RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

Latent Class Analysis in Research Policy and Higher Education

Lugano, 9th of September 2019

Benedetto Lepori

The issue

- Large amount of integrated data developed by RISIS
  - Combining data sources (projects, publications, etc.)
  - Adding integrative dimensions (actors, space, topics)
- Data exploitation is made **complex** due to several issues
  - Data are nested/multilevelled
  - **Data are heterogeneous and heterogeneity needs to be modelled**
  - Dataset are complex in terms of data availability (missing), type of data (categorical vs. continous) and distributions (lognormal, outliers)
  - Data could be affected/influeced by soft characteristics of context (governance, etc.) that impacts on their interpretation

Focus on heterogeneity:

- One of the main source of complexity in analysis RISIS data is heterogeneity, i.e. the fact that units of analysis have very different substantive nature.
- Different types of heterogeneity: individual, organizational, country, longitudinal;
- Heterogeneity could be a problem pooling data across observations is likely to produce misleading results.
- But might also be of substantive interest (i.e. understanding the types of universities in Europe).

-> Adoption of statistical tools and methods to deal with heterogeneity in RISIS2 data

How to deal with heterogenenity?

- Heterogeneity can be removed in panel data using fixed effects
  - Problematic when differences are of essential nature and represent most of the variance in the sample
- Heterogeneity can be modelled with reference to:
  - Observables variables
    - Directly scored/measured/observed
    - (continuous, discrete)

-> introduced directly in statistical models

  - Unobservable/unobserved variables (latent variables)
    - Can be inferred from observable ones
    - (continuous, discrete)

-> Latent variable modelling (LCA and LCRM)

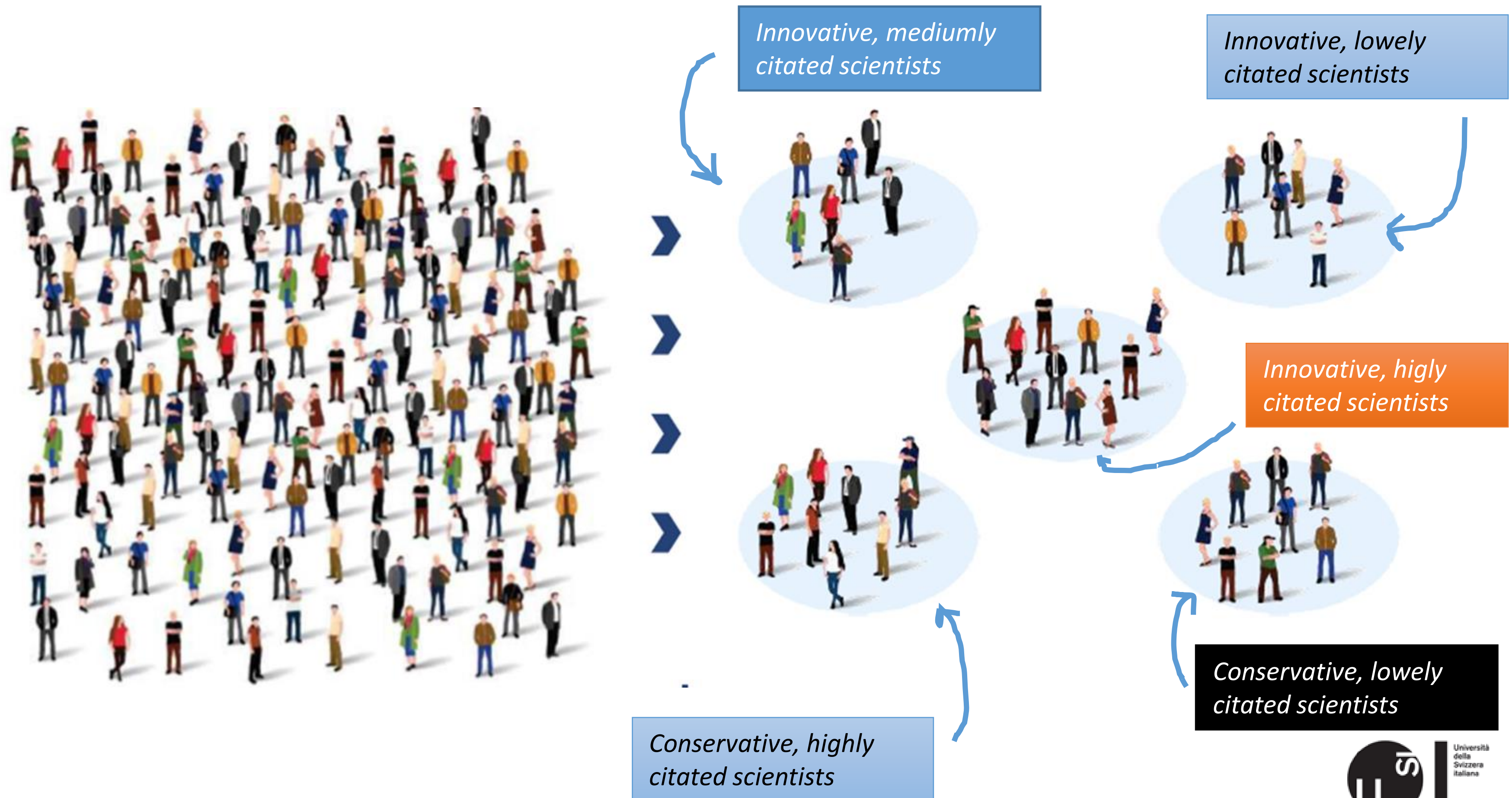
- These approaches are complementary and can be combined

