# Aggregating geocoded register data without losing geographical detail

Bo Malmberg

Dept Human Geography

Stockholm University

# Geocoded register data in Sweden

- Population registers since the reformation
  - Kept by parish priests into the 20th century
  - Then by the Swedish Tax Agency
  - Individuals are registered on real estate properties
- Starting in the 1960s real estate properties were geocoded
  - A proposal from Torsten Hägerstrand in the late 1950s
- Geocoded individual level register data available to researchers from the 1990s
  - 100 meter squares (alternatively 250 m squares in built-areas, otherwise 1000 m squares)
  - Researchers can use individual level data but are not allowed to identify individuals

# Challenges

- Computer performance in the early 2000s not good enough for handling the entire population (9 million individuals) in a convenient way
- What can one do with a coordinate?
  - Distances can be computed
  - But an individual location as such is not very interesting
  - As geographers we are more interested what a location stands for
- Geocodes are independent of administrative  boundaries (Hägerstrands's argument)
  - But how can one make good use of the coordinates?
  - Some form of aggregation is needed
  - But how to aggregate without loosing geographical detail?

# Solution

- Individualized neighborhoods
  - Other names: Egohoods, bespoke neighborhoods
- How to do it:
  1. Expand a buffer around a specific location until the buffer encompasses the k-nearest neighbors
  2. Compute aggregate statistics  for the population contained in the buffer
  3. Repeat for all locations.
- By choosing a large enough k, it is possible to ensure that privacy is maintained
- Still, it will be possible to provide detailed geographical information: "If you are at this exact location, 25% of the 200 nearest neighbors have a tertiary location".

# Advantages of individualized neighborhoods

- Better measure of geographical context than measures based on fixed geographical subdivisions
- Measures of segregation that are not influenced by boundaries of geographical subdivisions
  - Facilitates comparative studies. The same definition of neighborhoods can be implemented in different context.
- Show geographical context and segregation are scale dependent.
- Nice maps

# Limitations of individualized neighborhoods

- Buffers use Euclidian distances, barriers are not considered
- Neighborhood will be small in high density areas, large in low density areas
    - Though necessary to ensure privacy

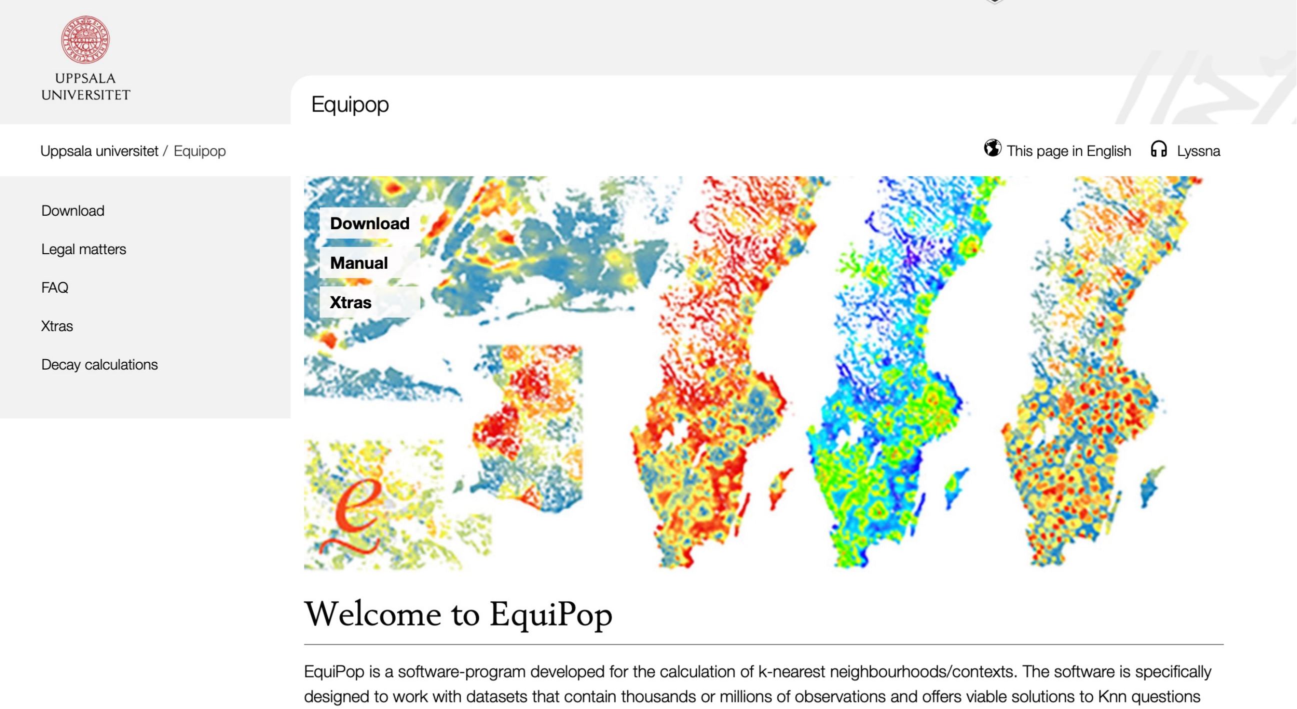
# How to do it?

- EquiPop software by John Östh, free proprietary

https://equipop.kultgeog.uu.se

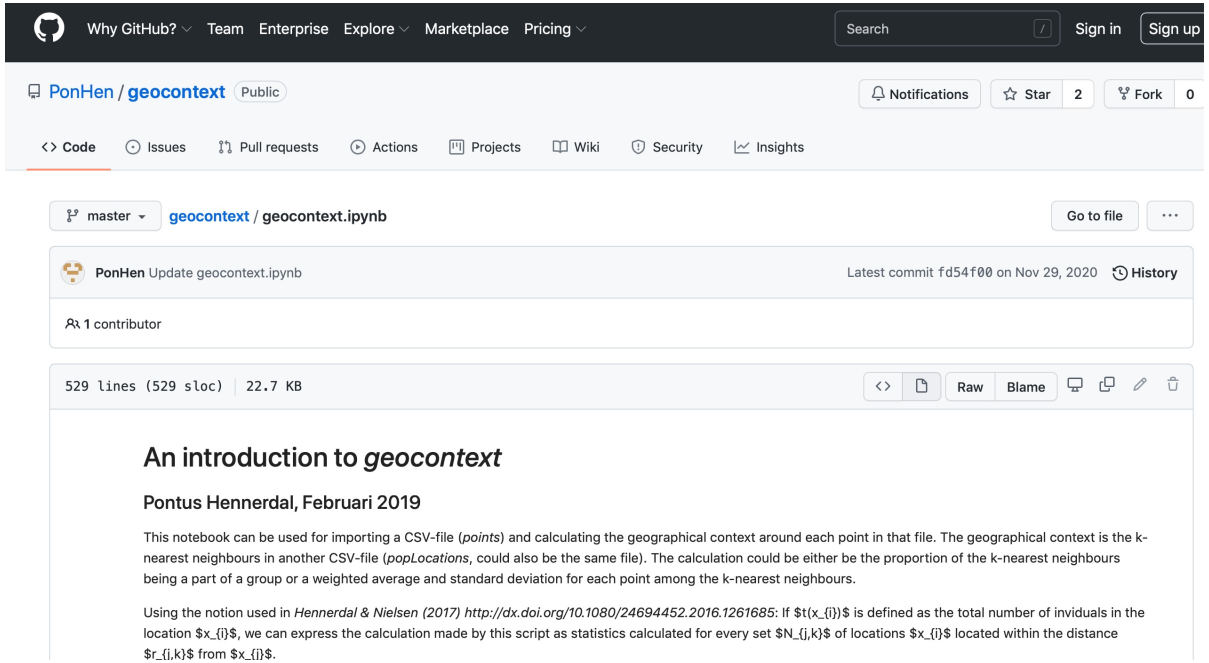
- Geocontext by Pontus Hennerdal, open source, Python script

