Migration studies with a Compositional Data approach: a case study of population structure in the Capital Region of Denmark

Javier Elío, Marina Georgati, Henning S. Hansen, Carsten Keßler

AALBORG UNIVERSITY
DENMARK

Twitter: @Elio_Javi
Email: javierdem@plan.aau.dk
Website: https://javierelio.netlify.com

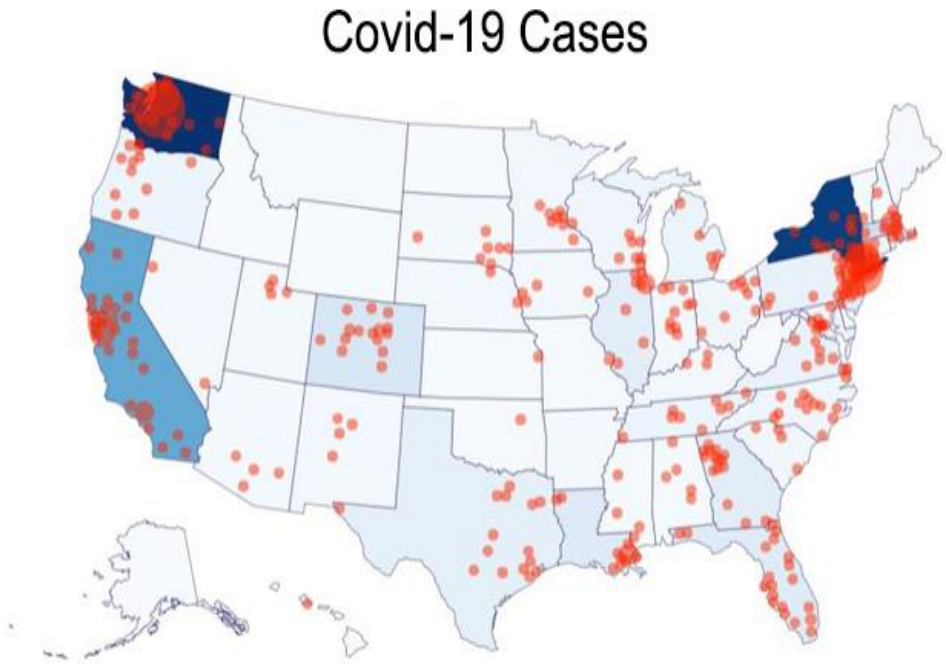
Introduction

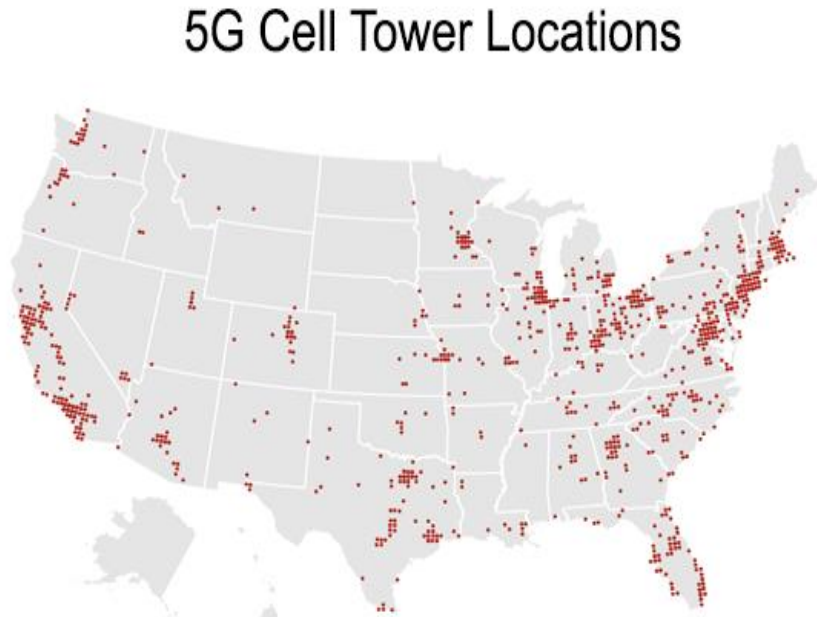
Why do we need CoDa in population geography?

- **Reason:** It is often more interesting to analyse proportions than counts, e.g., the percentage of people with X characteristics in a region rather than the number of people with those characteristics. Proportions do not depend on the total population of the region.

- **Drawback:** Data are constrained to a constant sum (e.g., 1 for proportions or 100 for percentages) and, therefore, they are not independent. If one increase, another one decreases to retain the sum.

- **Consequences:** Standard statistical procedures may lead to spurious correlations, predictions outside the range, or have problem with sub-compositional coherence.





Source: Julian Hazell (link) - ABC News & Scrapehero

Introduction

What is CoDa?

- **Compositional data** *"consist of vectors whose components are the proportion or percentages of some whole"* (J. Aitchison)*:*

$$x = [x_1, x_2, \dots, x_D]; \quad x_i > 0; \quad \sum_{i=1}^{D} x_i = C$$

- **Examples:**
  - Percentage of rented/owner households in a region
  - Population structure by:
    - Age (e.g. percentage of young, working age and elderly population)
    - Religion (e.g. percentage of Christians, Muslims, Hindus, Jews, other religion, and Unaffiliated)
    - Origin (e.g. percentage of Danes, Non-Western and Western migrants)
  - Time of the day spend at work, on leisure activities, and sleeping (sum up to 24h)

Introduction

Theoretical background

- The **main idea** is to transform the CoDa data in a way that they can be analysed with standard statistical tools, designed for unconstrained data.
- **Three main transformations:**

1. Additive log-ratio transformation: $\mathrm{alr}(\mathrm{x}) = \log(\frac{\mathrm{x}_1}{\mathrm{x}_\mathrm{D}}, \frac{\mathrm{x}_2}{\mathrm{x}_\mathrm{D}}, \ldots, \frac{\mathrm{x}_{\mathrm{D}-1}}{\mathrm{x}_\mathrm{D}})$
2. Centered log-ratio transformation: $\mathrm{clr}(\mathrm{x}) = \log(\frac{\mathrm{x}_1}{\mathrm{g(x)}}, \frac{\mathrm{x}_2}{\mathrm{g(x)}}, \ldots, \frac{\mathrm{x}_\mathrm{D}}{\mathrm{g(x)}})$
3. Isometric log-ratio transformations: $\mathrm{y} = \mathrm{ilr}(\mathrm{x}) = (y_1, y_2, \ldots, y_{D-1}) \in \mathbb{R}^{D-1}$ where $y_i = \frac{1}{\sqrt{\mathrm{i}\,(\mathrm{i}+1)}} \ln\left(\frac{\prod_{\mathrm{j}=1}^{\mathrm{i}} \mathrm{x_j}}{(\mathrm{x_i}+1)^{\mathrm{i}}}\right)$ for $\mathrm{i} = 1, \ldots, \mathrm{D}-1$

- There is not a best transformation and **all of them have their strengths and limitations**.