Examples of heterogeneity

- Students with different cognitive abilities
  - Impacting on learning outcomes
- Universities with different mission, legal status, internal governance
  - Impacting on their resourcing and profile
- Researchers with different mobility history
  - Influencing their productivity

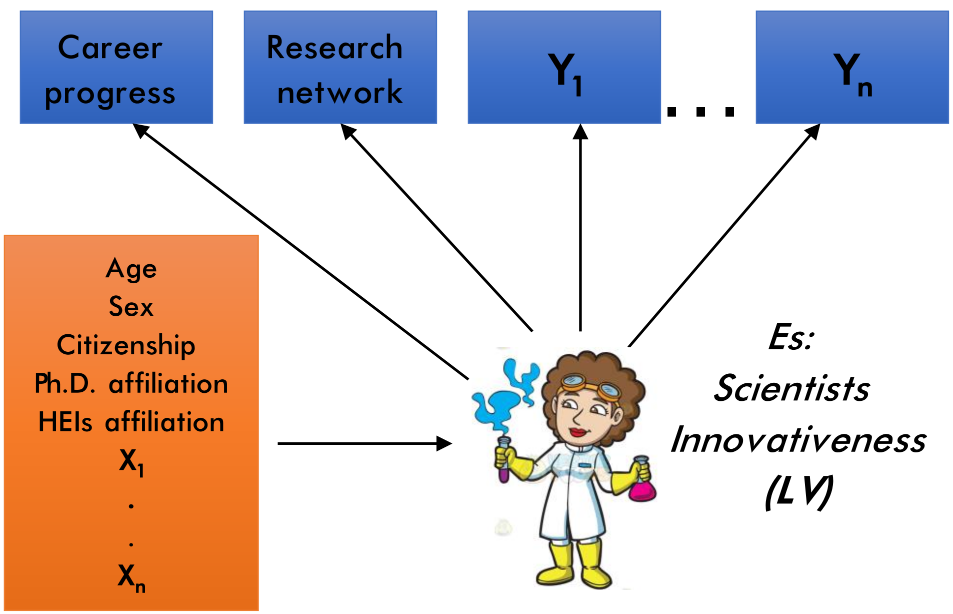
What does Latent Class modeling tell us?