https://github.com/PonHen/geocontext



Welcome to EquiPop

EquiPop is a software-program developed for the calculation of k-nearest neighbourhoods/contexts. The software is specifically designed to work with datasets that contain thousands or millions of observations and offers viable solutions to Knn questions



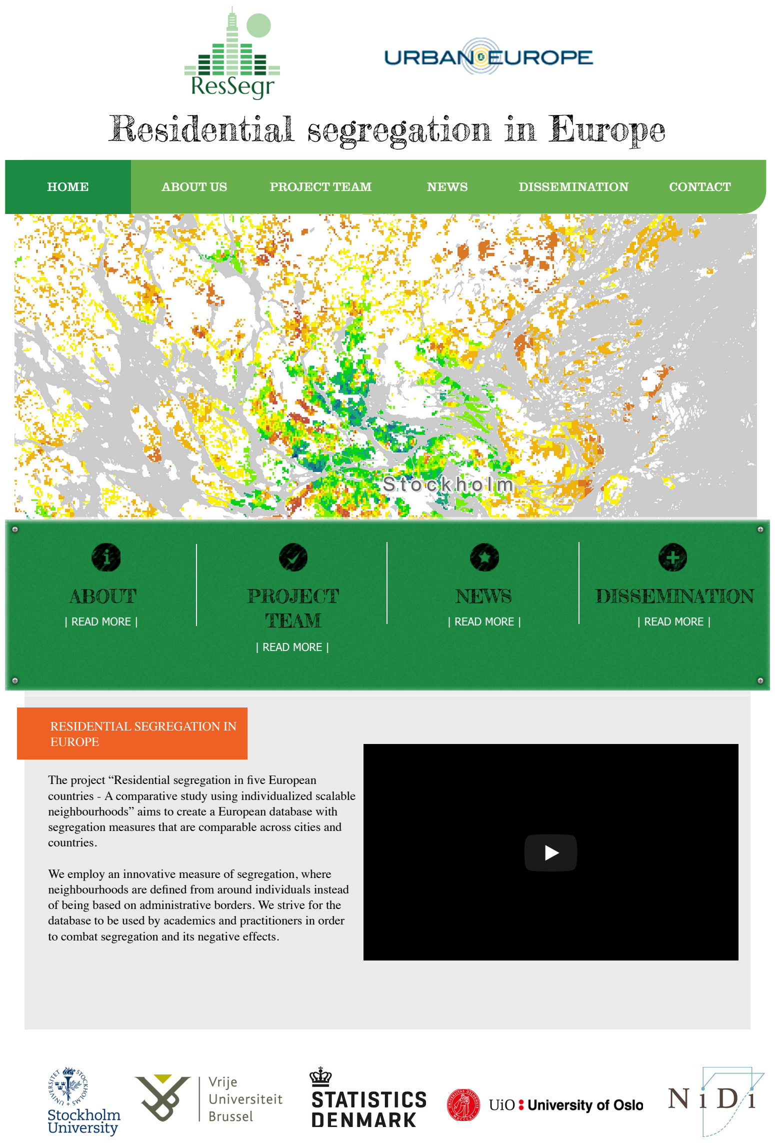
An introduction to *geocontext*

Pontus Hennerdal, Februari 2019

This notebook can be used for importing a CSV-file (*points*) and calculating the geographical context around each point in that file. The geographical context is the k-nearest neighbours in another CSV-file (*popLocations*, could also be the same file). The calculation could be either be the proportion of the k-nearest neighbours being a part of a group or a weighted average and standard deviation for each point among the k-nearest neighbours.

Using the notion used in *Hennerdal & Nielsen (2017) http://dx.doi.org/10.1080/24694452.2016.1261685*: If $t(x_{i})$ is defined as the total number of inviduals in the location $x_{i}$, we can express the calculation made by this script as statistics calculated for every set $N_{j,k}$ of locations $x_{i}$ located within the distance $r_{j,k}$ from $x_{i}$.