Introduction

Balances

- Special case of ilr transformation based on Sequential Binary Partitions
- SBP - vectors of n parts are partitioned into groups of parts presenting a certain affinity
- They represent the relationship between two groups of parts allowing interpretation within and between groups of parts

Six-part component

x_1	x_2	x_3	x_4	x_5	x_6
1	1	1	1	-1	-1
1	1	1	-1	0	0
1	1	-1	0	0	0
1	-1	0	0	0	0
0	0	0	0	1	-1

Parts

r (+)	S (-)
4	2
3	1
2	1
1	1
1	1

Balances

b
1
2
3
4
5

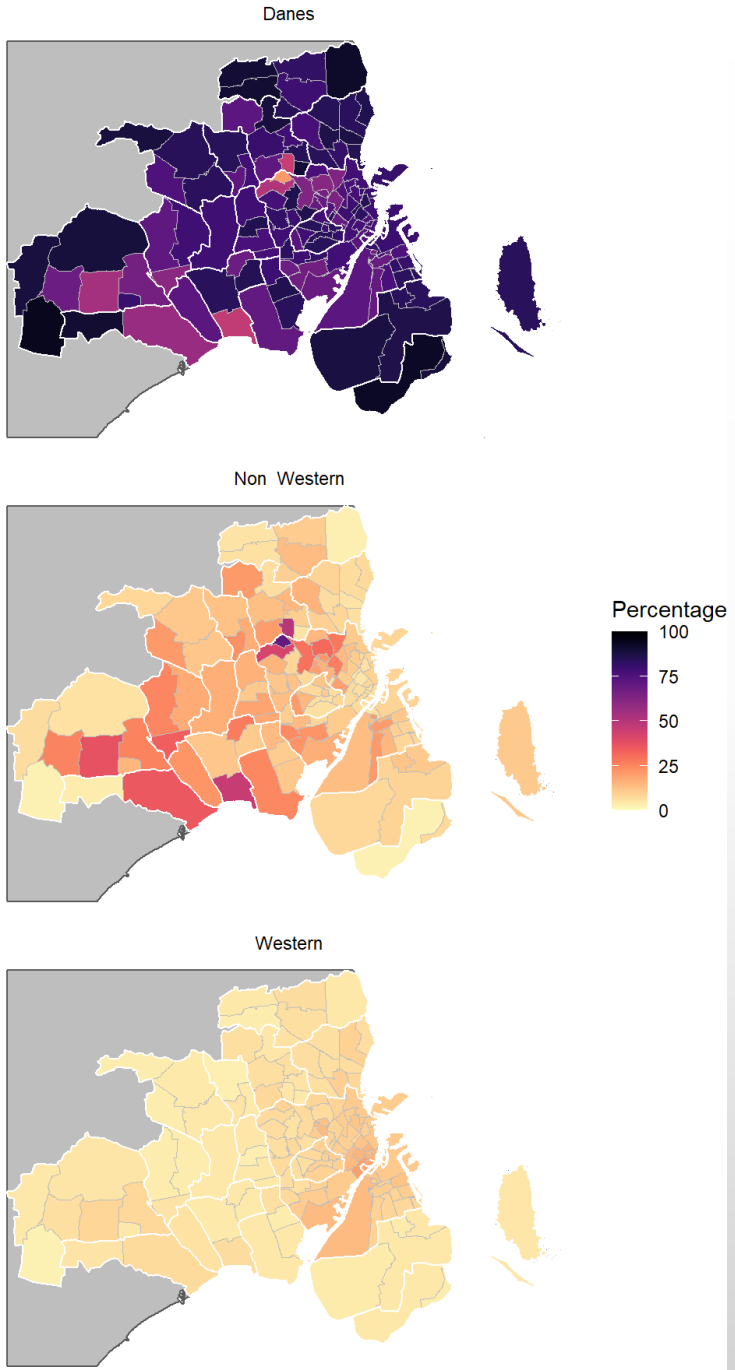
$$b_i = \sqrt{\frac{rs}{r+s}} \ln\left(\frac{(\prod_{+} x_j)^{\frac{1}{r}}}{(\prod_{-} x_k)^{\frac{1}{s}}}\right) \text{ for } i = 1, \ldots, D-1$$

First balance: $b_1 = \sqrt{\frac{8}{6}} \ln\left(\frac{(x_1 x_2 x_3 x_4)^{\frac{1}{4}}}{(x_5 x_6)^{\frac{1}{2}}}\right)$

Case Study

The Capital Region of Denmark

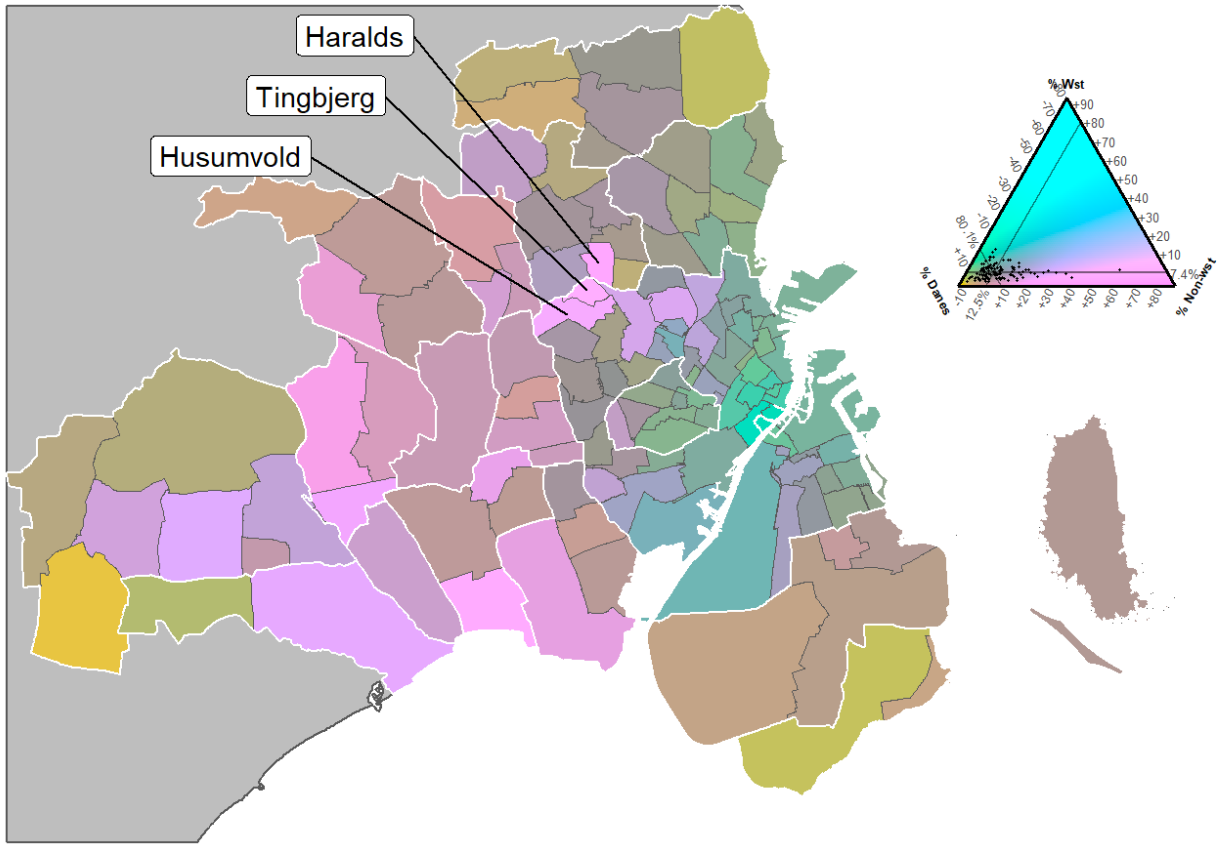
- **Population data at parish level**
  - Data from Denmark Statistics (i.e. table KMSTA001)
  - Population at the first day of the year by ancestry (2020)
  - Three categories:
    1. Persons of Danish origin
    2. Immigrants and descendants from non-western countries
    3. Immigrants and descendants from western countries