- It captures (and models) heterogeneity
- In latent class models, we use a latent variable that is categorical/discrete to represent the groups, and we refer to the groups as CLASSES.
- Observation within the same class are supposed to be homogeneous, while those in different class are dissimilar (identify groups of cases with similar **data patterns**)



Latent Class Modelling: basic idea

- There are groups in our population and members in these groups behave differently.
- We don't have a (observable) variable that identifies the groups, then differences in behavior is due to an unobservable variable (latent variable, **LV**).
- The value of a **LV** can be deduced (inferred, throughout this mathematical model) from observed (measured) variables.



- Observed Continuous (ys) or Categorical **Items** (us)
- Categorical Latent Class Variable (**LV**)

LCA

- Continuous or Categorical **Covariates** (x) (exogenous)

LCRM

This course

- Learn statistical methods that allow dealing with unobserved heterogeneity
  - latent variables and classes
  - Influencing the observed characteristics

- Implement these mothds for some real cases in science and higher education
  - Also to practice the methods and learn about potential issues

Programme of the course: Day 1

- 9:00-10.00 Latent Class Analysis in Research Policy and Higher Education (Benedetto Lepori)
- 10.00-11:00 Introduction to Latent Class Analysis (Francesco Bartolucci)
- 12:00-13:00 LCM with STATA (Barbara Antonioli Mantegazzini
- 13:00-14:00 Lunch
- 14:00-14:30 Introduction to Group Exercise (Barbara Antonioli Mantegazzini)
- 14.30-18.00 First session group work

Programme of the course: Day 2

- 9:00-11:00 Second session group work (Barbara Antonioli Mantegazzini)
- 11:00-12:30 Group Presentations
- 12:30-13:00 Closing Remarks and Recap (Benedetto Lepori)
- 13:00-14:00 Lunch

Empirical application of Latent Class Modeling

- “*The heterogeneity of European Higher Education Institutions. A typological approach*» (Lepori B., 2019)

The problem: heterogeneity in higher education

- European HEIs very diverse in terms of activity profile, subject orientation, size, etc.
  - public policy distinguishing between sectors of higher education
  - differentiation processes of HEIs and of scientific disciplines

- We have a poor understanding of such heterogeneity beyond the university/colleges distinction
  - Main lines of differentiation
  - Blurring between groups/types
  - Country differences

- Classifications as useful tools to analyze heterogeneity
  - Building groups homogeneous across some dimensions
  - Important also for the legitimacy and status of institutions

Existing approaches

RISIS

- Use exogenous variable (university vs. colleges)
  - Simple, but static and does allow dealing with blending
- Ex-ante classifications (Carnegie)
  - Based on in-depth knoweldge of systems
  - Very useful, but difficult to justify in terms of classes and indicators
- Clustering
  - Data-driven, no underlying model
  - No fit measure
  - Results difficult to interpret if clusters are not clear-cut

Università della Svizzera italiana

LCA

- Combines advantages of a priori and data-driven approaches
  - Explicit modelling of the relevant dimensions
  - Inclusion of exogenous variables in a flexible way
  - Optimize fit with the data
  - Fuzzy groups (probabilities)

Model

- Observed HEI characteristics $\boldsymbol{y}$ as a mixture of distributions contingent to the class probabilities π_i

$$\mathrm{f}(\boldsymbol{y}) = \sum_{ij} \pi_i f_i(\boldsymbol{y})$$

Class probabilities as a logistic function including exogenous variables

$$\pi_i = f_i(\mathbf{x}) = \frac{\exp(\gamma_i)}{\sum_1^g \exp(\gamma_i)} \qquad \textstyle\sum_i \ \pi_i \ = 1$$

$$\gamma_i = \ \theta_i + \ \mu_i \, (legal\ status) \ + \ \vartheta_i \, (research\ mandate)$$

Model optimizes the fit with the data for $\boldsymbol{y}$ with fixed number of classes and computes fit statistics (AIC) for the selection of the best number of classes

