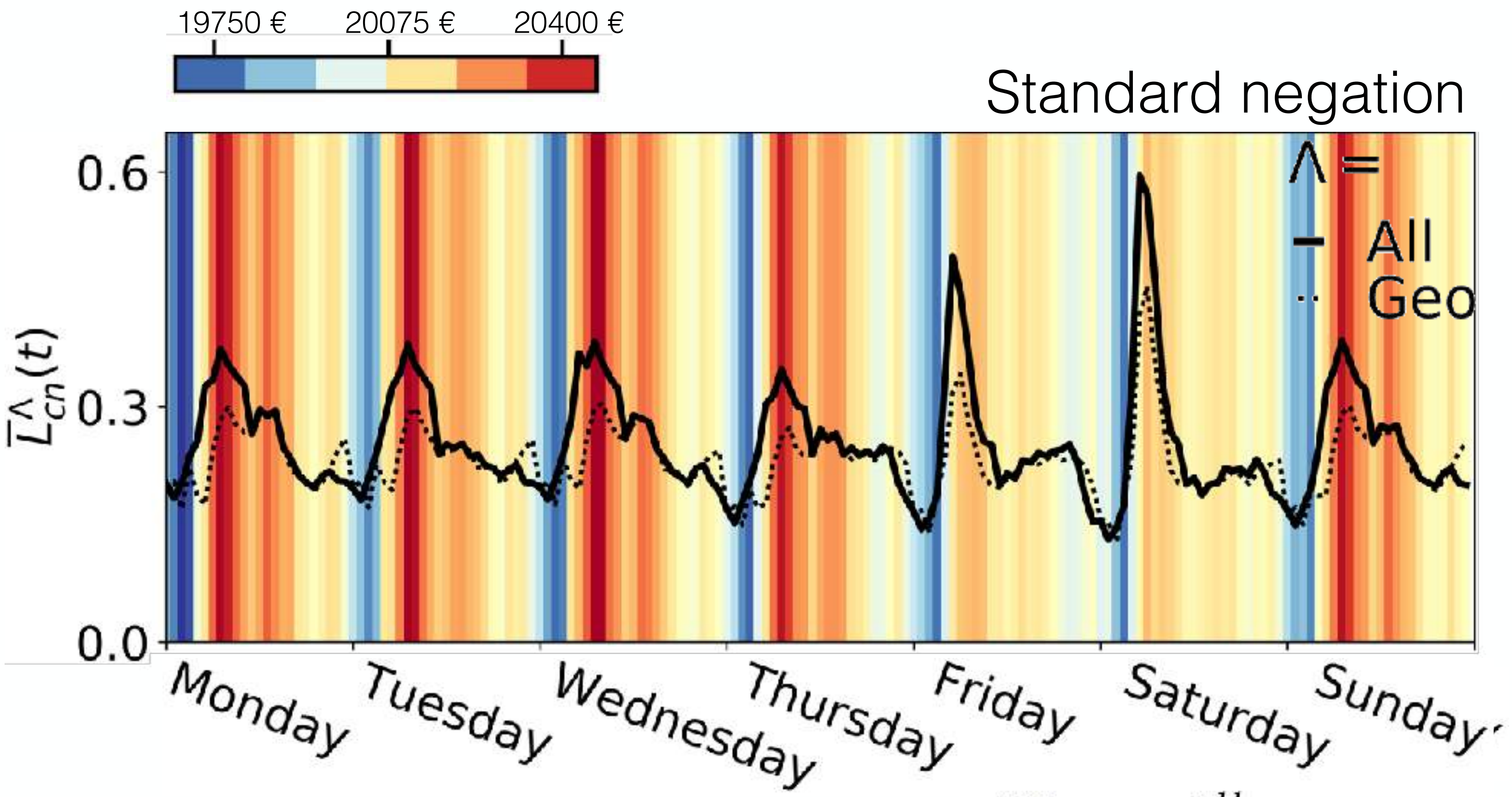


r($\overline{L}^{geo}_{\text{cn}}(t) \sim \overline{L}^{all}_{\text{cn}}(t)$) = 0.81