AALBORG UNIVERSITY
DENMARK

Case Study

Ternary plot



- Centred over the compositional mean (80.1%, 7.4%, 12.5% of Danes, Western, and non-Western population, respectively)
- We can identify some patterns:
  - ❖ Western migrants prefer parishes close to the city centre
  - ❖ Non-Western migrants tend to settle down in western peripheral parishes (with percentages up to 41.6%, 49.61%, and 69.85% for Husumvold, Haralds and Tingbjerg parishes)

Case Study

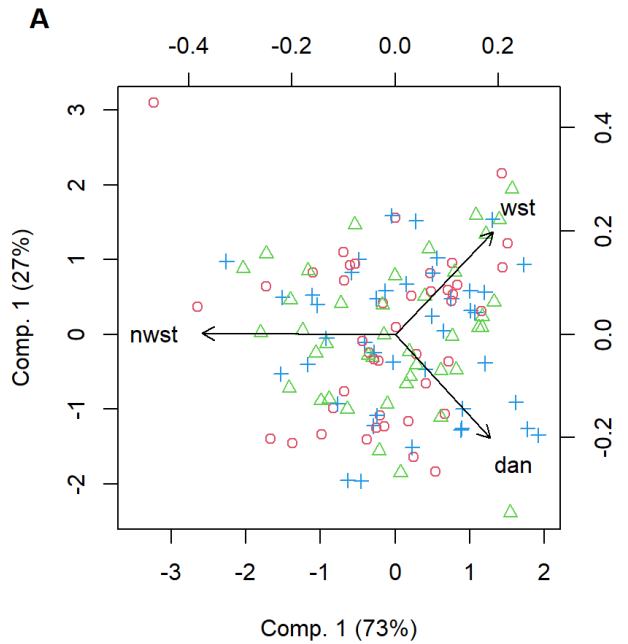
Estimate balances

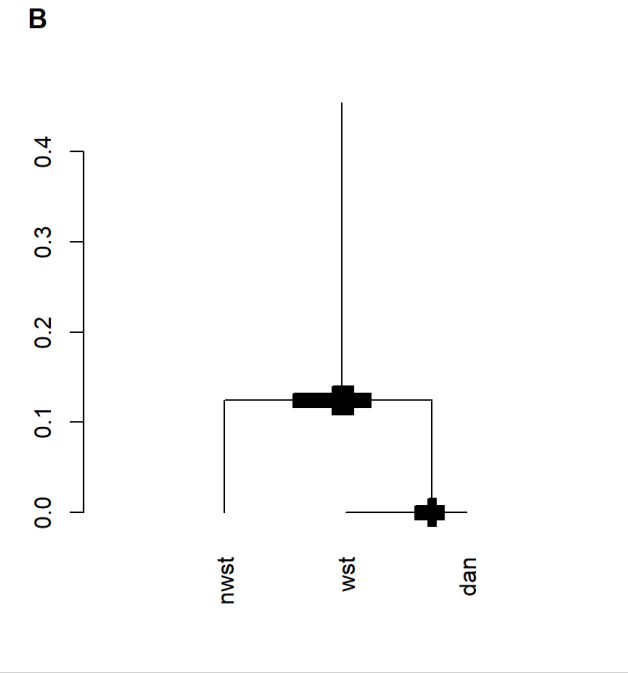
- **Partition scheme:** it is difficult to select the correct partitions for obtaining meaningful interpretations and it should be done base on expert knowledge and/or by compositional biplots (clr transformation).

Danes	Western	Non-Western	r	s	Balance
1	1	-1	2	1	b_1
1	-1	0	1	1	b_2

$$b_1 = \sqrt{\frac{2}{3}} \ln\left(\frac{(\text{Danes} \cdot \text{Western})^{0.5}}{\text{NonWester}}\right) \qquad b_2 = \sqrt{\frac{1}{2}} \ln\left(\frac{\text{Danes}}{\text{Western}}\right)$$

- **Analysis:** spatial autocorrelation, hierarchical cluster analysis, and regression analysis (house prices vs. migration)





