Estimating sub-national behaviour in the Danish microsimulation model SMILE

Marianne Frank Hansen
DREAM

Joint Eurostat/UNECE Work Session on Demographic Projections
Geneva, 18-20 April 2016

Outline

- Background information.
- The microsimulation model SMILE.
- Estimating transition probabilities by classification – the CTREE algorithm.
- Method challenged by ambition to identify geographical areas likely to be characterized by future exodus and depopulation
- Principal component analysis – PCA.
- Using PCA as a pre-process to aid convergence of classification algorithm.

Background information

- Danish Rational Economic Agents Model is an independent institution founded in 1997
- Purpose: Developing and maintaining tools to analyse structural policy, and fiscal sustainability
- Annual projections on future demography, educational attainment, and labour market participation.
- National population projection with Statistics Denmark since 2010.

The microsimulation model SMILE

- Simulation Model for Individual Lifecycle Evaluation
- Developed to project future housing demand
- Requires projection of the number of households
- Initial population consists of 2,8 mio. households comprising 5,5 mio. individuals.
- Transition between states is decided by event exposure and Monte Carlo simulation.
- Events: demography, education, labourmarket participation, retirement, change in cohabitation patterns, moving, dwelling choice.

The microsimulation model SMILE

- Transition probabilities associated with events are decribed by a vast range of high dimensional characteristics:

- Family structure and gender (3)
- Age (120)
- Origin (5/15)
- Children or not (2)
- Education level (6/12)
- Labourmarket status (2/3)
- **Geographic location (11)**
- Dwelling type (5)
- Dwelling category (9)
- Dwelling size (8)
- Dwelling building year (12)
- Dwelling area (5)

- Curse of dimensionality

Estimating transition probabilities by classification

- Curse of dimensionality challenges estimation of transition probabilities.
- Solved by classifying observations with similar responses by CTREE algorithm on pooled data.
- New Task: Identify geographical areas likely to be characterized by future exodus and depopulation
  - → expand geographical covariate from 11 regions to 98 municipalities.
  - → classification algorithm does not converge!!!

Estimating transition probabilities by classification

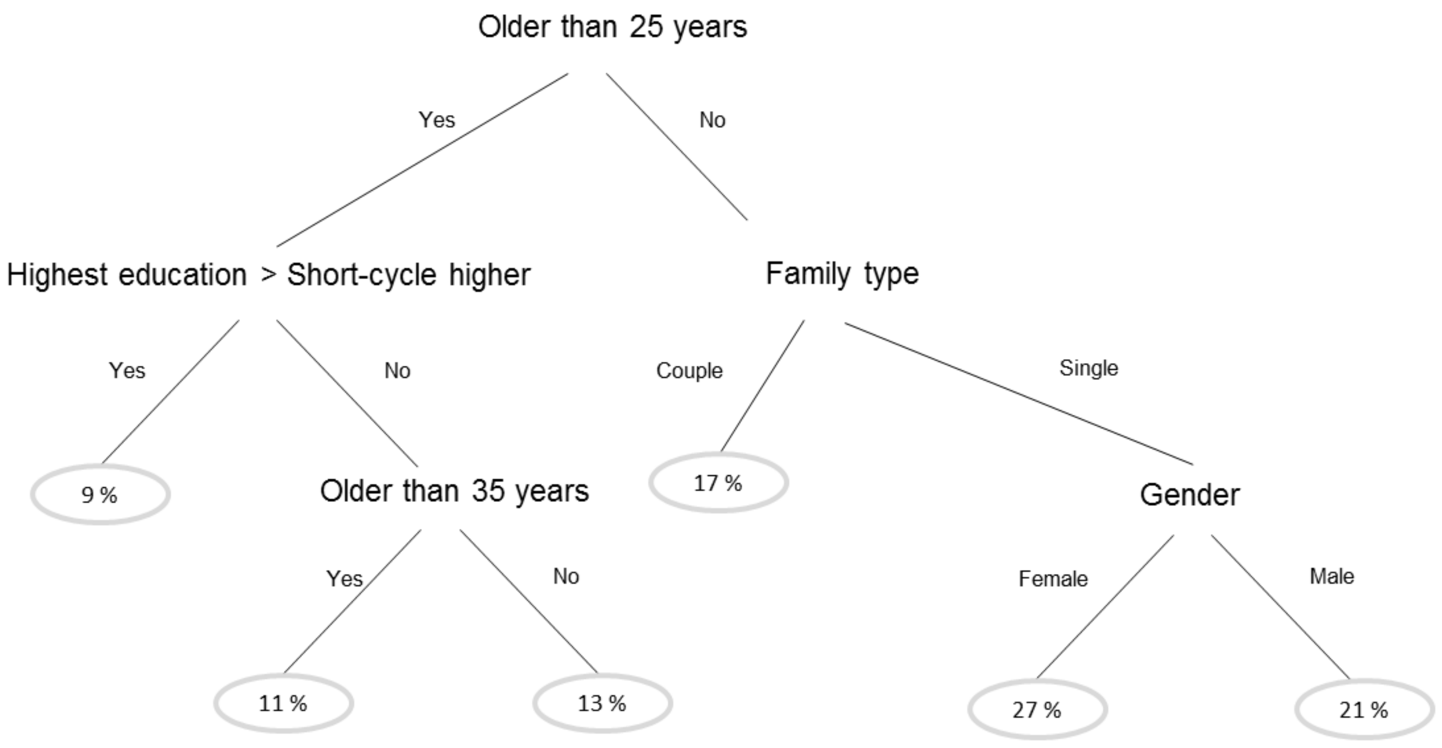
- Conditional inference trees (CTREEs) is an algorithm used to classify observations with similar behaviour/response from a range of characteristics.
- Decision tree grouping data by recursive binary splits
- Observations are grouped such that there is:
  - Minimum variation within a group
  - Maximum variation across groups
- Splits decided by statistical tests and stopping criteria.
- Calculate probability of event for each terminal group.