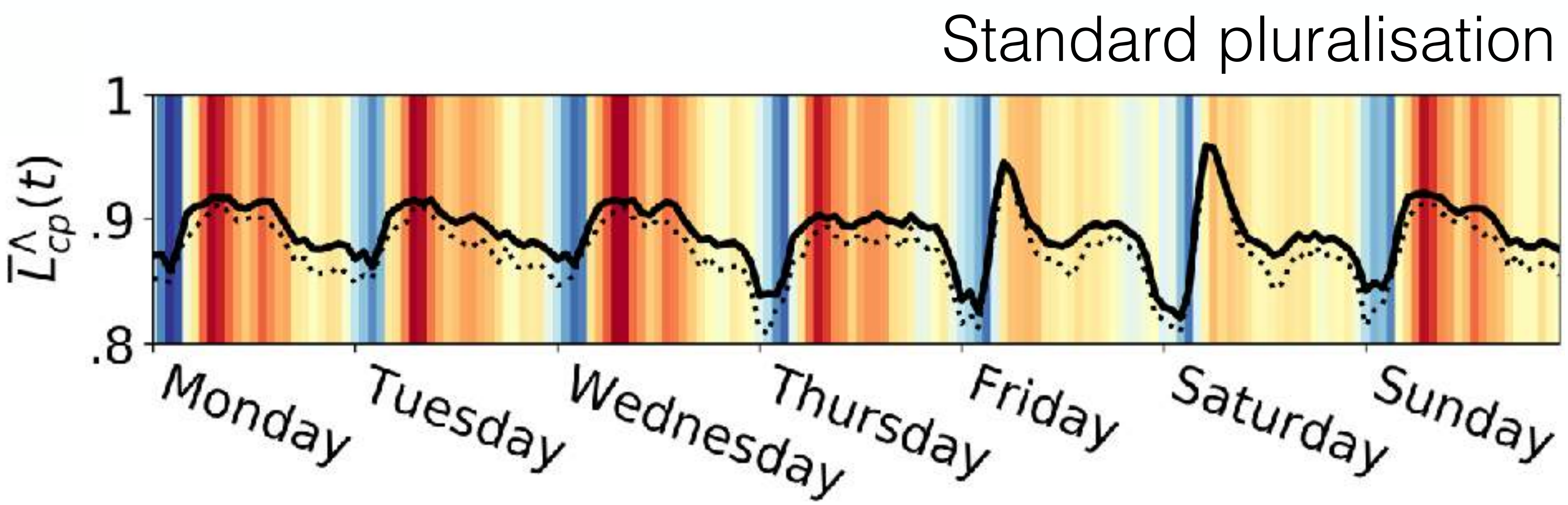
Temporal Correlations

Demographics of average users, which are active in Twitter through the day

- Measure average income of active users in a given hour over a week
- Significant **correlations between temporal variability of average linguistic variables and average income** of the active population on Twitter.
- Main factor behind observed linguistic variability: **People active during the day have a higher average income** than people active during the night



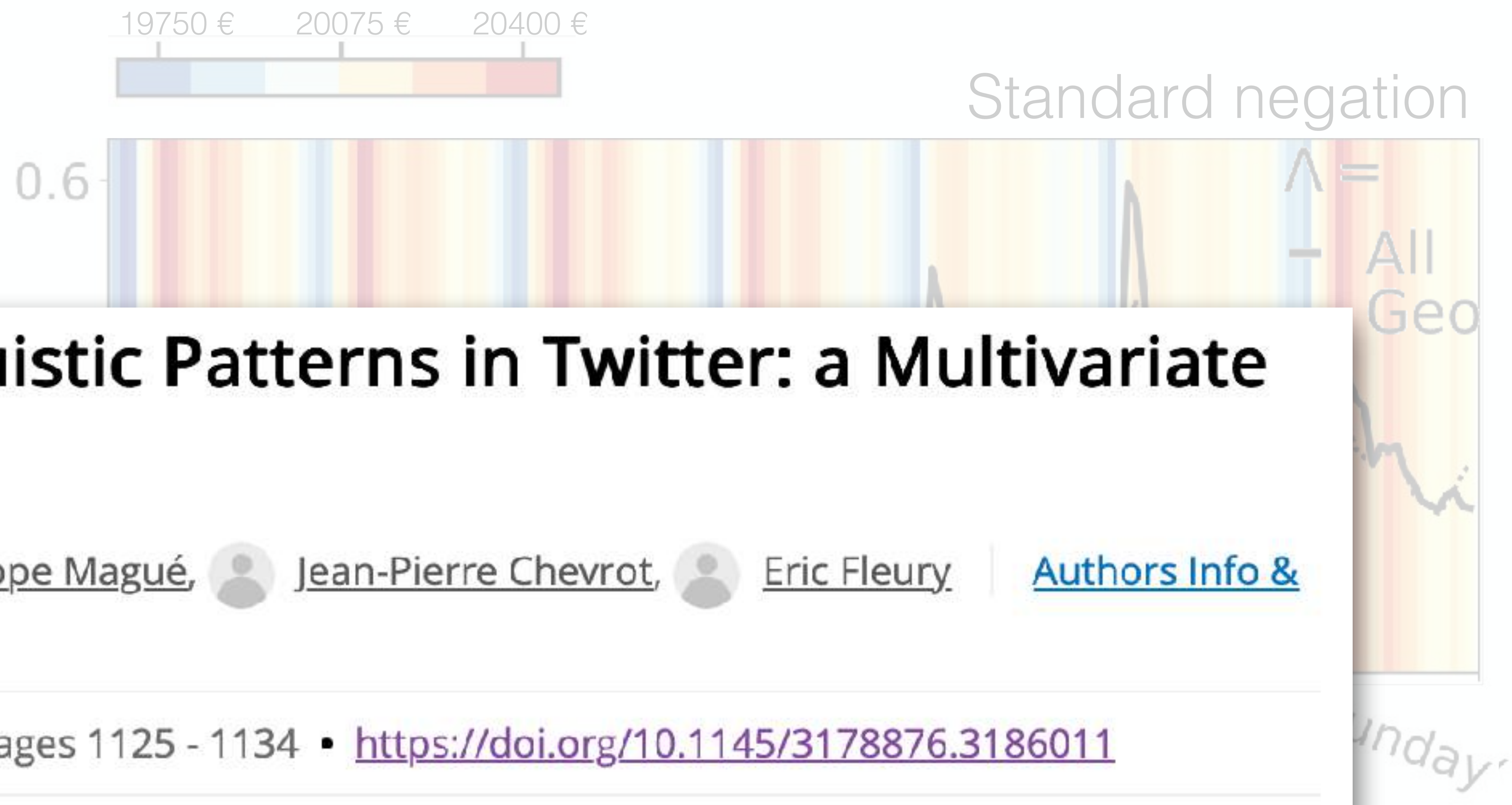
$r(\,\overline{L}^{geo}_{cn}(t) \sim \overline{L}^{all}_{cn}(t)\,) = 0.81$



Temporal Correlations

Demographics of average users, which are active in Twitter through the day

- Me… giv…
- Sig… **va… an…** on…
- Ma… variability: **People active during the day have a higher average income** than people active during the night



Socioeconomic Dependencies of Linguistic Patterns in Twitter: a Multivariate Analysis

Authors: Jacob Levy Abitbol, Márton Karsai, Jean-Philippe Magué, Jean-Pierre Chevrot, Eric Fleury | Authors Info & Claims

WWW '18: Proceedings of the 2018 World Wide Web Conference • Pages 1125 - 1134 • https://doi.org/10.1145/3178876.3186011