AALBORG UNIVERSITY
DENMARK

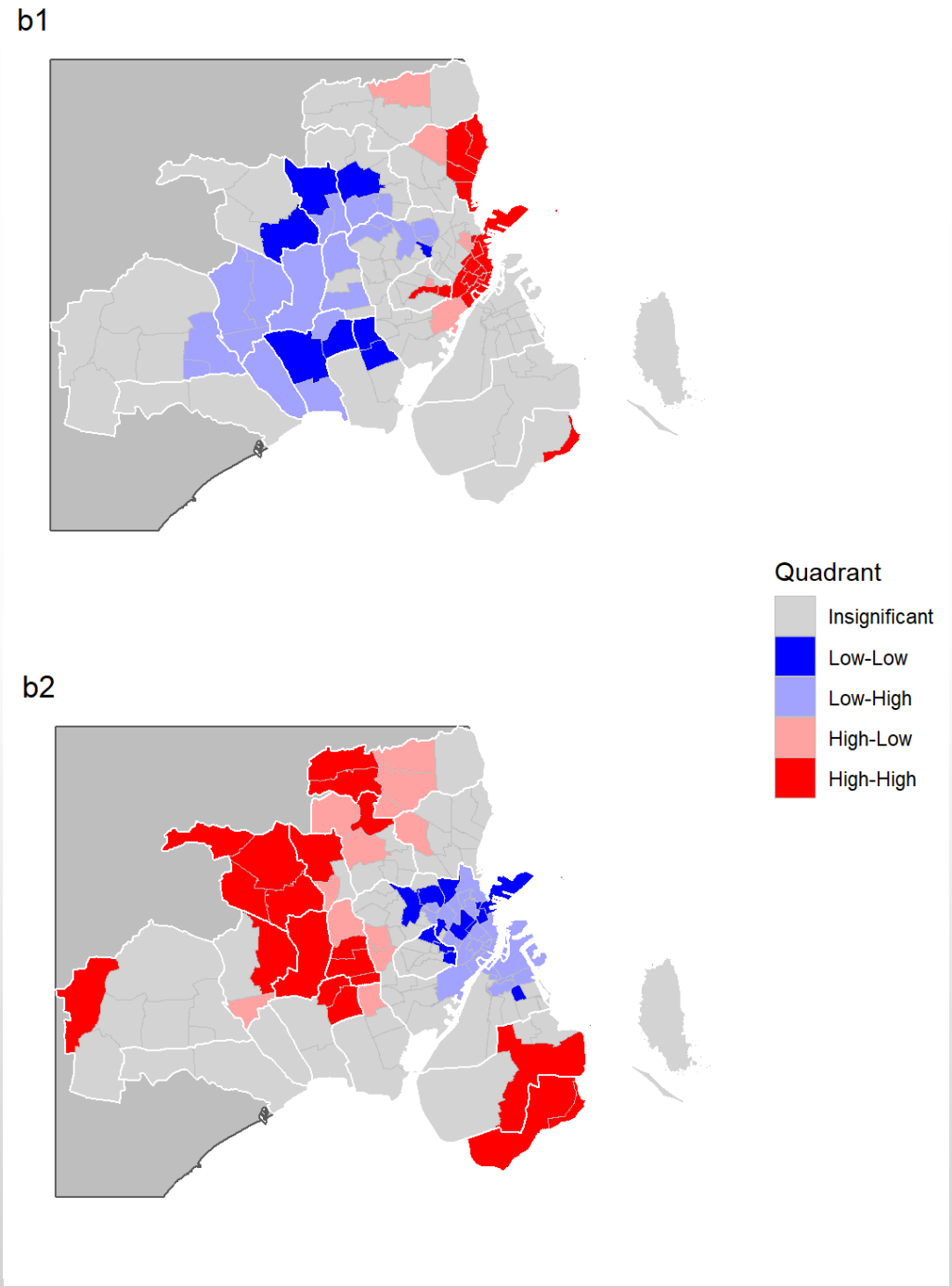
Case Study

Spatial autocorrelation

Moran I test under randomisation (alternative: greater)

Balance	Index	Expectation	Variance	Statistic	P value
b1	0.470	-0.008	0.003	8.563	5.503×10^{-18}
b2	0.542	-0.008	0.003	9.837	3.915×10^{-23}

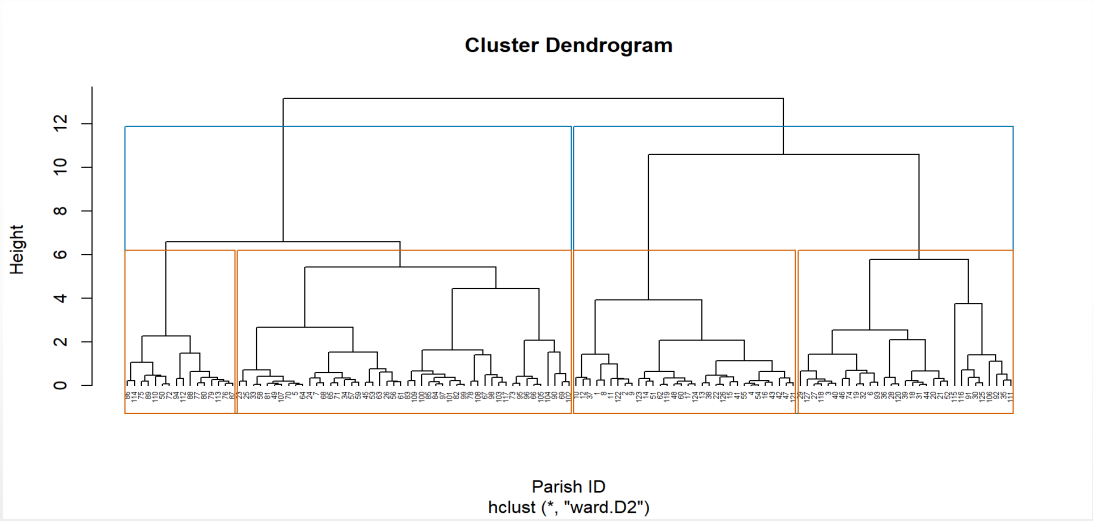
- **High values of b1** → Lower proportion of Non-Western
- **High values of b2** → Lower proportion of Western