CTREE algorithm

1. Test for independence between any of the explanatory variables and the response. Stop if p>0,05.
2. Otherwise select the input variable with the strongest association to the response.
3. Find optimal binary split point for the selected input variable.
4. Recursively repeat from 1) until a stopping criterion is reached.

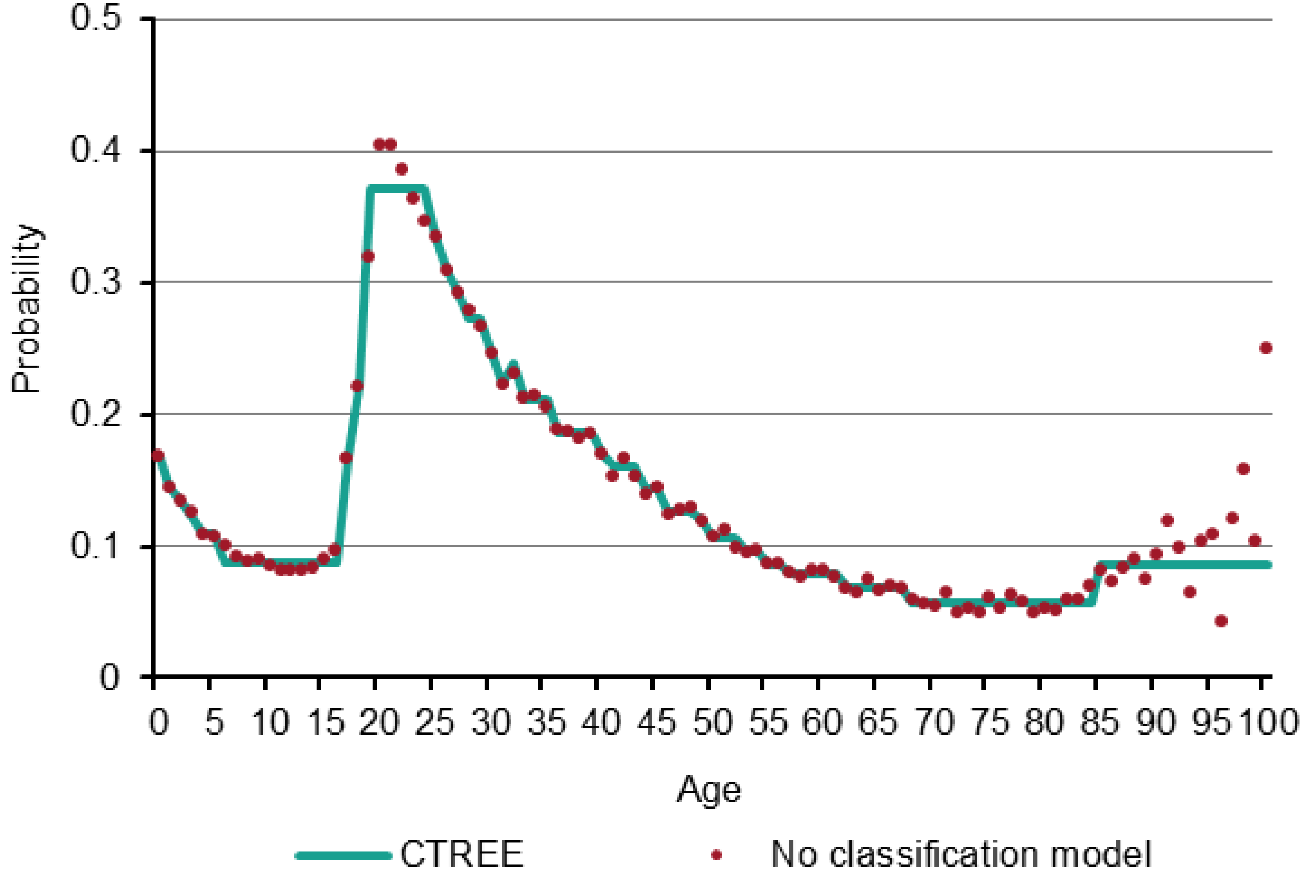
Example: Decision tree

- The probability of moving - binary response.



DREAM

Example: Decision tree



Convergence issues

- Induced when expanding the dimension of the geographic variable to 98.
- Solutions:
  - Use smaller sample when classifying
  - Change stopping criteria or test to allow for fewer splits
  - Order/rank elements of geography variable (municipalities)
- Ordered vs. non-ordered variables.