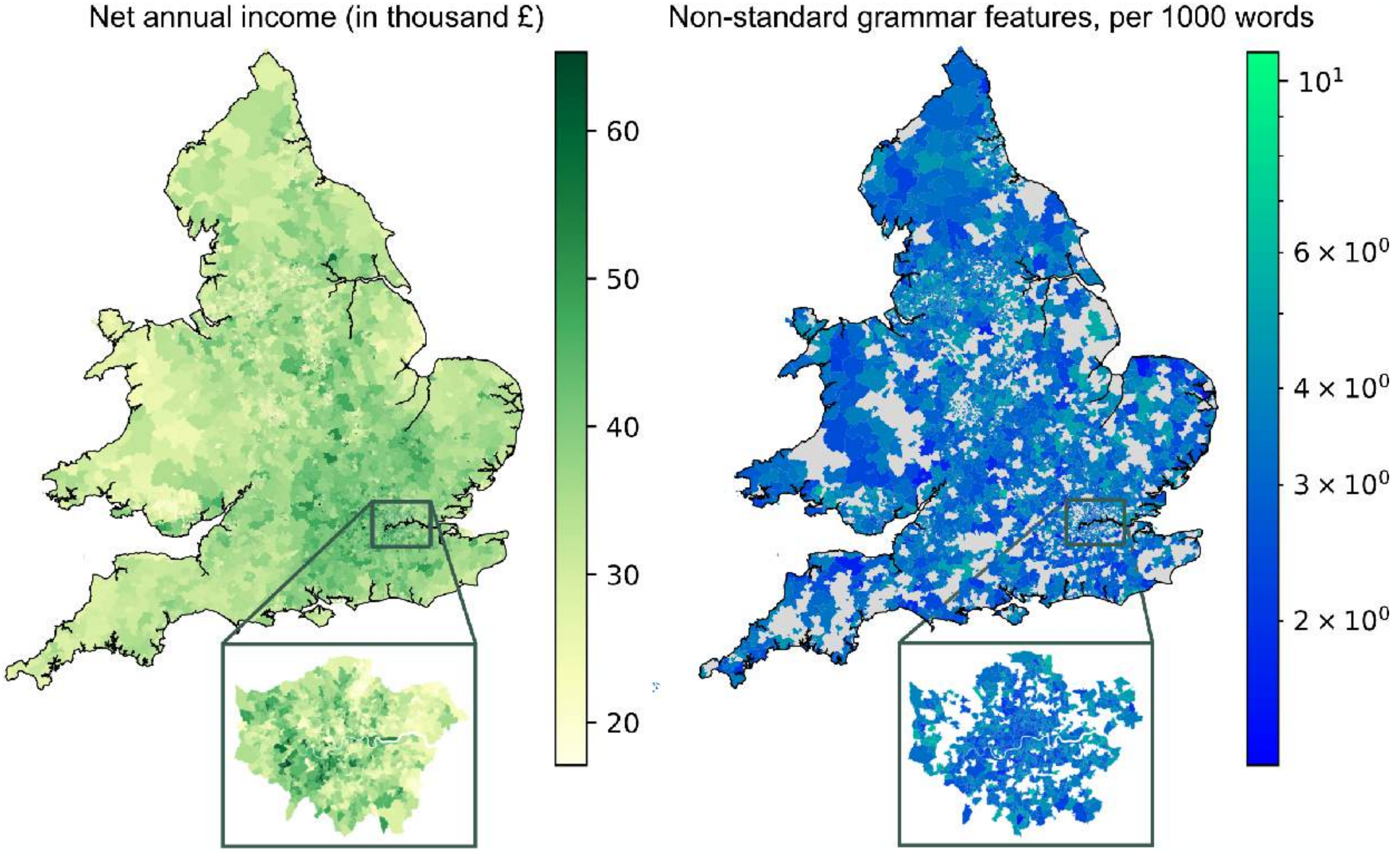
Published: 10 April 2018 Publication History

Check for updates

Standard language use and socioeconomic mixing in the UK

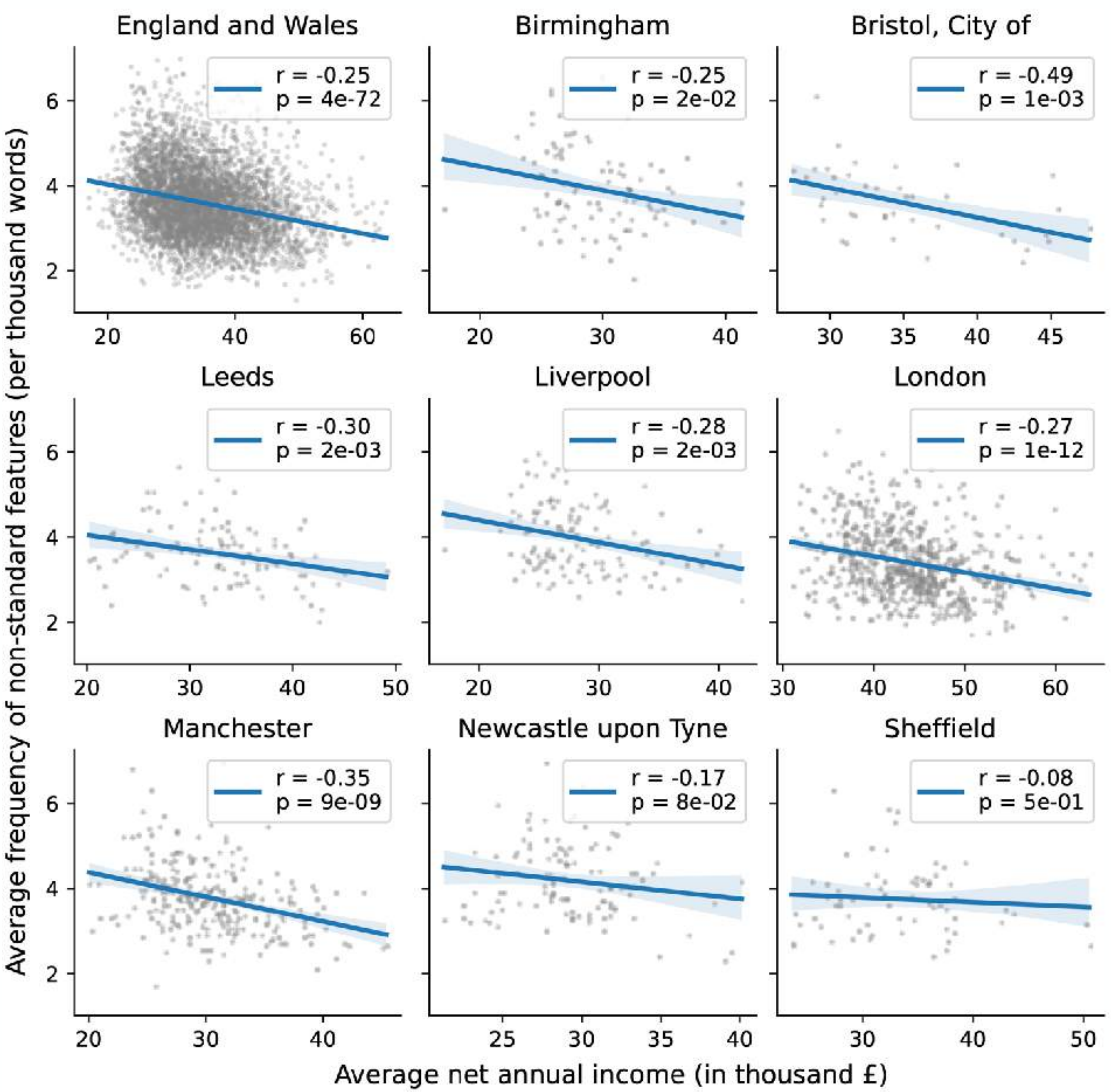
Data combination:

- 550M geolocated Tweet in the UK and Wales between 2015 and 2021
- Average income map from NSI
- LanguageTools - frequency of deviation from standard language rules per word per user
- Home location and visited place inference
- Final set: 130K user from 4879 census tracts in 8 metropolitan areas



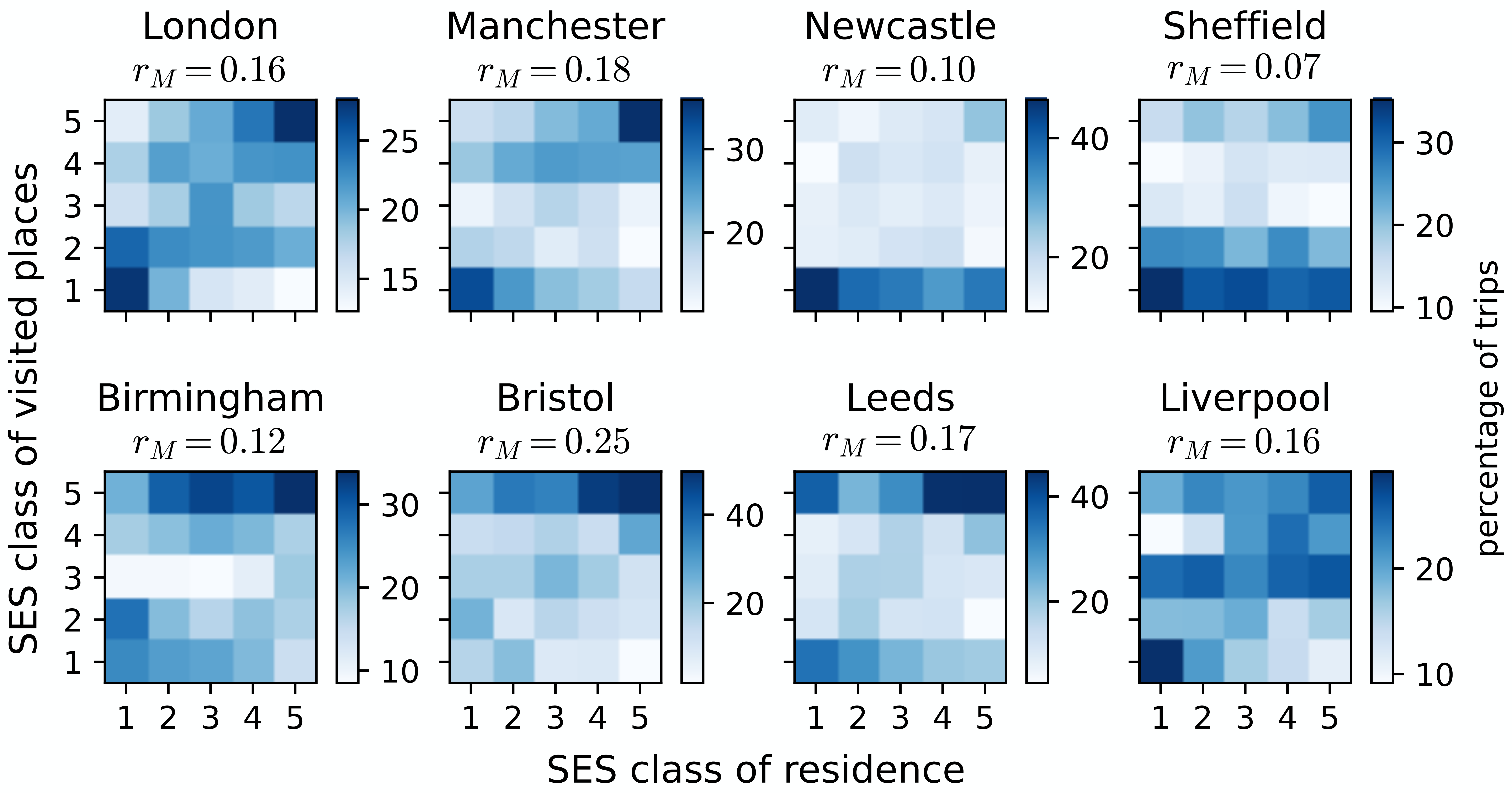
Language variation vs. socioeconomic status