# Examples

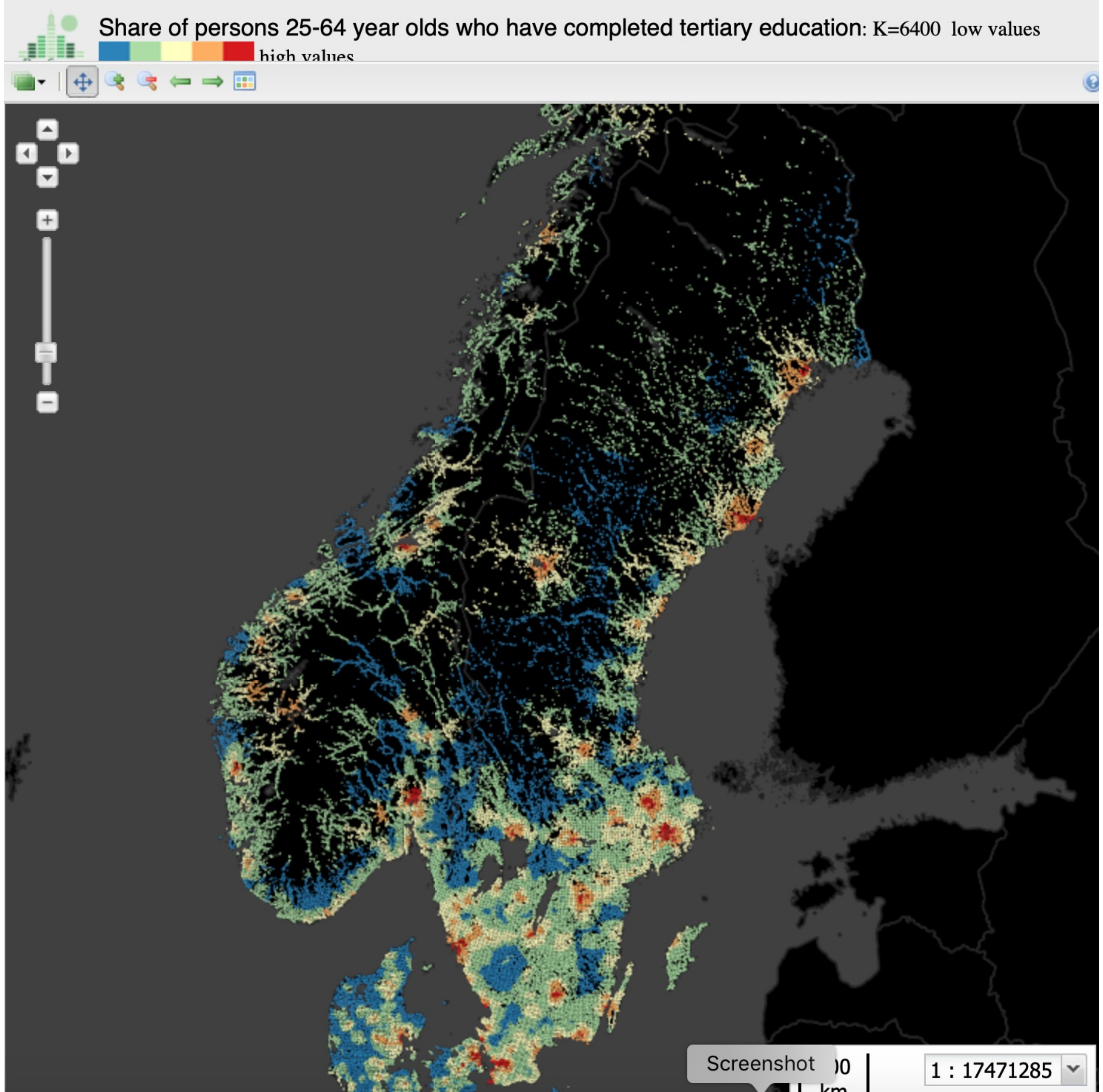


# Residential segregation in Europe

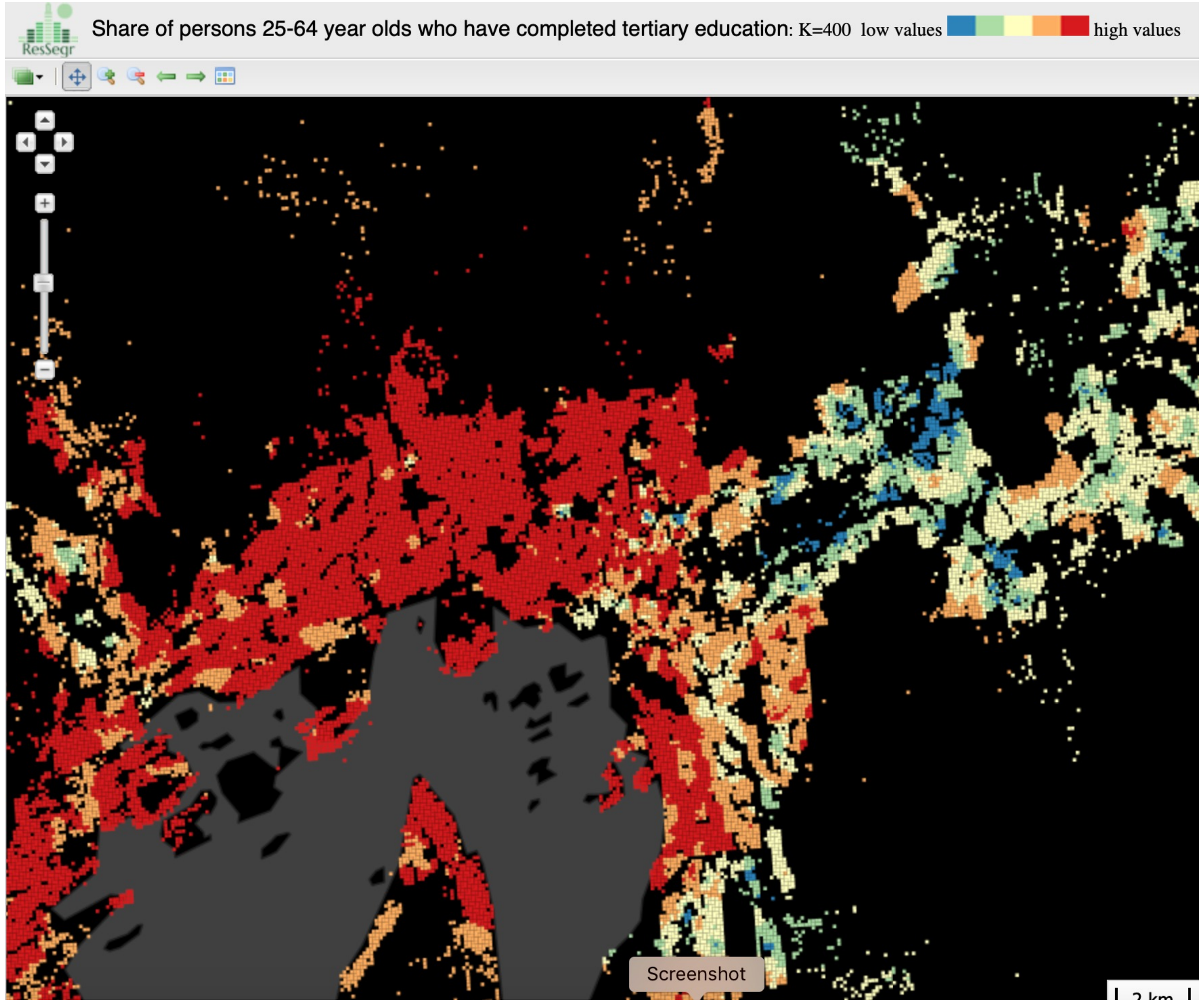
RESIDENTIAL SEGREGATION IN EUROPE

The project “Residential segregation in five European countries - A comparative study using individualized scalable neighbourhoods” aims to create a European database with segregation measures that are comparable across cities and countries.