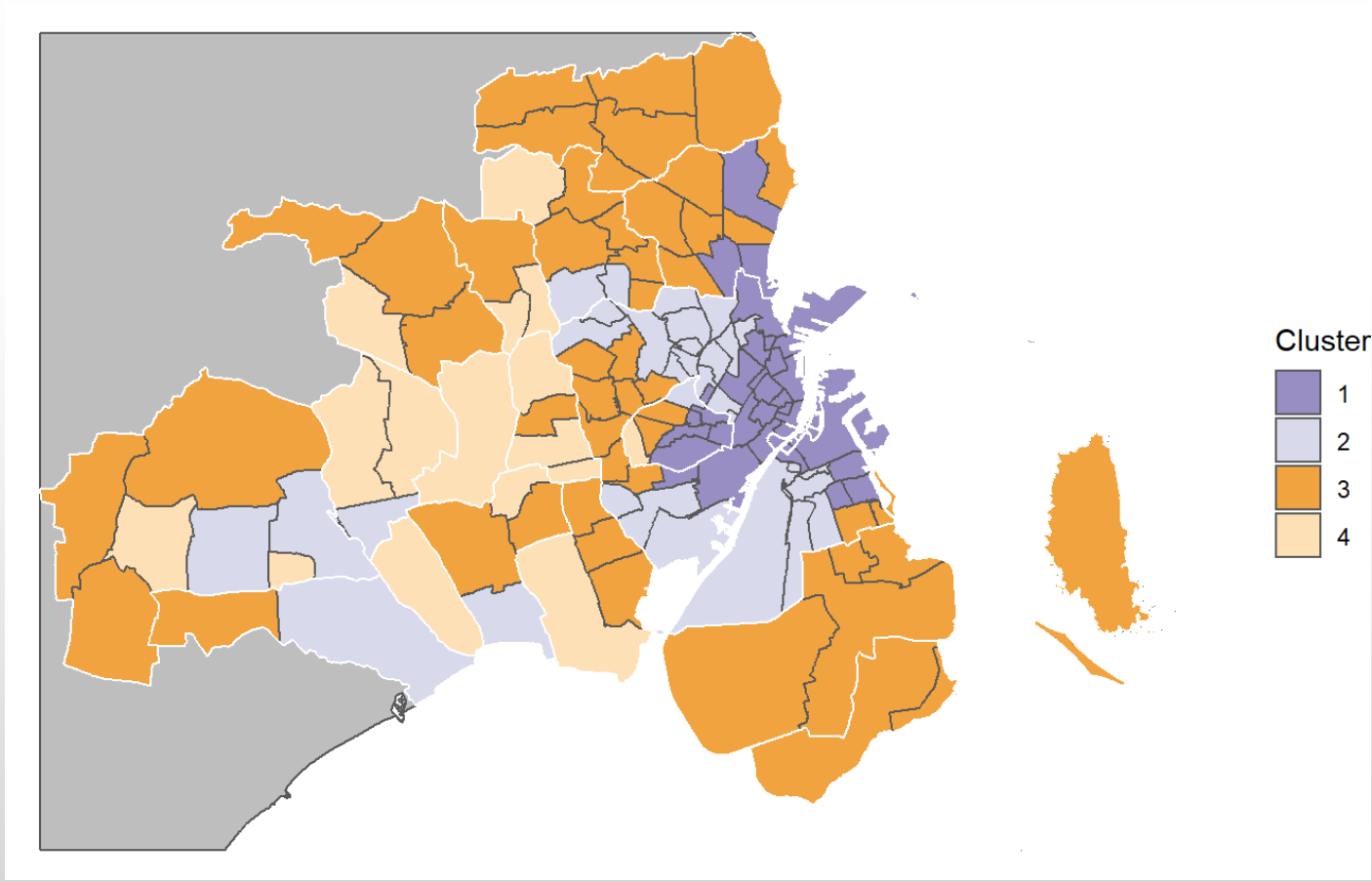


AALBORG UNIVERSITY
DENMARK

Case Study

Hierarchical cluster analysis





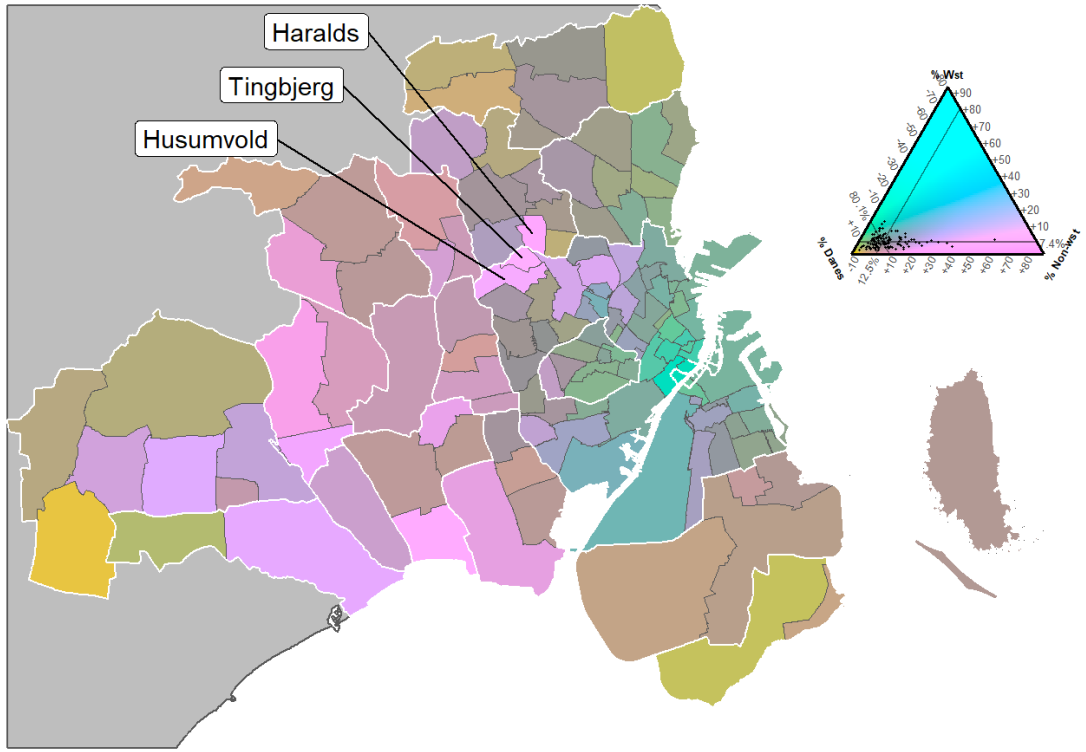
Compositional mean (%)

Cluster	N	D	Wst	No-Wst
1	32	81.0	11.2	7.8
2	31	66.6	9.3	24.1
3	48	85.9	5.2	8.9
4	16	74.8	4.7	20.5

