Linking Administrative Data across Administrations: The Danish Case

Lorenzo Cappellari
Università Cattolica Milano and SFI Copenhagen

SIEP, INPS, Rome, 23 May 2018

Introduction

- Denmark is an interesting case study for all those countries wanting to improve the accessibility of admin data to the research community
- Like in other Nordic countries, the key feature is the possibility to link information across the archives produced by several different administrations
- The National Statistical Office (Statistics Denmark, DST) plays a key role for collecting, validating and disseminating the data
- Nowadays, researchers worldwide in social and medical science use these data

Outline

- I will provide a description of the key aspects of the system as I see them from a user perspective
  - History
  - Available data
  - Access conditions
  - Working on the data
  - History of access
  - Empirical illustration

History

- Denmark has a long history of collecting, information on births, deaths, immigration and emigration, disease incidence, and social conditions.
- High-quality data covering the whole population during long periods of time is therefore available.
  - causes of death since 1875
  - compulsory schooling and continuing education for cohorts born after 1945
  - twins are registered for cohorts born after 1870
  - cancer incidence has been registered for the whole country since 1943
- The idea then was using this wealth of information for producing **register-based national statistics**
- Data from separate administrations were conferred to DST, for validation processing and publication of national statistics
- One element of the validation process is cross-validation between registers, requiring a key to trace persons across archives

The CPR

- Denmark introduced the Personal Identification Number (CPR) in 1968
- 10-digit sequence: birth date (DD-MM-YY) + random 4 digits number [odd for men, even for women] [the equivalent of the Italian tax number]
- It was used in a census for the first time at the Population and Housing Census in 1970.
- During the 1970s the first attempts were made to base the production of statistics on registers
- 1976: a register-based population census was conducted as a pilot project
- 1981: a proper register-based population and housing census was conducted containing most of the conventional population and housing census information.

From register-based statistics to microdata access

- Having the data ready, a legal framework for granting access to researchers was still missing
- 1987 Public Administration Act
  - A public authority can impose the **duty of non-disclosure** on non-public servants
  - Disclosure/anonymity/confidentiality with breach punishable by imprisonment
  - Statistics Denmark thereby allows researcher access
- 2000 Act on Processing of Personal Data
  - Duty of notification to data protection agency before collection/use
  - Approves disclosure of data for scientific purposes
  - Applies whether using a register or collecting a survey
  - Private individuals & public agencies can process CPRs for scientific work
  - Subject's cannot oppose the use of personal data for scientific work

Data available for researchers

- Most of the data derive from the administrative registers of governmental agencies; high data quality for the entire Danish populations of persons, buildings and companies
- Data can be combined in endless ways
- Allow researchers to produce unique analysis of dynamic processes and fluctuations, using the Danish population as their study population.
- All together, data from 250 subject-areas

The Candy Shop

Table II. Examples of Danish registers on economic and social issues.

Danish register	Content	Start year of registration	Data administrator
The Student Register	Grade-level information on compulsory schooling (primary and lower secondary education), upper secondary education and vocational education	1974	Statistics Denmark
The Population's Education Register	Information on individuals' highest completed education	1981	Statistics Denmark
The Employment Classification Module (AKM)	Information on attachment to the labour market at a given moment or throughout the year	1976	Statistics Denmark
The Integrated Database for Labour Market Research	Information on persons and establishments and their relation	1981	Statistics Denmark
The Income Statistics Register	Includes anyone who is economically active in Denmark. Variables describing wages, entrepreneurial income, taxes, public transfer payments, public pensions, capital income, private pension contributions and payouts, home ownership and fortunes	1970	Statistics Denmark
The Building and Housing Register	Information on ownership, type of housing, rental terms, living area, number of rooms, condition concerning bathroom, kitchen and toilet, and year of construction	1880/1981	Statistics Denmark

IDA

- A frequently applied research database is the **Integrated Database for Labour Market Research** (IDA)
- Created to solve a difficult problem of definition: Identity of enterprises over time, a task that individual researchers were unable to handle for reasons of both time and funding
- Ten man-years were spent on the development, which was jointly funded by the Danish Social Science Research Council and DST
- Since the establishment of IDA, Statistics Denmark has handled the updating of the database financed by user fees
- IDA contains information on the total population of people and enterprises in Denmark from 1980 and onwards