- Choice of ranking measure?
- Using the position of the municipality in a principal component vector will allow ranking to spring from multiple features…

Principal Component Analysis

- Data matrix *X*, *n* obs., *p* variables/features
- A principal component, Z_k, $k = \min(n-1, p)$

is a linear combination of columns in X

$$Z_k = \phi_{1k} X_1 + \phi_{2k} X_2 + \ldots + \phi_{pk} X_p$$

$$z_{ik} = \phi_{1k} x_{i1} + \phi_{2k} x_{i2} + \ldots + \phi_{pk} x_{ip}$$

- *Loadings* and *scores*
- Principal component are determined successively

Principal Component Analysis

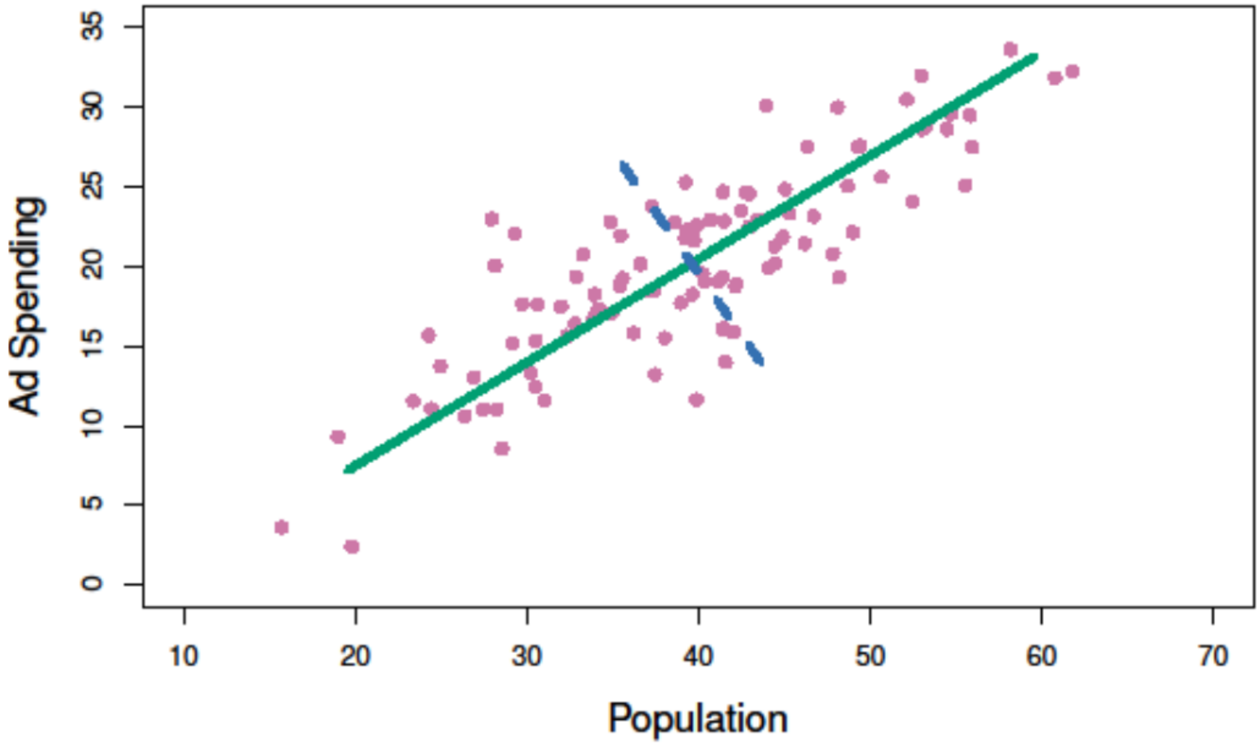
- Center $X_1, \ldots, X_p$ to have zero mean
- 1st PC (Z_1):

$$\max\left\{\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2\right\} \quad \text{s.t.} \quad \sum_{j=1}^{p}\phi_{j1}^2 = 1$$