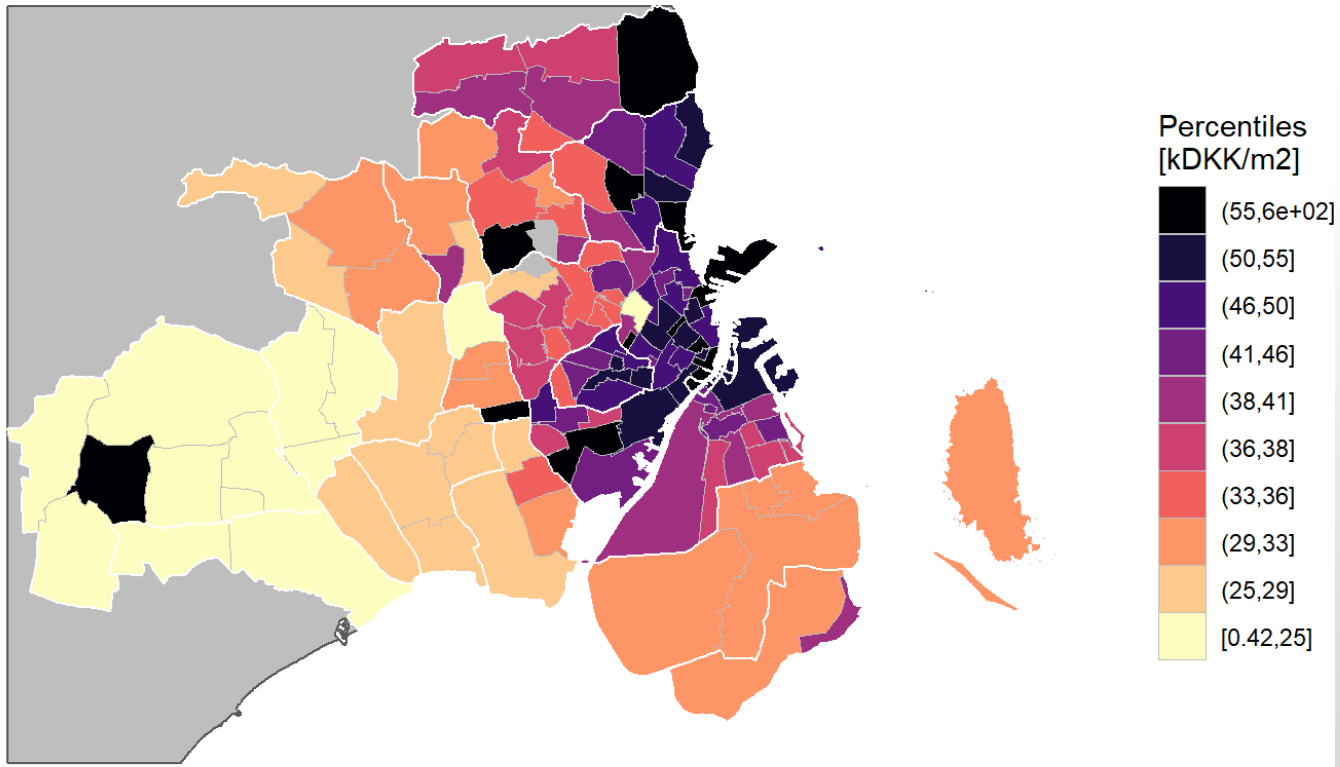
Case Study

Housing prices and migration

Population structure



Median house prices

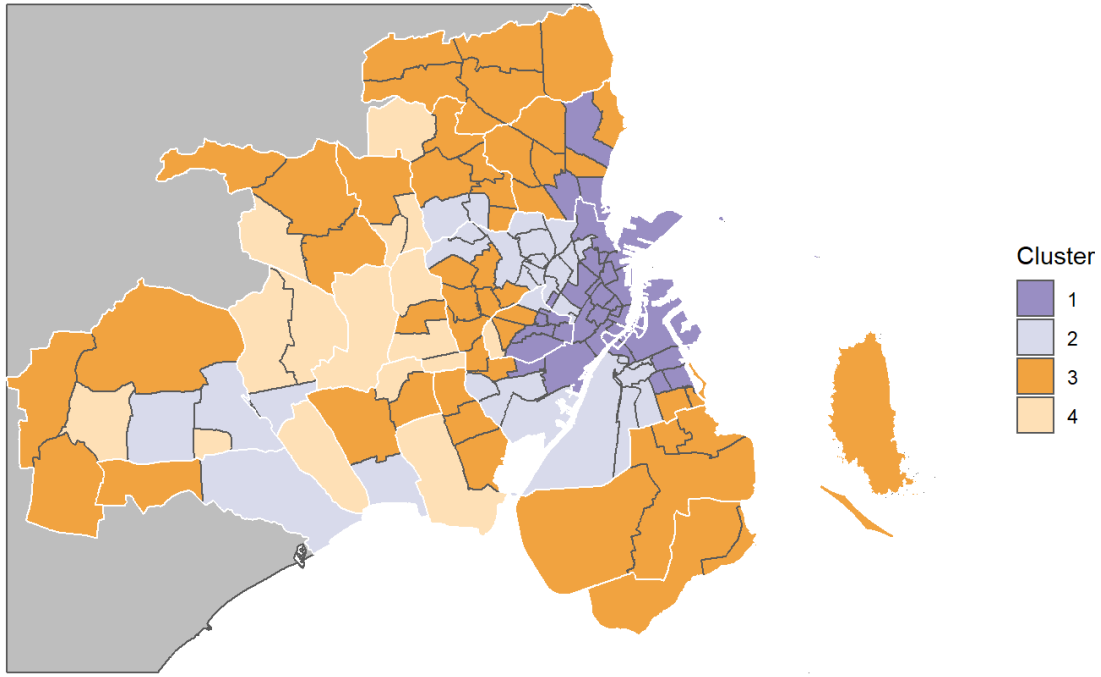


Case Study

Housing prices and migration

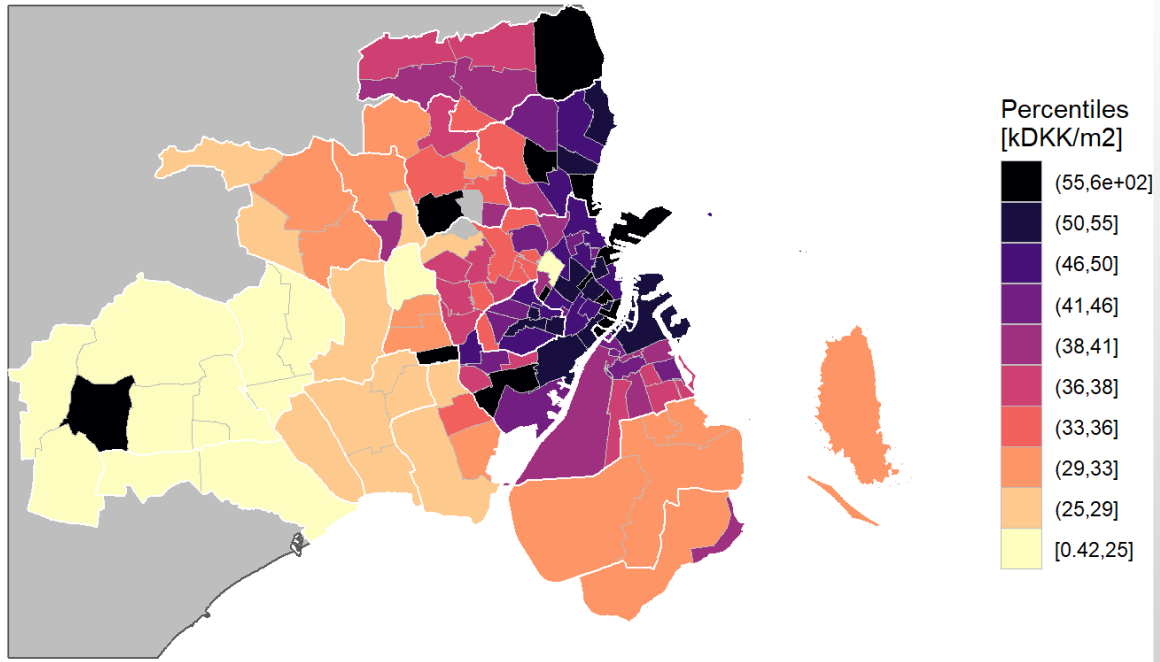
Clusters based on balances

Cluster	N	D	wst	nwst
1	32	81.0	11.2	7.8
2	31	66.6	9.3	24.1
3	48	85.9	5.2	8.9
4	16	74.8	4.7	20.5