- Average income correlations with frequency of non-standard language use in different metropolitan areas
- Not strong but mostly significant **negative correlations**
- Strongest correlation: Bristol
- Weakest correlation: Sheffield (not even significant)



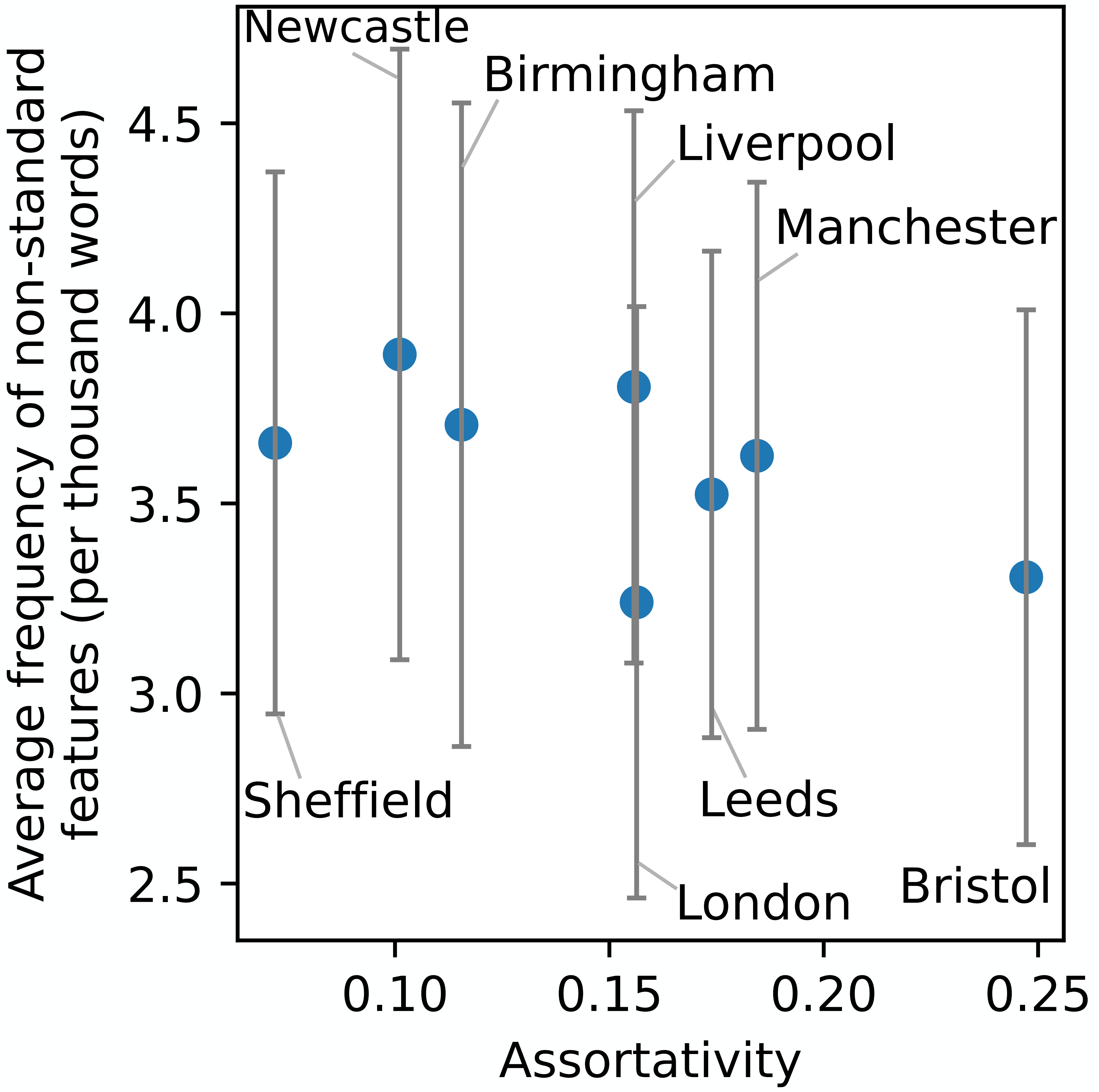
Socioeconomic mixing in mobility

- Assortative mixing in all metropolitan areas
- Some are biased towards on class due to geographical constraints of urban design