- 2nd PC (Z_2):

$$\max\left\{\frac{1}{n}\sum_{i=1}^{n} z_{i2}^2\right\} \quad \text{s.t.} \quad \sum_{j=1}^{p}\phi_{j2}^2 = 1 \text{ and } corr\left(\mathrm{Z}_1, \mathrm{Z}_2\right) = 0$$

Principal Component Analysis



Source: Figure 6.14, Hastie & Tibshirani (2013).

Ranking municipalities by principal component score values

- Set of principal components constitutes a low-dimensional representation of data.
- Z_k spans the variability in data, hence observations *s* and *t* are more similar than observations *s* and *u*, when

$$|z_{sk} - z_{tk}| < |z_{sk} - z_{uk}|$$

→Principal component score values can be used to rank geographic areas based on multiple features.

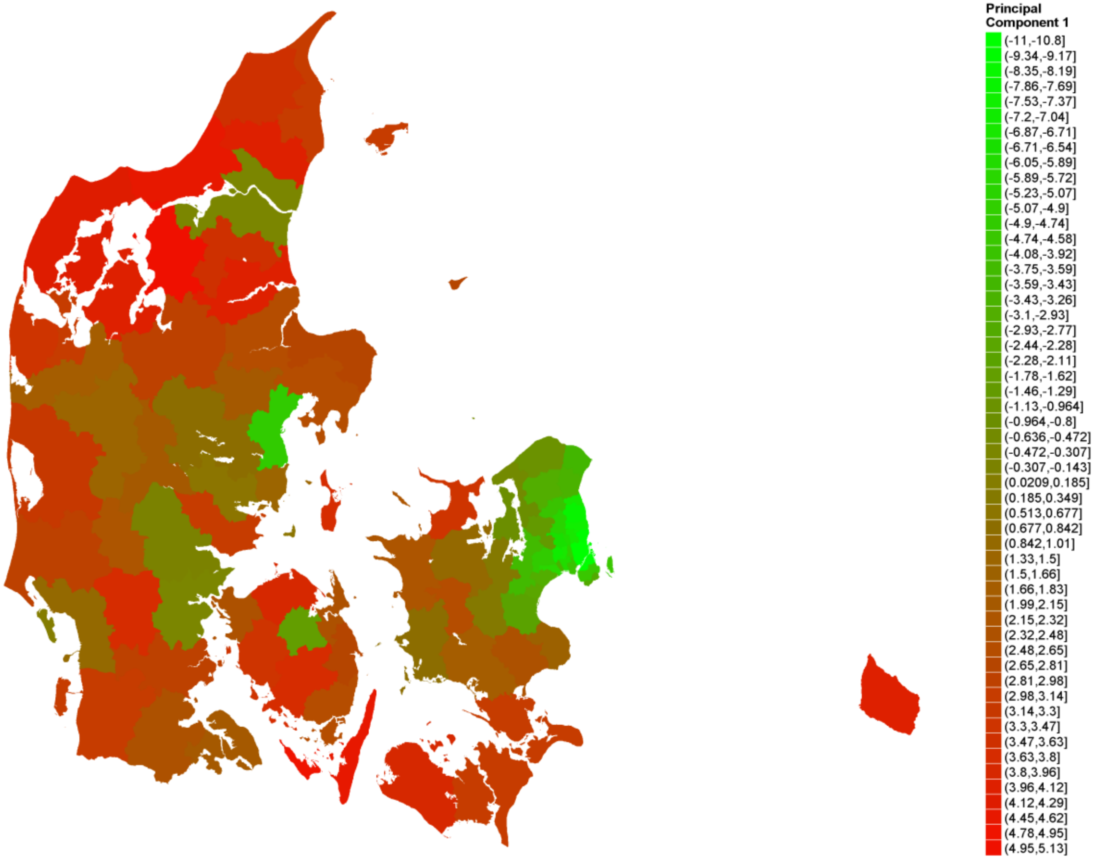
Ranking municipalities by principal component score values

- Municipality Fundamentals Database of The Ministry of Social Affairs and the Interior.
- 60 variables describing demographic, socioeconomic and economic features of each municipality.
- Data $x_{ij},\ i=1,\ldots,98,\ j=1,\ldots,60$
- Perform PCA on *X*.
- Rank the 98 municipalities by elements in Z_k
- Estimate CTREE with ranked elements of geographic feature variable(s).