Selection of variables

- theoretical reasoning on the relevant dimensions
  - literature-based
- Discriminating power of variables
  - Statistical analysis
- Data availability
  - Despite imputation in GSEM

Dimensions

- Activity profile
  - Education vs. research vs. third-mission
- Subject scope
  - Generalist vs specialist
  - SSH vs. NATSCI
- Resources
- Legal status and institutional mandate (research vs. education)

Data

- Data from the European Tertiary Education Register (www.eter-project.com), 2014 edition
- Enriched with data from WoS, EU-FP EUPRO database and PATSTAT
- Final sample (excluding cases with missing staff data): 2,243 observations in 30 European countries

Variables

- Size: ln(staff)
- Education: education intensity; masterorientation
- Research: research intensity (composite), citations per staff
- Third mission: patent intensity
- Subject scope: subject concentration, students SSH, students natsci
- Exogenous: legal status, PhD awarding

Methods: latent class clustering

- Modeling the distribution of the observed variables
- Mixture of normal distribution contingent to the observation belonging to a class
- Probability of a class contingent of the regulatory variables (logistic regression)

$$\mathrm{f}(\boldsymbol{y}) = \sum_{ij} \pi_i f_i(\boldsymbol{y})$$

$$\pi_i = f_i(\mathbf{x}) = \frac{\exp(\gamma_i)}{\sum_1^g \exp(\gamma_i)}$$

$$\gamma_i = \theta_i + \mu_i\,(legal\ status) \quad + \vartheta_i\,(research\ mandate)$$

- The model computes the distribution parameter and the distribution of cases by class
- Optimal number of classes can be identified using fit statistics (AIC/BIC)
- Attributing cases to classes with the highest probabilities

Results

RISIS

- Five class model can be interpreted in a straightforward way

- Two main discriminant dimensions
  - Research vs. education
  - Subject composition: generalists vs specialists

- Associated with regulatory dimensions
  - But exceptions (theological schools, etc.)

- Five classes

1: Specialised colleges, small, SSH, no research, mostly no PhD
2: Research universities, generalists, research, PhD
3: SSH universities, strong education
3: Technical universities, research, PhD, patents
5: Large generalist colleges and some universities