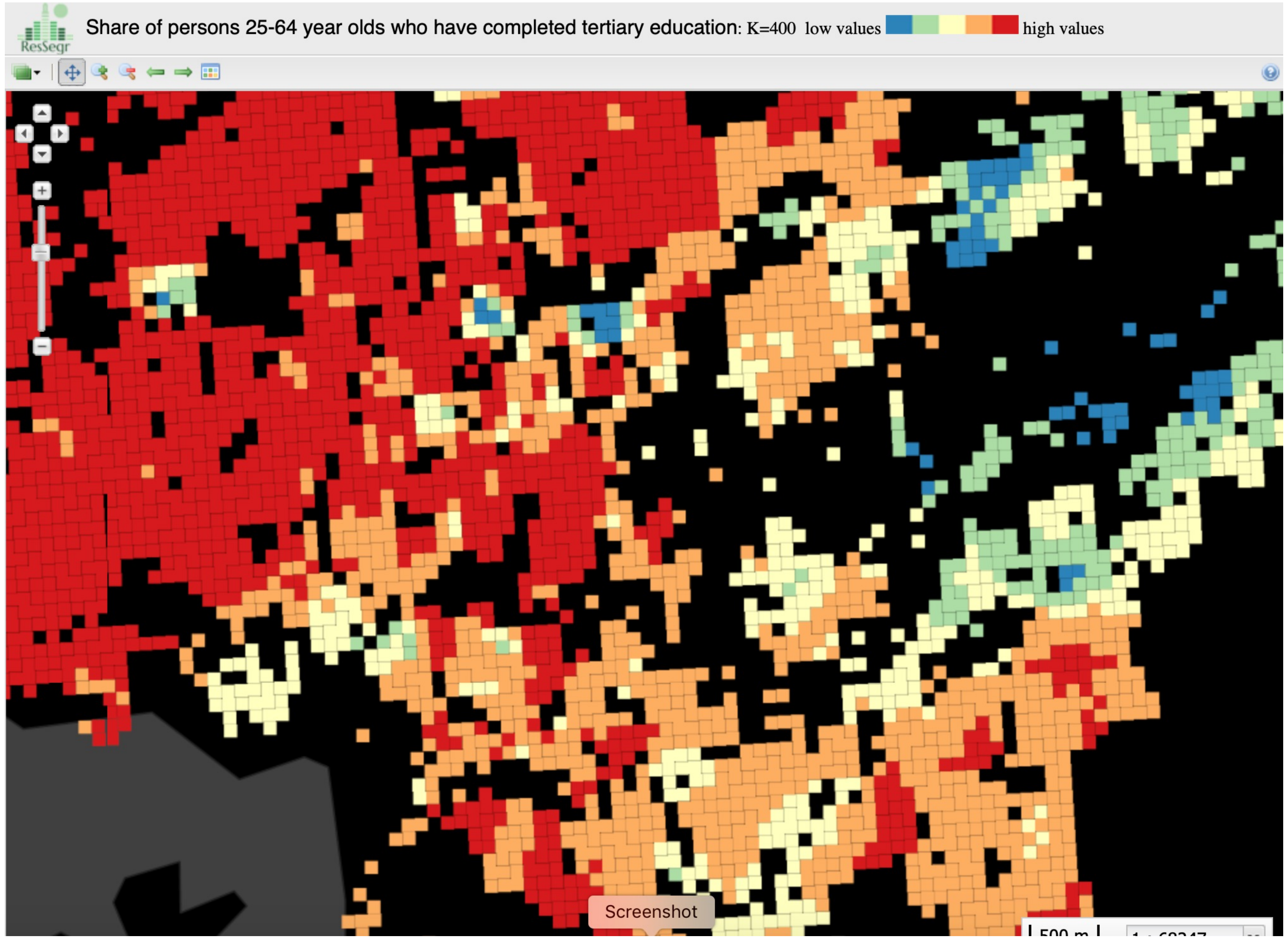
We employ an innovative measure of segregation, where neighbourhoods are defined from around individuals instead of being based on administrative borders. We strive for the database to be used by academics and practitioners in order to combat segregation and its negative effects.



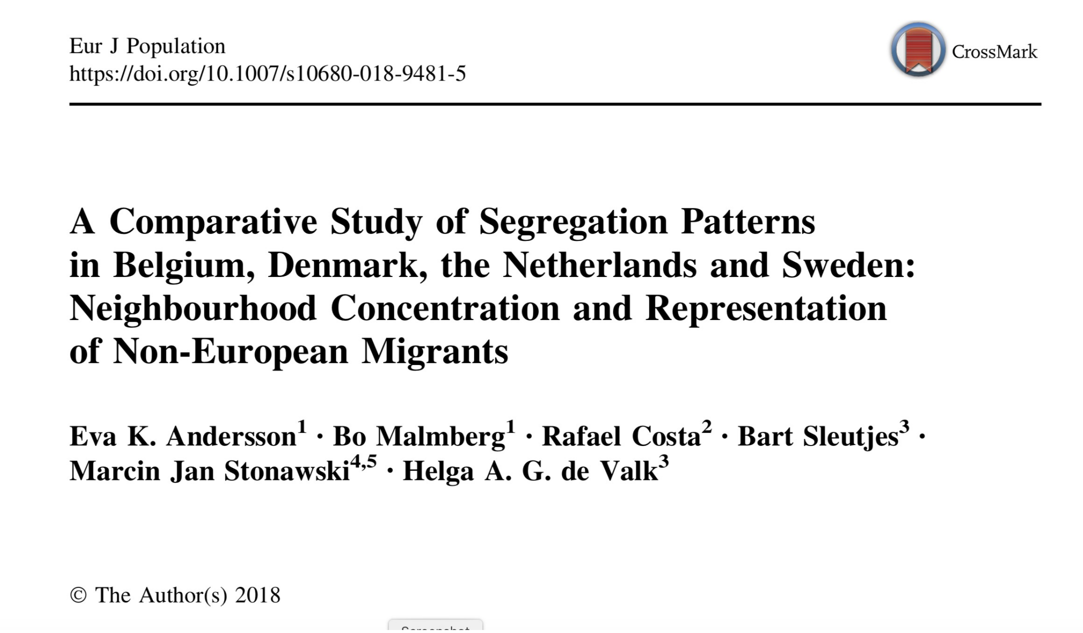
Share of persons 25-64 year olds who have completed tertiary education: K=6400 low values 🟦 🟩 🟨 🟧 🟥 high values



Share of persons 25-64 year olds who have completed tertiary education: K=400 low values 🟦🟩🟨🟧🟥 high values



Share of persons 25-64 year olds who have completed tertiary education: K=400 low values 🟦🟩🟨🟧🟥 high values



Eur J Population
https://doi.org/10.1007/s10680-018-9481-5

# A Comparative Study of Segregation Patterns in Belgium, Denmark, the Netherlands and Sweden: Neighbourhood Concentration and Representation of Non-European Migrants

Eva K. Andersson[1] · Bo Malmberg[1] · Rafael Costa[2] · Bart Sleutjes[3] · Marcin Jan Stonawski[4,5] · Helga A. G. de Valk[3]