Score values of 1st PC
Level of urbanisation

Negative correlation:

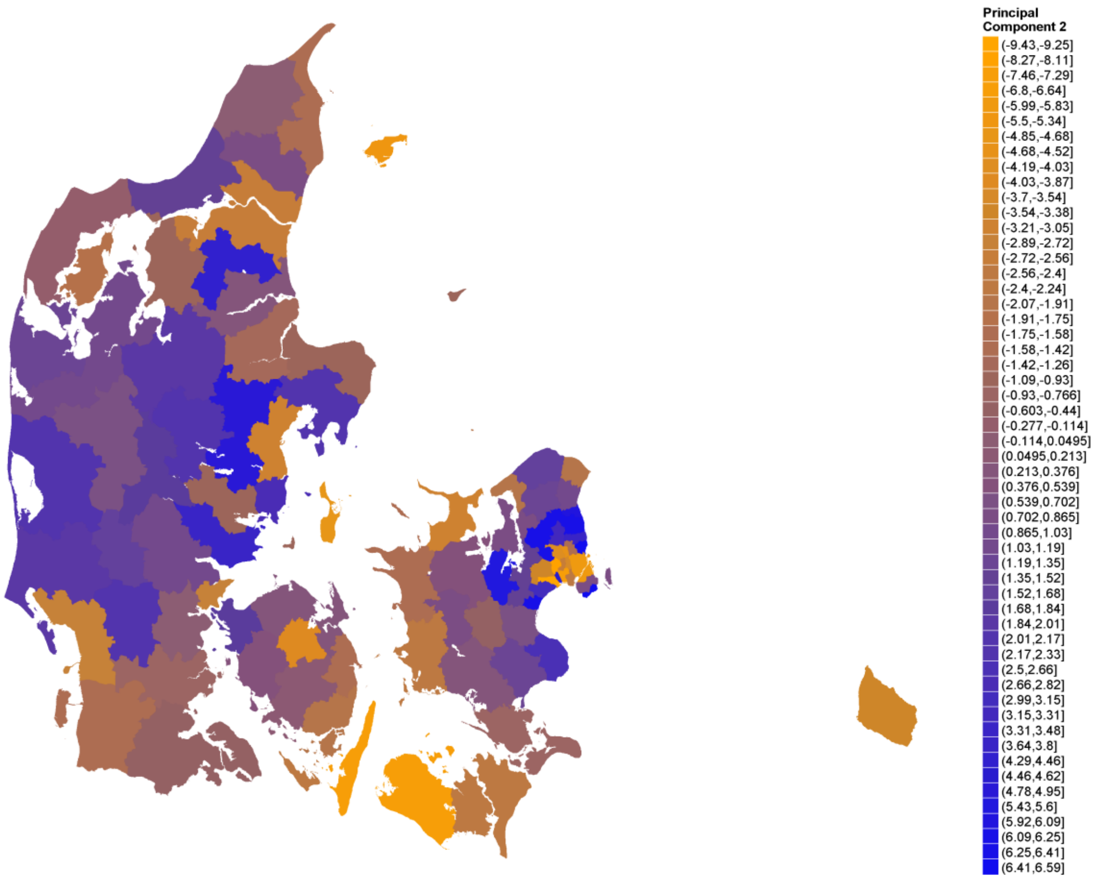
- Share of pop. in urban housing.
- Share of pop. commuting.
- Per capita revenue from real estate and income taxes.
- Share living in social housing.
- Land value per capita.
- High level of education.
- Population density.
- Share of Western immigrants.



Score values of 2nd PC Budget balance

Negative correlation:
- Exp. per capita
- Share of pop. with basic education
- Cash benefit recipients
- Unemployment

Positive correlation:
- Income tax per capita
- Land value per capita
- Owner-occupied housing
- High level of education.



Conclusion

- Classification is a useful tool when estimating behaviour based on a large number of high dimensional covariates.
- Expert knowledge of behavioural patterns or model tuning not required.
- PCA can be used to aid convergence of CTREE by ranking variable elements from multiple features.
- Allows for the introduction of detailed sub-national behaviour in SMILE

For more information please visit

www.dreammodel.dk