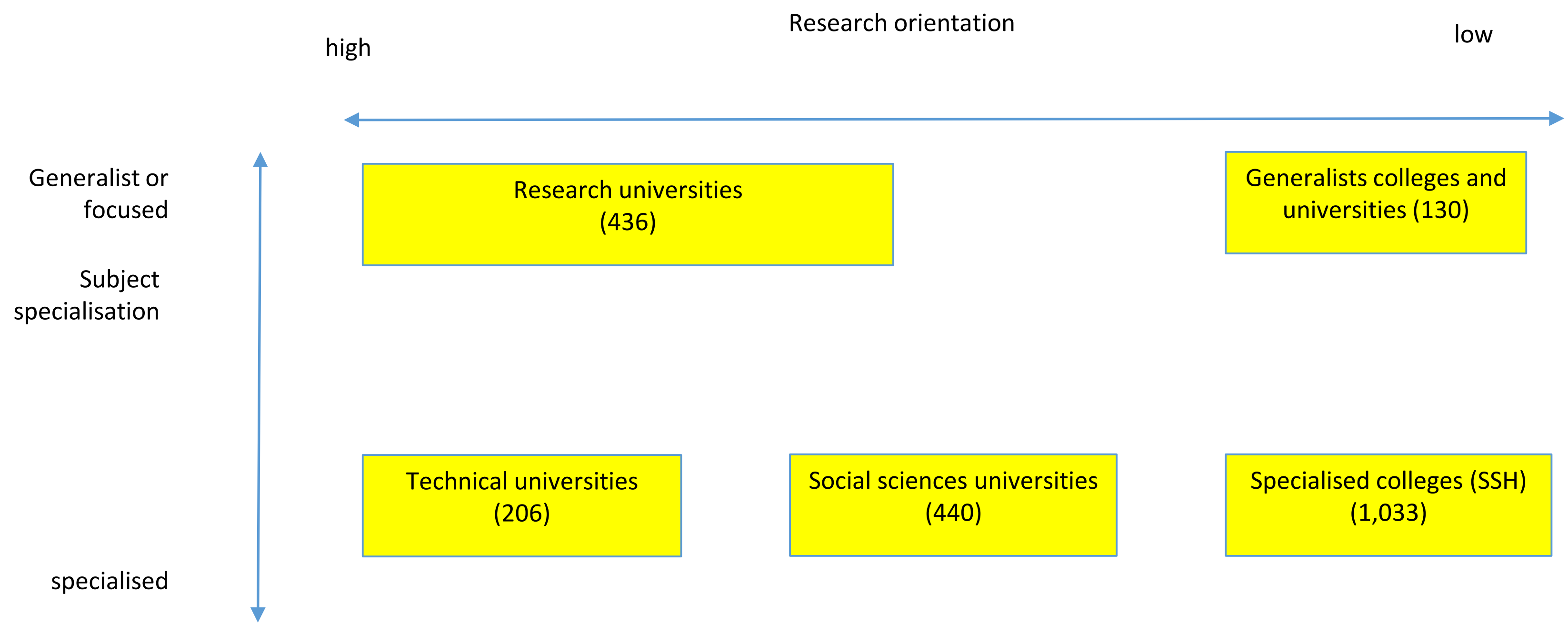
Class characteristics

RISIS

Class		N	Regulatory characteristics			
			PhD		Legal	
			no	yes	Public	private
1	Specialised colleges (SSH)	1'033	819	214	527	506
2	Research universities	436	15	421	421	15
3	Social sciences universities	440	79	361	335	105
4	Technical universities	206	36	170	193	13
5	Generalist colleges and universities	115	76	39	102	13