Data available for researchers

- All variables have an affiliated quality declaration
- The declarations describe quality and content of data (in Danish!)
- Data from DST can easily be linked to data from other sources, like survey data (SHARE, ECHP, LFS) or data from other governmental agencies (e.g. Health Authorities; Twins Registry)
- Links with data external to DST are carried out under the approval by the Danish Data Protection Agency
- Each researcher pays the costs of data production/updates (based on # variables; #years). About € 10K per year for a labor economist

Rules for researchers´ access to micro data

- Access to micro data can be granted to researchers from **research environments pre-approved** by DST.
  - Universities
  - Ministries
  - Sectoral institutes
  - Non-profit foundations
  - Private consultancies
  - NGOs
- DST will evaluates the proposed organization carefully, especially rigid with private sector ones
- DST will not grant authorization to single persons, single persons are authorized by their own environments
- Access is given according to a so-called “need to know”-principle: can only get access to the data needed to fulfill their research purpose

Foreign researchers

- Foreign researchers are given access to micro data through an affiliation to a Danish authorised environment.

- Affiliation can only take place if the authorised environment is willing to take the responsibility for the foreign researcher making sure that all existing rules governing access to micro data are observed.

- The research environment must also appoint a contact person who will undertake the responsibility for all contact between the foreign researcher and Statistics Denmark.

Working with micro data on the research servers

- All access to micro data for research is given through a research server placed at DST
- Note: *placed at* does not mean *owned by*.
- The vast majority of research environments buy their own servers and place them at DST
- This reduces congestion by order of magnitudes
- Servers are maintained by DST personnel and research institutions pay maintenance fees
- Research servers are separated from DST production network and only contain de-identified micro data for research purposes.

Working with micro data on the research servers

- When research data for a project have been prepared by DST's Division of Research Services, they are transferred to the research server where remote access is given via the Internet
- The researcher has to sign an agreement with DST before remote access is granted
- States that no attempts to identify people or enterprises – or to export micro data must be made and is considered a very serious breach of the agreement between the researcher and DST
- All results are sent to the researchers automatically by e-mail. This is a continuous process (every five minutes) and has shown to be quite effective [but more restrictions recently]
- Violations are punished with bans to the whole environment of the infringing reseracher
- Example of violation: sending out a log listing individuals observations or statistics based on few units

History of access conditions

- 1986 DST on-site (Copenhagen) access for researchers
  - ”Need to know” principle for access to anonymized data based on a project description
  - Trade-off of # variables & sample size
  - Researcher signs confidentiality agreement
- 1995 DST Århus office opens
- 2001 Own workplace access once the analysis environment authorized
- 2003 Remote access by Danes from abroad via Danish analysis environment authorization (and from home for Danes in Denmark)
- 2007 Foreigner remote access from abroad via the same
- 2018 status – 250+ environments & 2000+ researchers

An illustration from own ongoing research

On the origins of income inequality: evidence from children of twins

Paul Bingley
Lorenzo Cappellari
Konstantinos Tatsiramos

Questions

- Nature vs nurture in shaping outcomes
- Differential influence within vs between generations

Research design

- Existing studies use twin designs with information on zygocity to separate nature vs nurture
  - Within a generation
  - Common environment for MZ and DZ

- Children of Twins (CoT): data on twins (with zigocity), their children and their spouses
  - Both within and between generations
  - Relax common environment assumption

- Essentially: CoT allows observing income links with someone that is genetically identical to the father, but is not the father

Data

- DST registers
  - Income
  - Familiy links
  - Education

- Twin register

- Consider nature vs nurture in income and years of education
- Approx. 2000 twin families

Decomposition results

	Permanent incomes		Years of education	
	Coeff.	S.E.	Coeff.	S.E.
IGE (β)	0.1829	0.0328	0.2387	0.0156
Share pre-birth	0.6483	0.4633	0.4814	0.2810
Sibling correlation DZ (ρ_{DZ})	0.2766	0.0601	0.6577	0.0448
Share pre-birth	0.4514	0.3274	0.1759	0.1023
Sibling correlation MZ (ρ_{MZ})	0.4286	0.0577	0.6825	0.0525
Share pre-birth	0.3778	0.2807	0.1927	0.0892