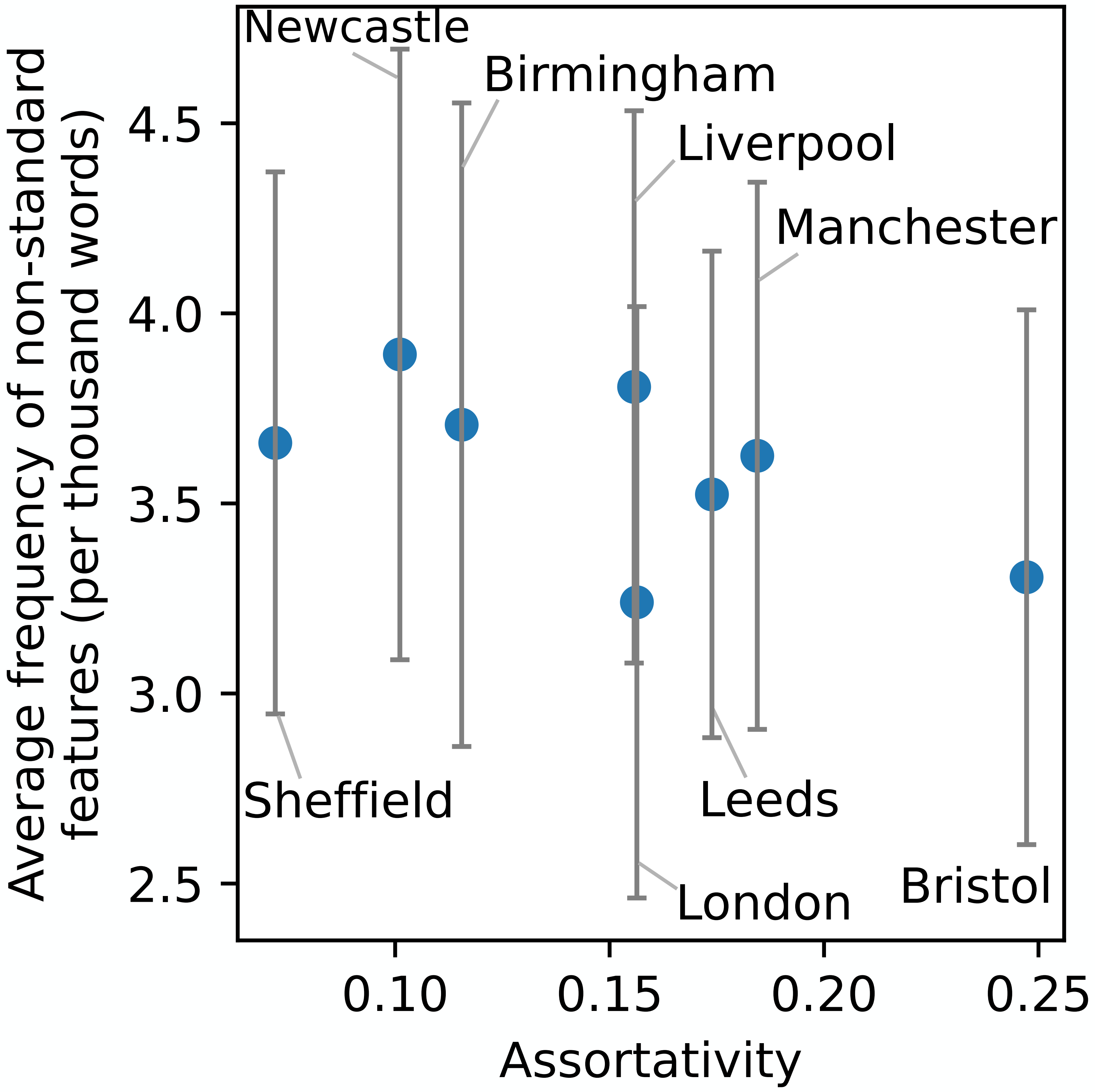
Language variation vs. socioeconomic mixing