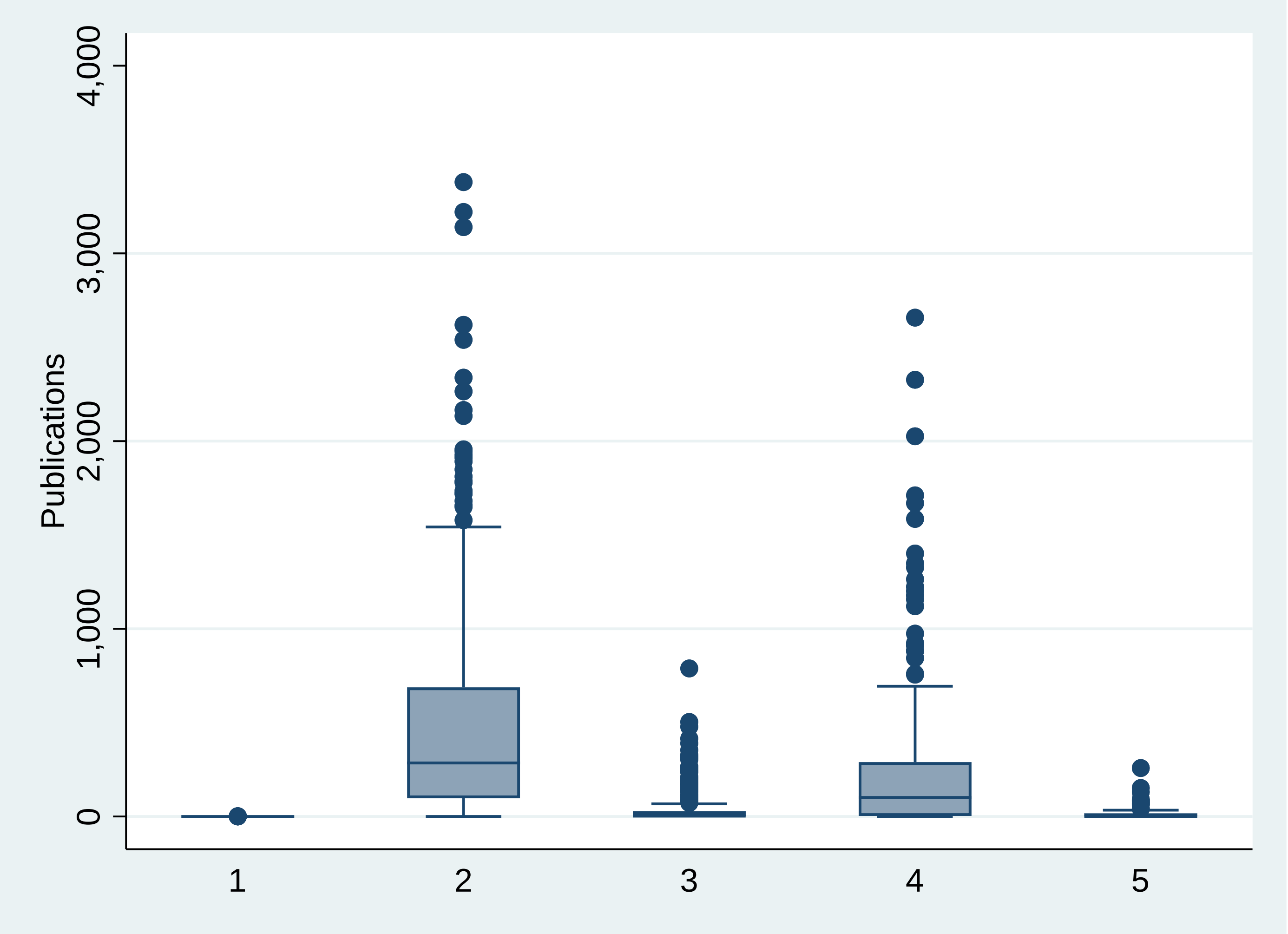
Class		Median characterizing variables								
		academic staff	education intensity	masterorientation	research intensity	citationsstaff	patentintensity	HF students	students SSH	students natsci
1	Specialised colleges (SSH)	52.0	18.5	0.14	0.000000	0.000	-	0.64	0.93	0.00
2	Research universities	1092.0	16.0	0.32	0.000059	0.269	0.018	0.18	0.57	0.26
3	Social sciences universities	257.4	21.0	0.23	0.000014	0.003	-	0.38	0.82	0.09
4	Technical universities	415.0	14.8	0.40	0.000059	0.194	0.041	0.44	0.16	0.71
5	Generalist colleges and universities	387.4	21.5	0.15	0.000004	0.000	0.005	0.29	0.44	0.40

Università della Svizzera italiana


Research orientation
high
low
Generalist or focused
Subject specialisation
specialised
Research universities (436)
Generalists colleges and universities (130)
Technical universities (206)
Social sciences universities (440)
Specialised colleges (SSH) (1,033)