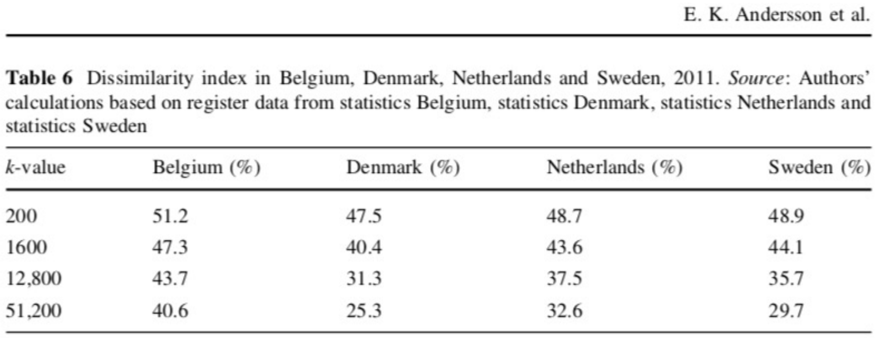
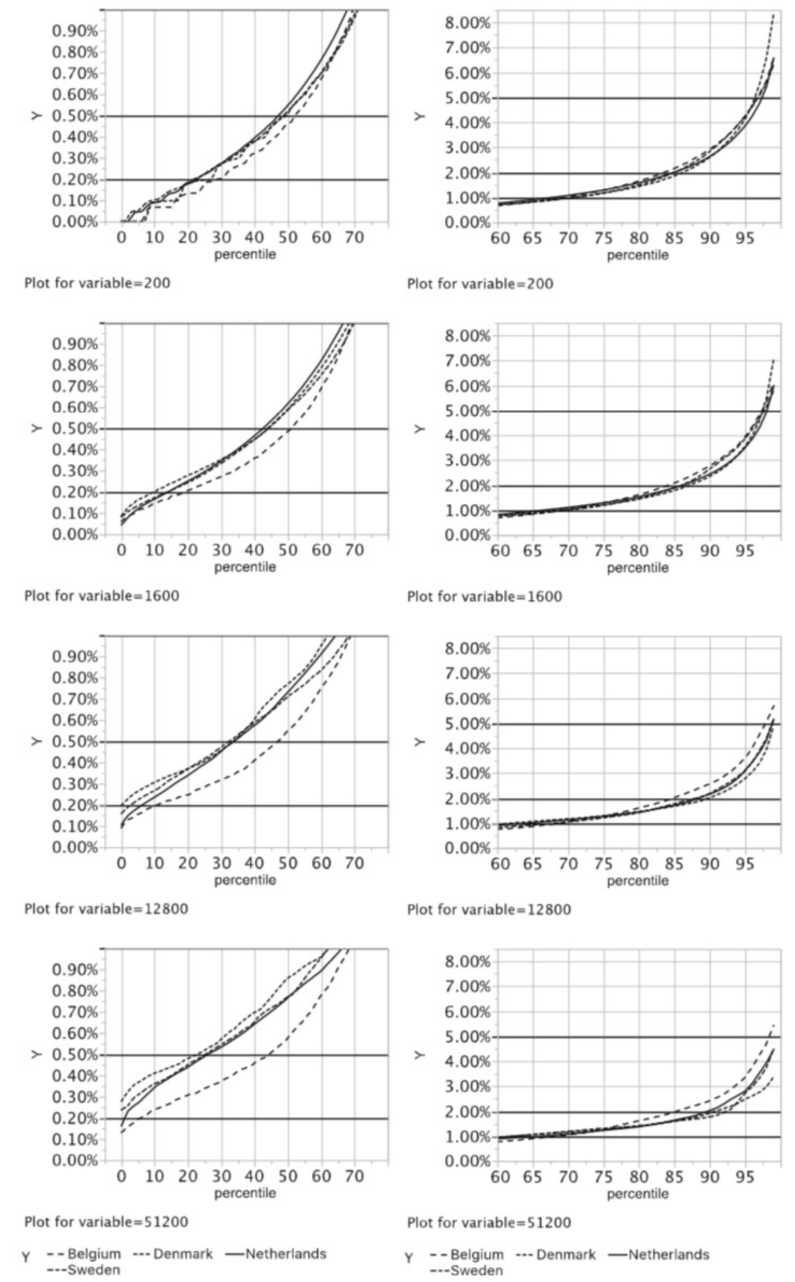


**Table 6** Dissimilarity index in Belgium, Denmark, Netherlands and Sweden, 2011. *Source*: Authors' calculations based on register data from statistics Belgium, statistics Denmark, statistics Netherlands and statistics Sweden

| *k*-value | Belgium (%) | Denmark (%) | Netherlands (%) | Sweden (%) |
|---|---|---|---|---|
| 200 | 51.2 | 47.5 | 48.7 | 48.9 |
| 1600 | 47.3 | 40.4 | 43.6 | 44.1 |
| 12,800 | 43.7 | 31.3 | 37.5 | 35.7 |
| 51,200 | 40.6 | 25.3 | 32.6 | 29.7 |



Plot for variable=200

Plot for variable=200

Plot for variable=1600

Plot for variable=1600

Plot for variable=12800

Plot for variable=12800

Plot for variable=51200

Plot for variable=51200

Article

Urban Studies

*Urban Studies*
2015, Vol. 52(12) 2117–2133
© Urban Studies Journal Limited 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0042098014542487
usj.sagepub.com
SAGE

# Contextual effects on educational attainment in individualised, scalable neighbourhoods: Differences across gender and social class

**Eva K Andersson**
Stockholm University, Sweden

**Bo Malmberg**
Stockholm University, Sweden

## Abstract

This paper analyses whether a multi-scale representation of geographical context based on statistical aggregates computed for individualised neighbourhoods can lead to improved estimates of neighbourhood effect. Our study group consists of individuals born in 1980 that have lived in Sweden since 1995 and we analyse the effect of neighbourhood context at age 15 on educational outcome at age 30 controlling for parental background. A new piece of software, Equipop, was used to compute the socio-economic composition of neighbourhoods centred on individual residential locations and ranging in scale from including the nearest 12 to the nearest 25,600 neighbours. Our results indicate that context measures based on fixed geographical sub-divisions can lead to an underestimation of neighbourhood effects. A multi-scalar representation of geographical context also makes it easier to estimate how neighbourhood effects vary across different demographic groups. This indicates that scale-sensitive measures of geographical context could help to re-invigorate the neighbourhood effects literature.