Language variation
vs. Assortativity

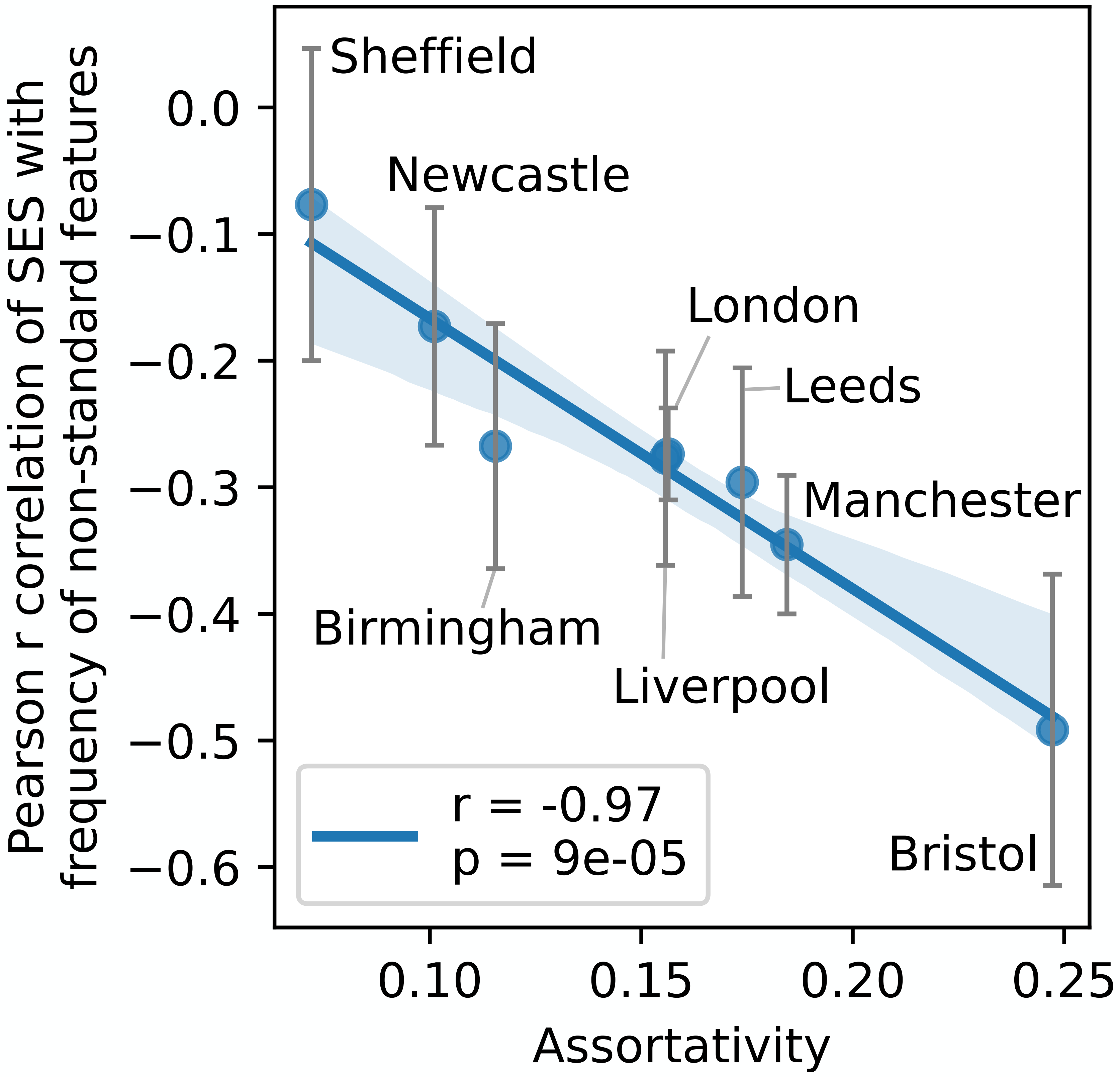


Language variation vs. socioeconomic mixing

Language variation vs. Assortativity

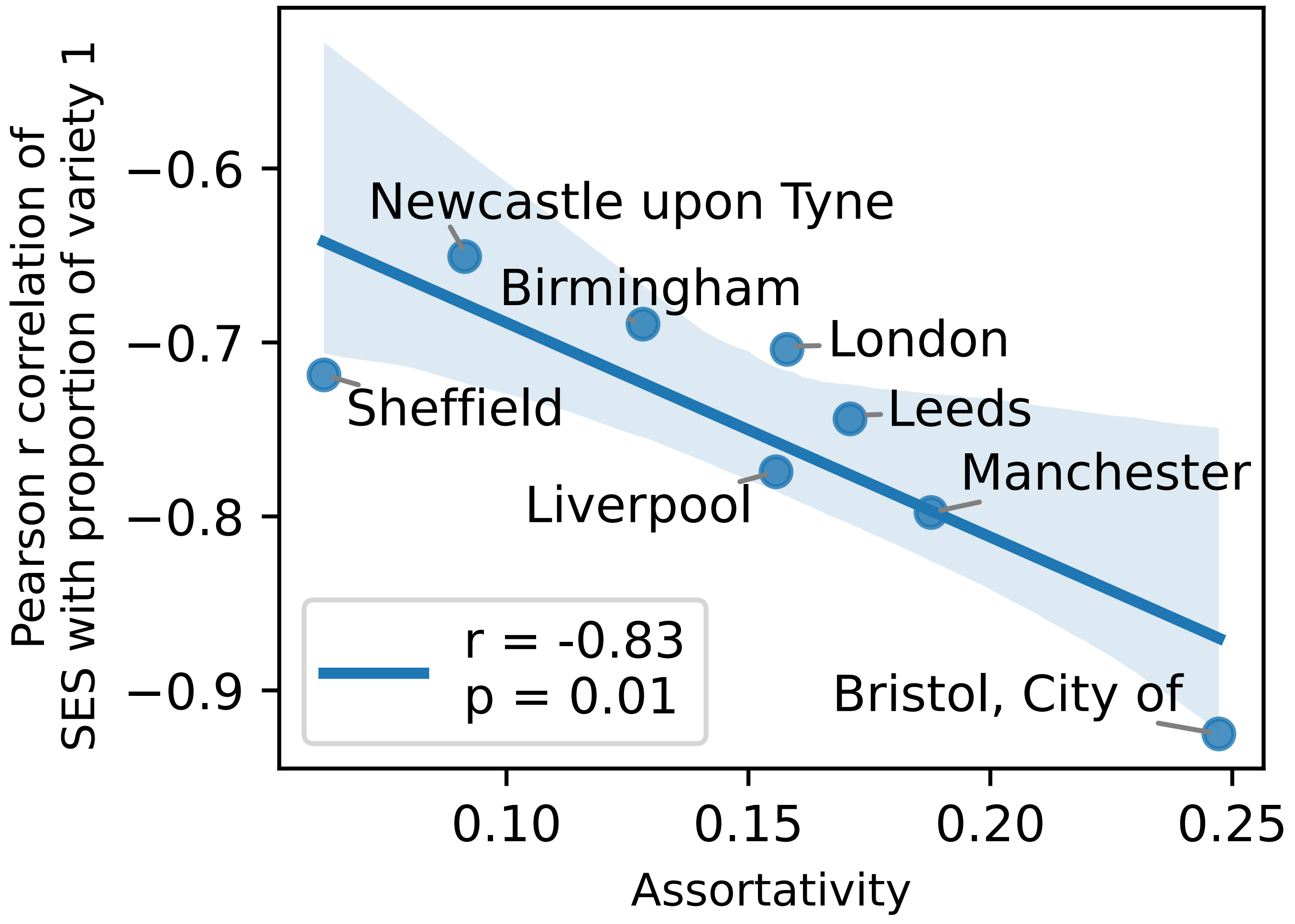


Language variation - SES correlations vs. Assortativity



Paper spoiler

- Data driven agent-based model can reproduce the observed correlation



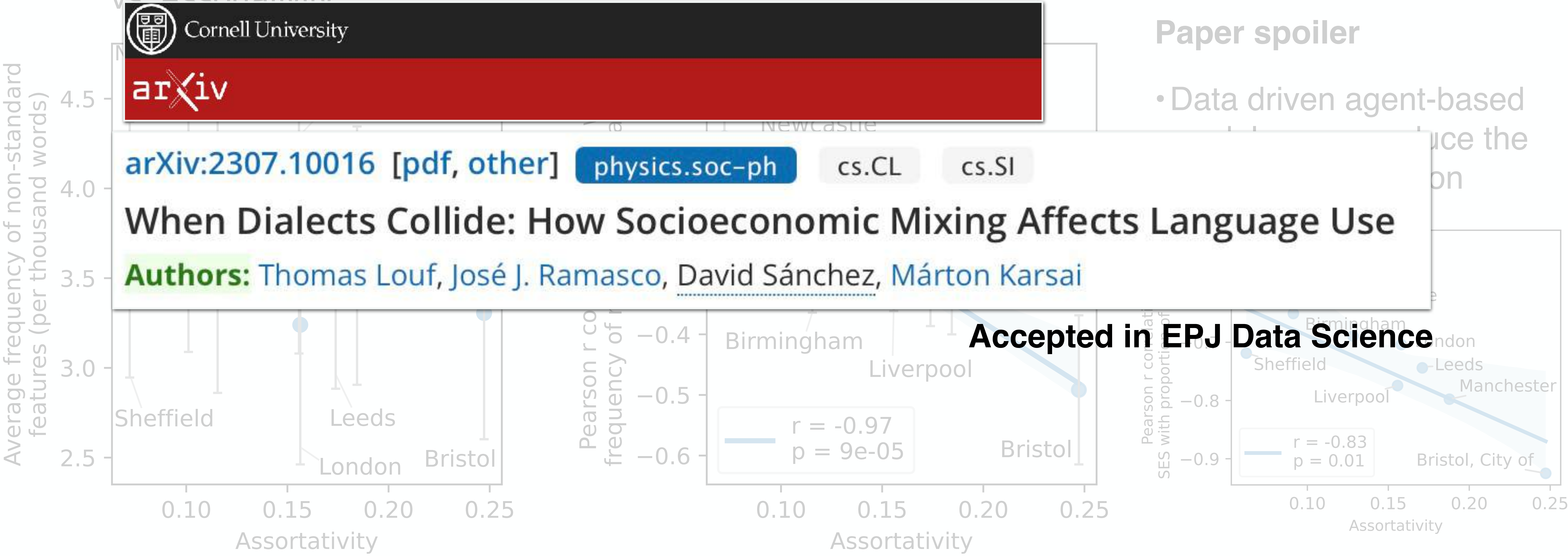
Language variation vs. socioeconomic mixing

Language variation vs. Assortativity

Language variation - SES correlations vs.

Paper spoiler

- Data driven agent-based ... uce the ... on



Take home

Socioeconomic inequalities matter in social and mobility network formation and language usage

The observation of socioeconomic networks is difficult but possible via deep learning methods using multiple data sources

Socioeconomic patterns are rigid but can re-organise rapidly due to external shocks

The Computational Human Dynamics Team @ CEU

Márton Karsai

Principal Investigator

Lisette Espín Noboa

Postdoctoral Fellow - CEU

Remi Vaudaine

Postdoctoral Fellow - ENS Lyon

Gergely Ódor

SNSF Postdoctoral Fellow - CEU

Arash Badie-Modiri

Postdoctoral Fellow - CEU (co-supervised with János Kertész)

Sámuel Balogh

Postdoctoral fellow Rényi Institute

Rafiazka Hilman

PhD student - CEU

Ludovico Napoli

PhD student - CEU

Adriana Manna

PhD student - CEU

Elsa Andres

PhD student - CEU

Timur Naushirvanov

PhD student - CEU

András Hubai

Research Engineer Rényi Institute

Csegő Kolok

Visiting PhD student Rényi Institute - ELTE

In Collaboration With…

Eszter Bokányi
Jean-Pierre Chevrot
Eric Fleury
Ignacio A Hamelin
Gerardo Iniguez
Yannick Leo
Jacobo Levy Abitbol
Manuel Garcia-Herranz
Serena Giurgola
Sándor Juhász
János Kertész
Júlia Koltai
Balázs Lengyel
Thomas Louf
Jean-Philippe Magué
Yelena Mejova
André Panisson
Nicola Perra
Jose Ramasco
David Sanchez
Carlos Sarraute
Vedran Sekara
Michele Tizzoni

karsaim@ceu.edu
@MartonKarsai
www.martonkarsai.com

Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, M. Karsai, Socioeconomic correlations and stratification in social-communication networks. *J. R. Soc. Interface* 13 125 (2016)