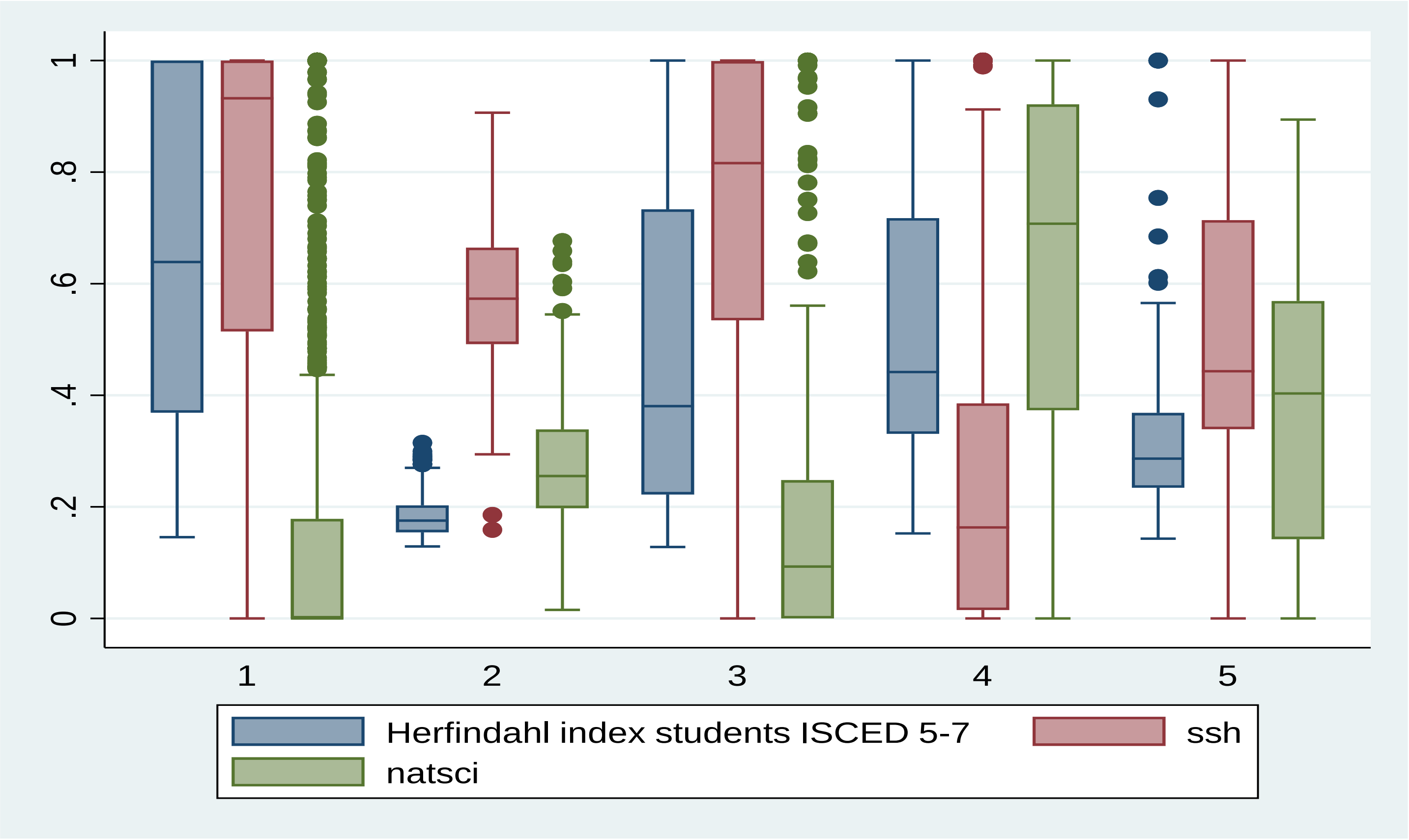
Class composition and characteristics

RISIS



Università della Svizzera italiana

Class composition and characteristics



Discussion

- Classes have their own profile and are interpretable in meaningful terms
- Method is more flexible than assignment based on hard criteria
- Two key dimensions of distinction within the system
- Identification of specific groups of specialists
- Split some groups are highly heterogeneous (generalist universities) and would need more detail
  - What about global universities?

Further work/developments

- Incorporate more priors indicative of status:
  - For example LERU, Coimbra group, etc.
  - Or network centrality measures

- Integrate measures of internationalization:
  - Education, respectively research (international publications, network centrality)

- Use ex-post expert opinions
  - to check and correct misclassifications

- Develop the interpretation of classes in terms of audiences and market positioning

- Link groups prevalence with national specificities

Methodological remarks

- Method is very flexible
  - Allows incorporating priors such as group membership
- Model design is key to results
  - To get stable and interpretable results

RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

THANK YOU !

CONTACT@RISIS2.EU

@RISIS_EU

FACEBOOK.COM/RISIS.EU

RISIS2 EU PROJECT