Median house prices

Cluster	1 (N = 32)	2 (N = 29)	3 (N = 48)	4 (N = 16)
Mean	68	37	36	51
Median	**51**	**39**	**36**	**27**
IQR	47 - 55	34 - 42	30 - 40	25 - 32
Range	37 - 603	0.4 - 71	20 - 56	5 - 367



Case Study

Housing prices and migration

Linear model: Median house prices at parish level as dependent variable and the two balances as the independent variables.

$$\ln(HP_i) = \beta_0 + \beta_1 \cdot b_{1i} + \beta_2 \cdot b_{2i} + \varepsilon_i$$

Issues:

1. We cannot interpret the coefficients (β) as the increase/decrease of Y due to an increase/decrease of X; instead we need to think in the relative behaviour of the components.
2. Only the coefficient of the balance that explain the ratio between one component and the others is interpretable

Regression coefficient (β) of pivot coordinates

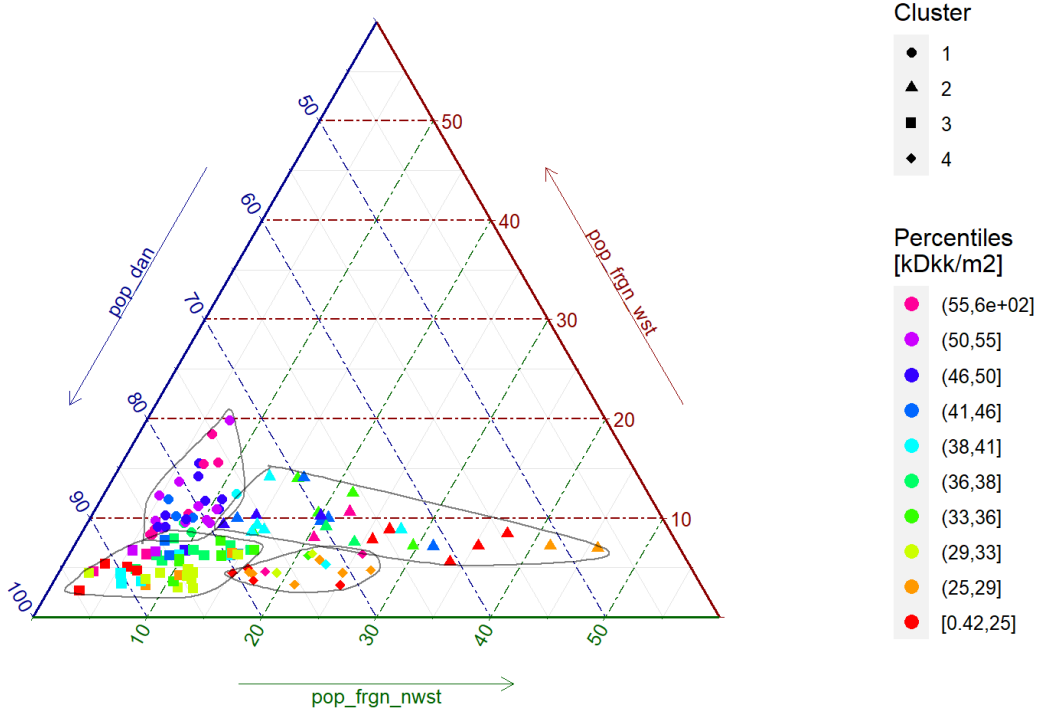
Model	Estimate	CI (lower)	CI (upper)	Std. Error	t value	Pr(>|t|)	
Model 1 (dan & wst vs. nwst)	0.359	0.166	0.551	0.097	3.683	<0.001	***
Model 2 (dan & nwst vs. wst)	-0.430	-0.699	-0.162	0.136	-3.170	0.002	**
Model 3 (nwst & wst vs. dan)	0.072	-0.217	0.360	0.146	0.494	0.622	
Note: CI = 95% confidence interval							

Case Study

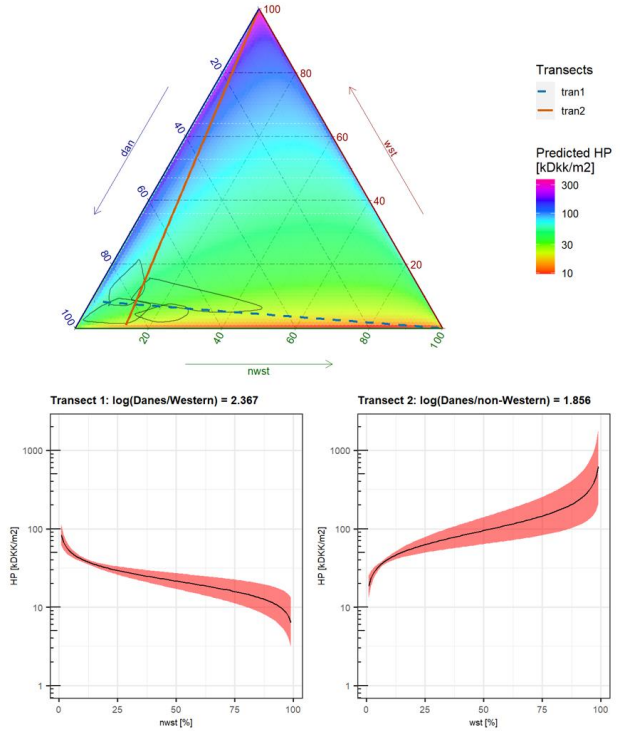
Housing prices and migration

Non-Western and Western migrants are associated with the median house prices at parish level. However, they are not independent predictors by themselves and their interpretation should be done in comparison with the other variables (i.e. Danes, Wester, Non-Western)

Ternary diagram



Predicted median housing price



Case Study

Main findings

- We have detected **four main clusters** according to population structure:
  - Danes are the main population in the region but they tend to avoid the city centre
  - Western migrants, on the other hand, prefer the central areas
  - Non-Western migrants increase in the western peripheral parishes
- We have also found that **migration patterns might help to interpret some of the variation in the median house prices:**
  - Prices are negatively correlated with Non-Wester population, while there is a positive correlation with Western population
  - It seems however there are also some location influences in the median house prices: CL1 and CL2 have higher values than CL3 and CL4
- Can we interpret the **causes of our observations**?
  - Do Immigrants tend to settle down in areas with an ethnicity background similar to their country of origin?
  - Are there socio-economic (status) segregations?
  - Are Western migrants wealthier than Non-Western?
  - Are there negative effects of Non-Western population on the neighbour?