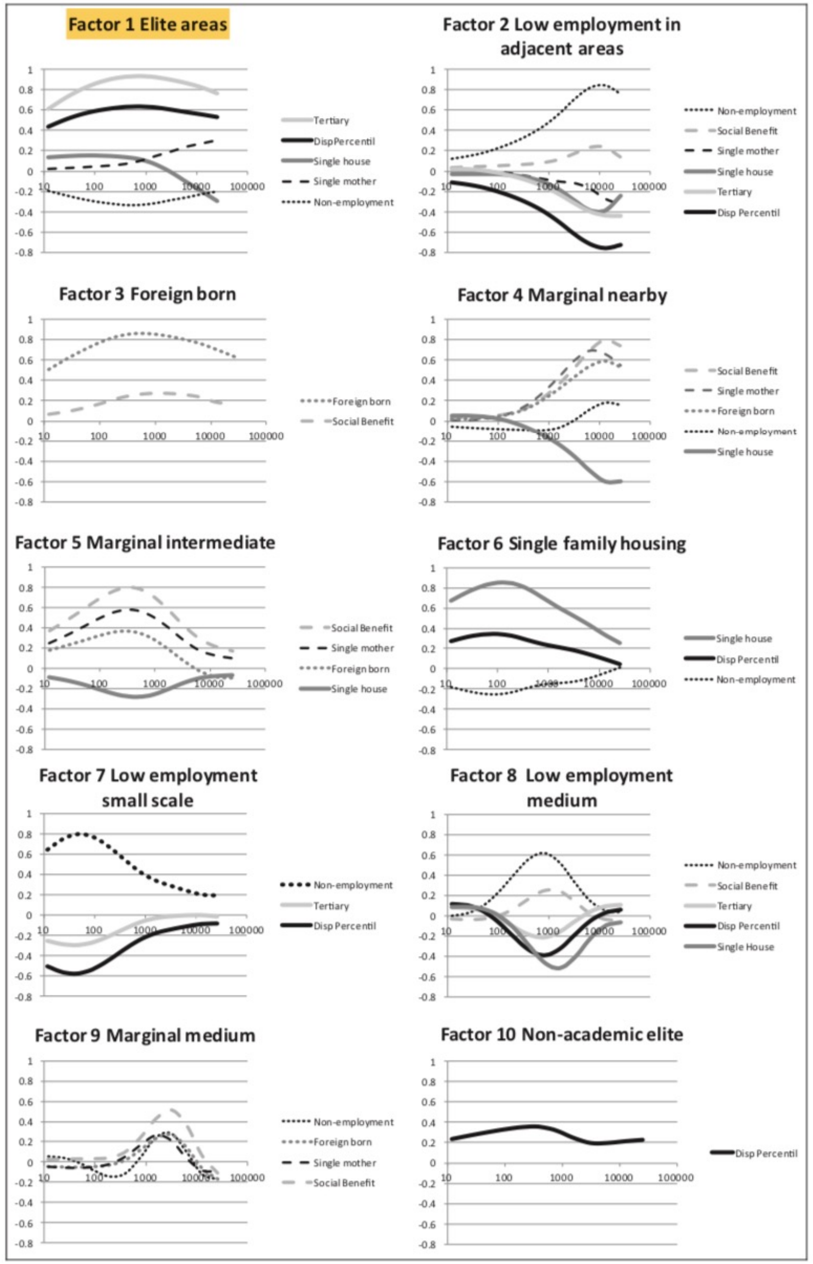
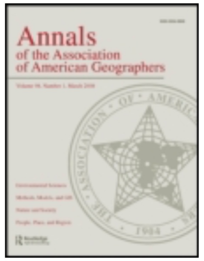


**Figure 1.** Factors and loadings. (To reduce clutter, these graphs only show factors that for at least one *k*-le than 0.2 or lower than −0.2.)



**Annals of the Association of American Geographers**

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/raag20

# Composite Geographical Context and School Choice Attitudes in Sweden: A Study Based on Individually Defined, Scalable Neighborhoods

Bo Malmberg[a], Eva K. Andersson[a] & Zara Bergsten[b]

[a] Department of Human Geography, Stockholm University,

[b] Institute for Housing and Urban Research, Uppsala University

Published online: 04 Jun 2014.

CrossMark

Click for updates

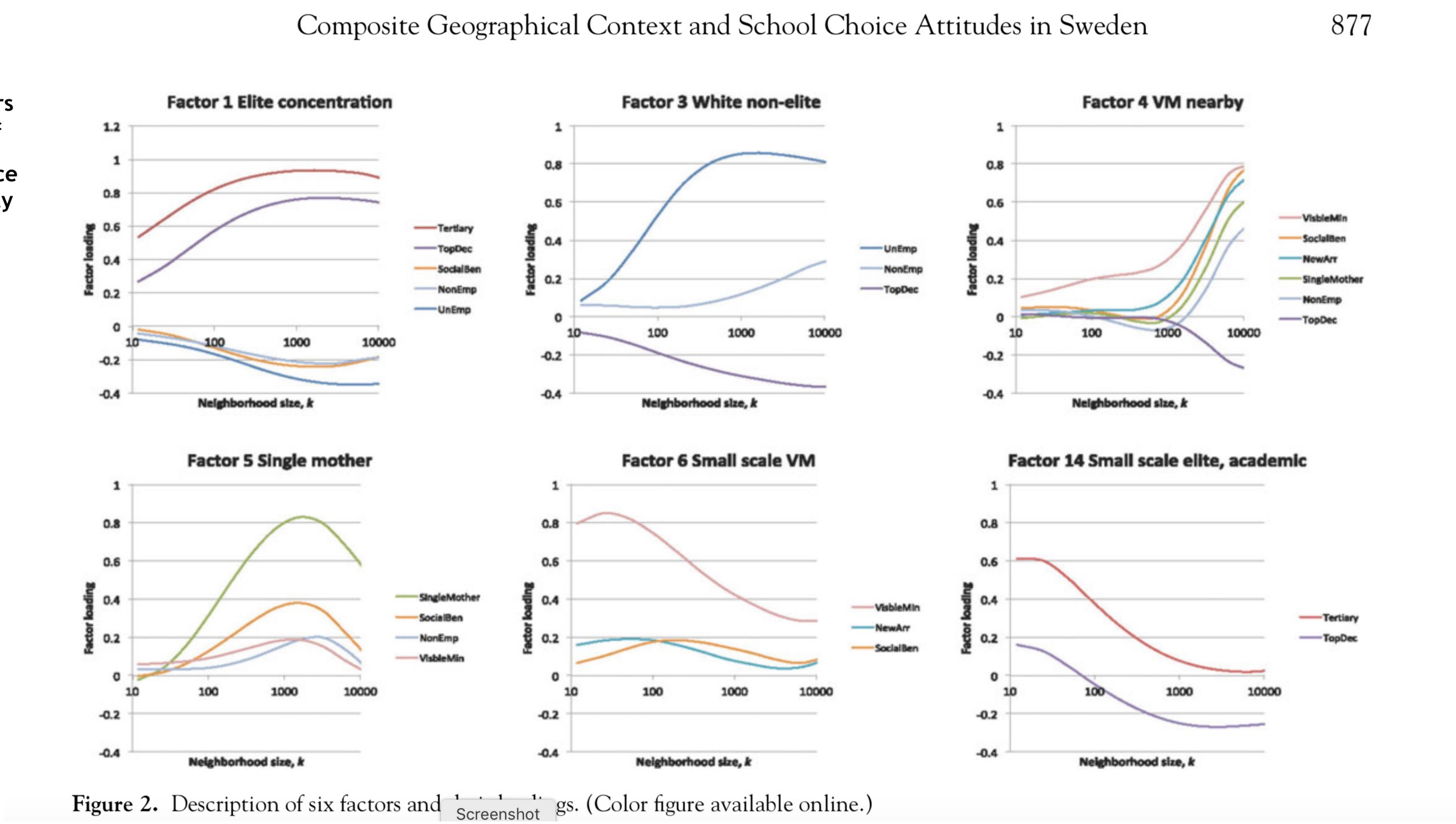
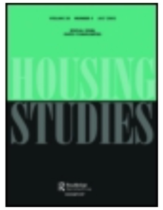