Y. Leo, M. Karsai, C. Sarraute, E. Fleury, Correlations of consumption patterns in social-economic networks. *IEEE/ACM ASONAM* p. 500-507, San Francisco CA, August 18-21 (2016)

Y. Leo, M. Karsai, C. Sarraute, E. Fleury, Correlations and dynamics of consumption patterns in social-economic networks. *Soc. Netw. Analys. Min.* 8, 9 (2018)

J.L. Abitbol, M. Karsai, J-P. Magué, J-P. Chevrot, E. Fleury, Socioeconomic dependencies of linguistic patterns in Twitter: a multivariate analysis. *WWW'18 IW3C2 International World Wide Web Conference* (2018).

J.L. Abitbol, M. Karsai, E. Fleury, Location, Occupation, and Semantics based Socioeconomic Status Inference on Twitter. *IEEE ICDM–IWSC* (2018).

J.L. Abitbol, E. Fleury, M. Karsai, Optimal Proxy Selection for Socioeconomic Status Inference on Twitter. *Complexity*, 6059673 (2019)

J. L. Abitbol, M. Karsai, Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2, 684–692 (2020)

E. Bokányi, S. Juhász, M. Karsai, B. Lengyel, Universal patterns of long-distance commuting and social assortativity in cities. *Scientific Reports* 11, 20829 (2021)

R. Millanida Hilman, G. Iñiguez, M. Karsai, Socioeconomic biases in urban mixing patterns of US metropolitan areas. *EPJ Data Science* 11, 32 (2022)

M. Tizzoni, et al., Addressing the socioeconomic divide in computational modeling for infectious diseases. *Nature Communications* 13 2897 (2022)

S. Giurgola, et al., Mapping urban socioeconomic inequalities in developing countries through facebook advertising data. *Frontiers in big Data 5:1006352 (2022)*

L. Espín-Noboa, J. Kertész, M. Karsai, Interpreting Wealth Distribution via Poverty Map Inference using Multimodal Data. In Proceedings of the ACM Web Conference 2023 WWW '23, April 30–May 04, 2023, Austin, TX, USA.

Napoli, L., Sekara, V., García-Herranz, M., & Karsai, M. Socioeconomic reorganization of communication and mobility networks in response to external shocks. *Proceedings of the National Academy of Sciences*, *120*(50), e2305285120 (2023).

T Louf, JJ Ramasco, D Sánchez, M Karsai When dialects collide: how socioeconomic mixing affects language use arXiv:2307.10016