Conclusions

CoDa analysis in in population geography

- **CoDa techniques are robust** and more appropriate than standard statistical and geostatistical methods when we are analysing close data (e.g. percentages).
  - ❖ Alleviate issues with spurious correlations
  - ❖ Avoid problems with sub-compositions (i.e., results with all the dataset do not contradict results obtained with only a portion of them).
- Compositions only **give information about the relative magnitude of its components**. Different phenomena can lead to the same proportions in the data, and therefore additional information would be needed in order to make inferences with the absolute values.
- **Balances help to data interpretation**. However, there is still some complexity in the interpretation of models based on balances, and they can been seen, somehow, as black-boxes (especially when we have more than three components in our dataset).

Future work

Changes in migrating compositions over time

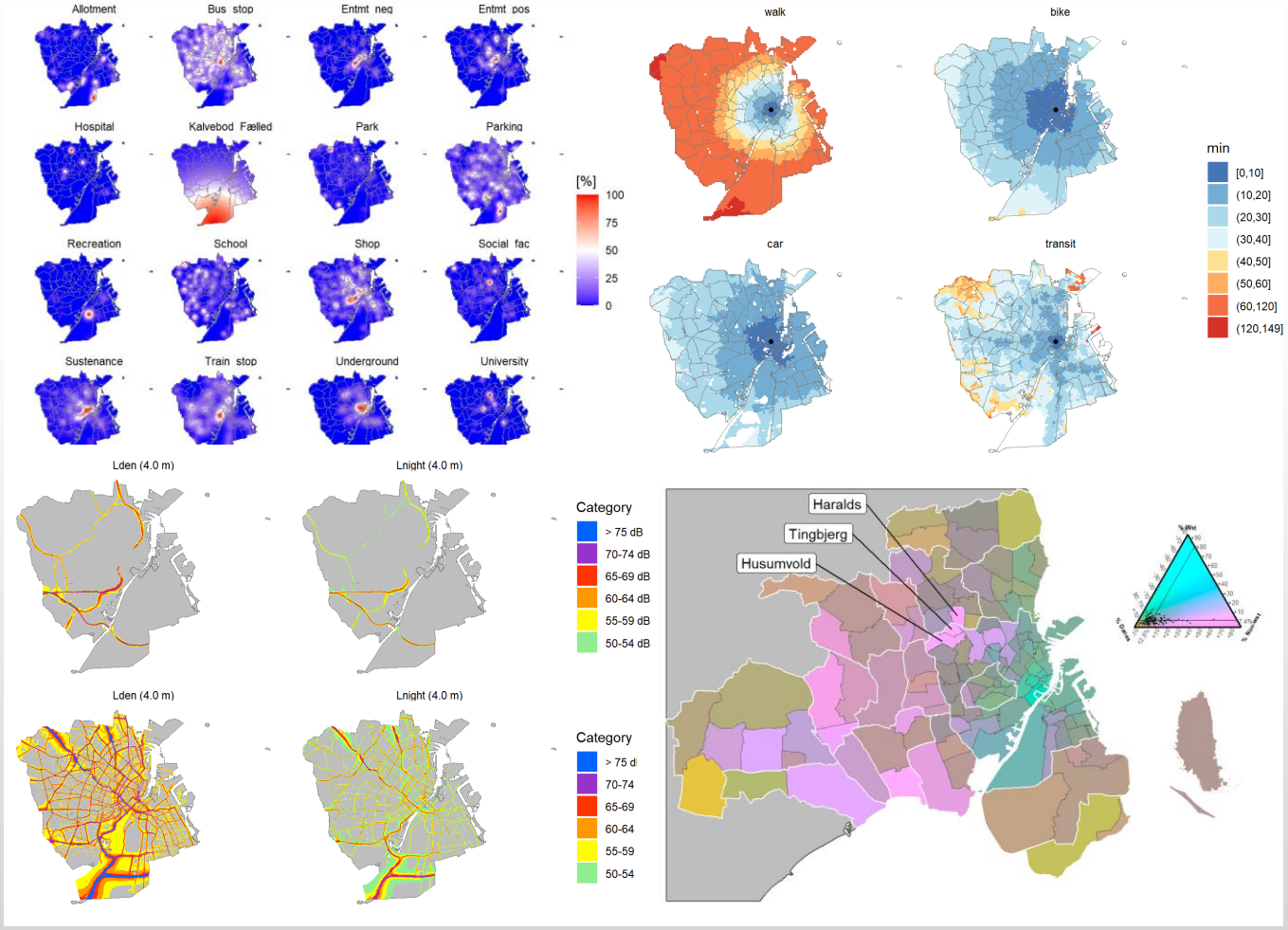
- Until May 2018, Denmark defines an area as "**ghetto**" if at least three of the following five criteria were fulfilled (Wikipedia):
  1. The share of inhabitants aged 18–64 neither in employment nor education is higher than 40%, as an average over the span of 2 years.
  2. **The share of immigrants and their descendants from non-Western countries is higher than 50%**.
  3. The share of inhabitants aged 18 and over convicted for infractions against the penal law, weapons law or drug regulations is greater than 2.7%, as an average over the span of 2 years.
  4. The share of inhabitants of aged 30–59 with only primary education or less, is greater than 50%.
  5. The average gross income for inhabitants aged 18–64 excluding those in education is less than 55% of the average gross income for the region in question.
- **Research question: What are the impacts of ghetto policies?** Are they really working (e.g. reduction of the share of non-western migrants in those areas)?

Future work

Migration and housing prices

Hedonic model: house prices can be modelled based on:

- **Structural characteristics of the house:** age, size, building materials, floor level, etc.
- **Location**: proximity to urban services, distance to Central Business District - CBD, accessibility, etc.
- **Surrounding environment:** neighbourhood services and socio-economic aspects of its inhabitants, leisure facilities, noise levels, **population structure**, etc.



Migration studies with a Compositional Data approach: a case study of population structure in the Capital Region of Denmark

Javier Elío, Marina Georgati, Henning S. Hansen, Carsten Keßler

Thanks!!

Any question?

AALBORG UNIVERSITY
DENMARK

Twitter: @Elio_Javi
Email: javierdem@plan.aau.dk
Website: https://javierelio.netlify.com