**Figure 2.** Description of six factors and [illegible]gs. (Color figure available online.)



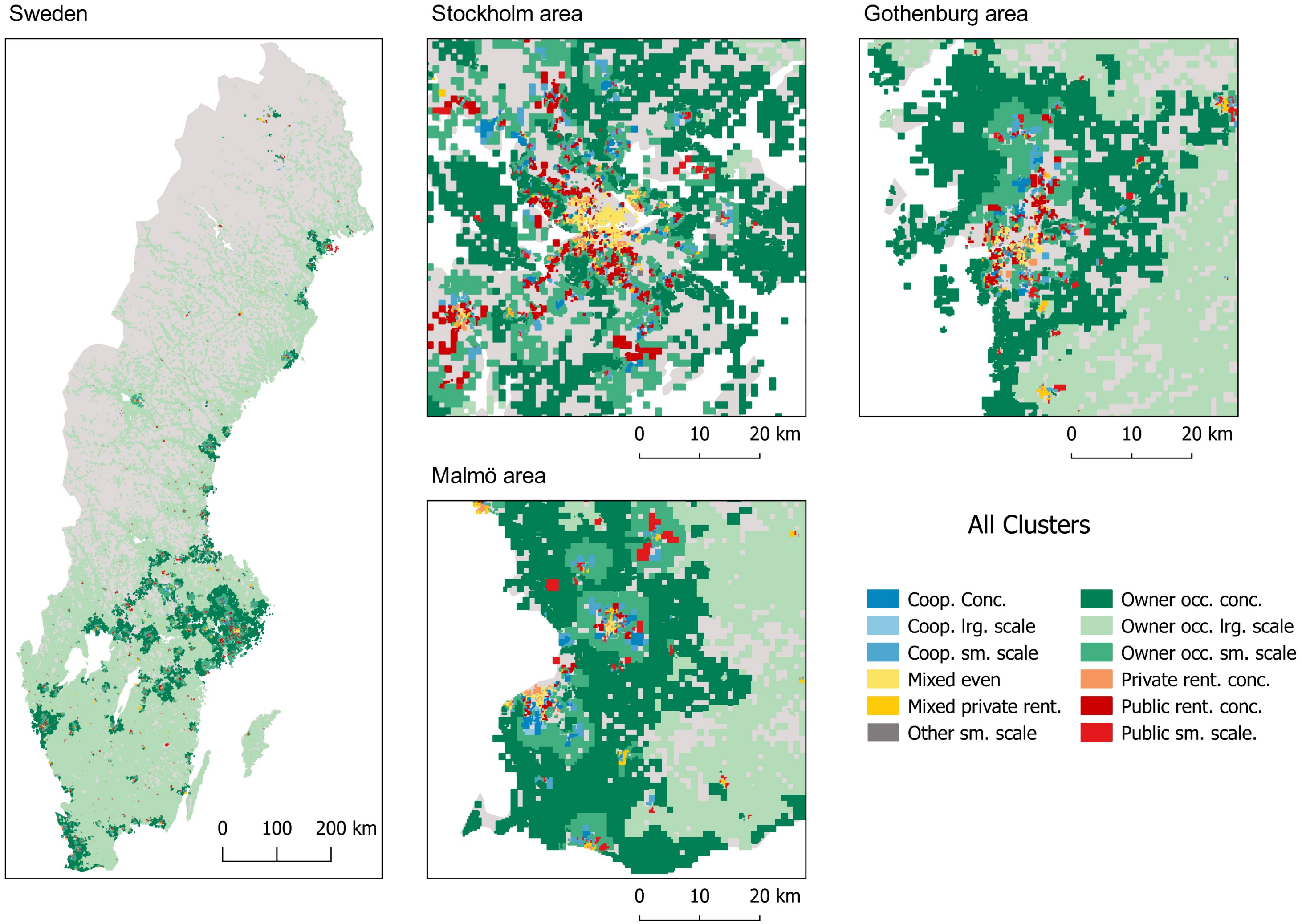
Routledge
Taylor & Francis Group

Housing Studies

ISSN: 0267-3037 (Print) 1466-1810 (Online) Journal homepage: https://www.tandfonline.com/loi/chos20

# Tenure type landscapes and housing market change: a geographical perspective on neo-liberalization in Sweden

Thomas Wimark, Eva K. Andersson & Bo Malmberg



Sweden

Stockholm area

Gothenburg area

Malmö area

All Clusters

Urban Geography

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rurb20

# Socio-economic segregation in European cities. A comparative study of Brussels, Copenhagen, Amsterdam, Oslo and Stockholm

Karen Haandrikman, Rafael Costa, Bo Malmberg, Adrian Farner Rogne & Bart Sleutjes

2021

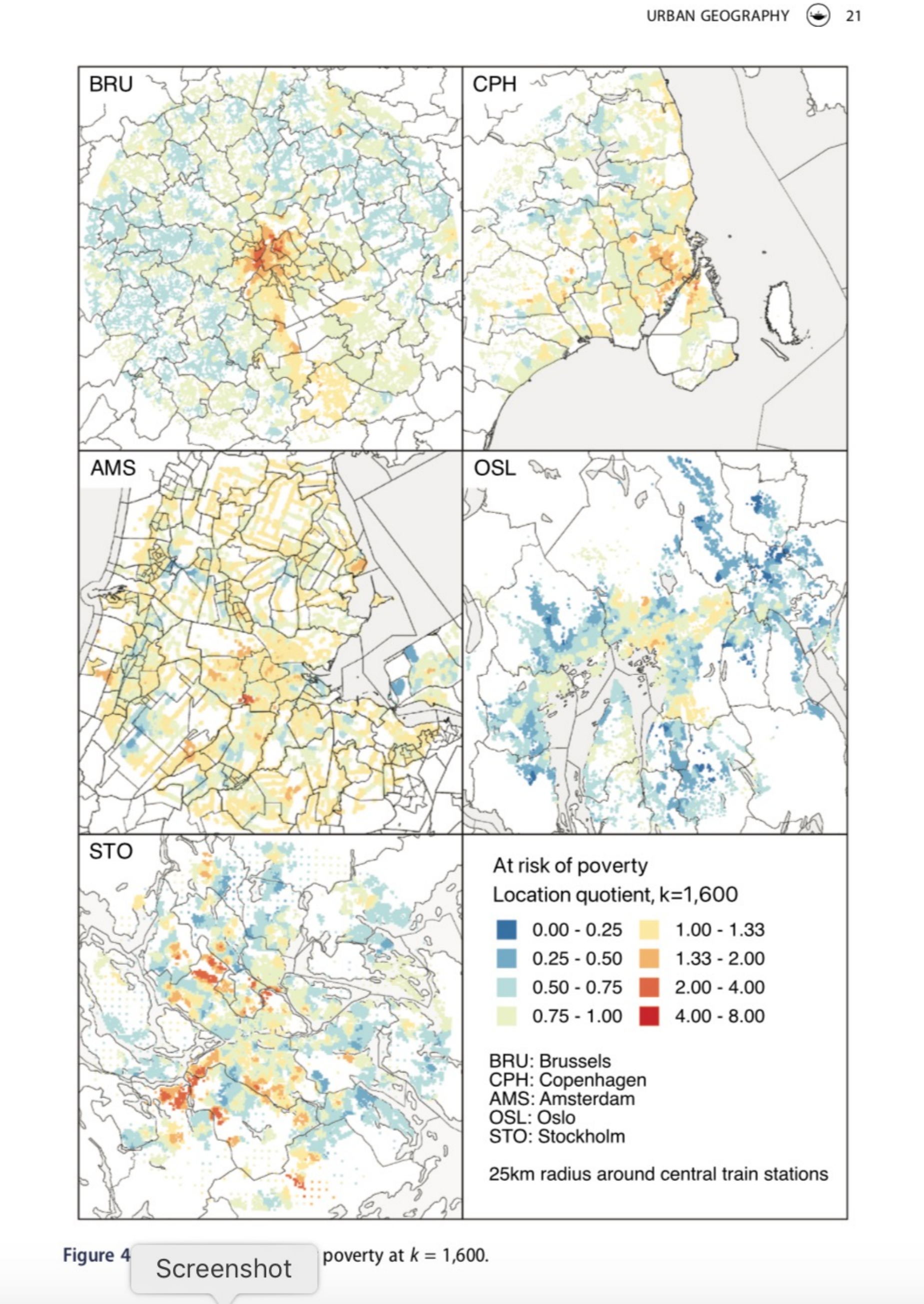

URBAN GEOGRAPHY 21

Figure 4 poverty at $k$ = 1,600.

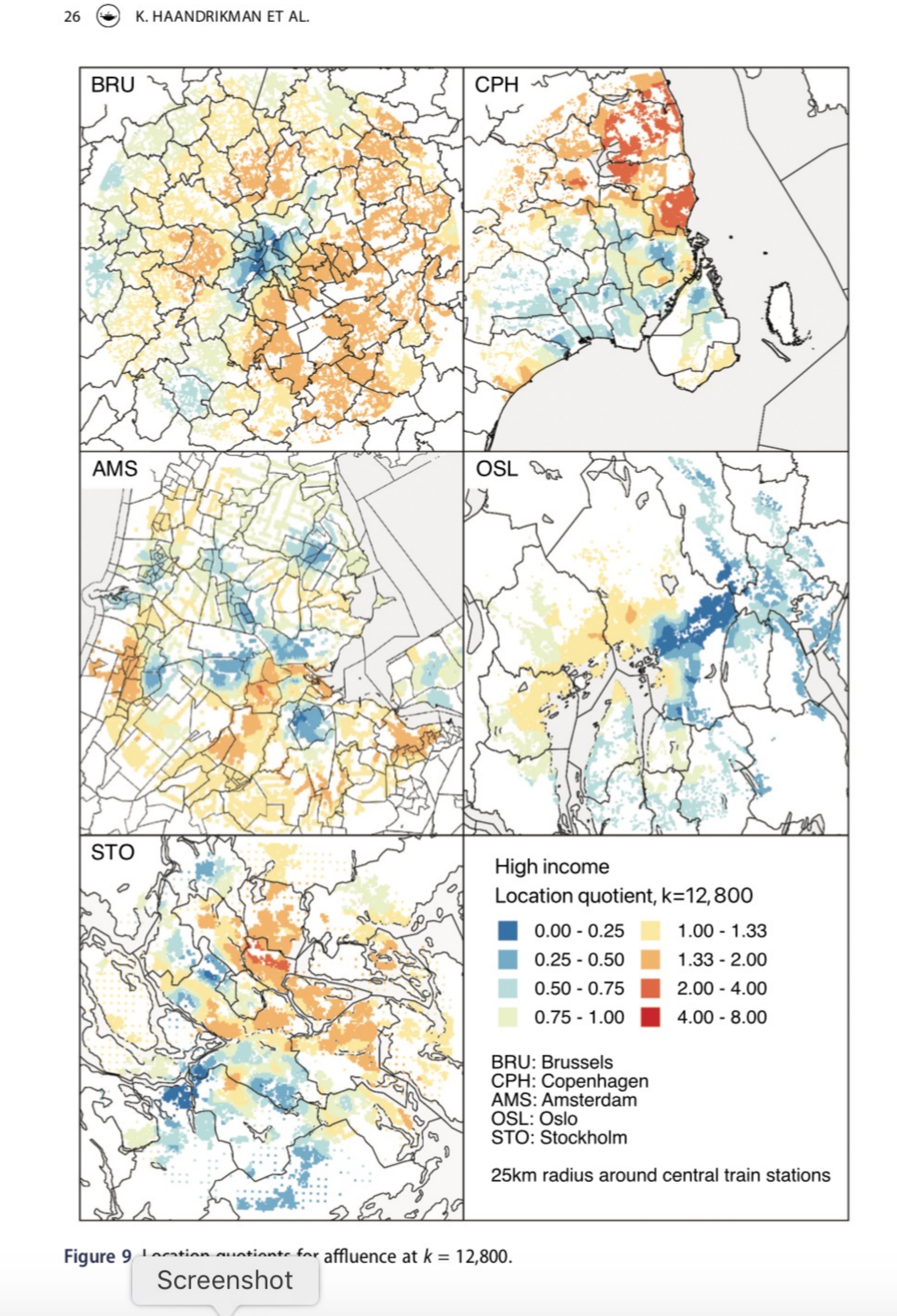

26 K. HAANDRIKMAN ET AL.

Figure 9. Location quotients for affluence at $k$ = 12,800.

# Summary

- The individualized neighborhood approach is a great tool for geographical analysis
  - New patterns can be explored and discovered
  - At different scale levels

# Geoprivacy

- Our argument: *If geocoded individual level register data has been aggregated using individualized neighborhoods, data with high levels of geographical data can be shared*
  - Has been accepted by Statistics Sweden, Statistics Denmark, and Statistics Norway
  - Statistics Netherlands more cautious
- Would be good to have an evaluation of this argument by fellow academics
- Could help data sharing

# Thank you!