Swedish register based research in the social sciences

A practitioner's view

Martin Hällsten, PhD

SUNSTRAT, Stockholm University

Overview

- Definition
- Background
  - Development of register based research in the social sciences
- Registers in the social sciences
  - Brief outlook to medicine and public health
- One (1) example
- Experience of the MONA-system
- Practical problems
- Pros and cons for the future

Definition

- What is register data?
  - By-product of registers held for administrative purposes
  - Dedicated for governmental planning (and research)
  - Population level surveys (e.g., censuses)
- What register data is not
  - Small scale surveys, labour force surveys

Background

- Personal IDs (1947)
- Extensive book-keeping within Swedish organizations
  - Since 17th century (military planning)
  - Automatic data processing introduced in 1968 circa
- Legislation allows registers to be matched if ethically approved and data security can be maintained
  - Always anonymized data (individual, workplace, organization)
  - Owner of register must approve (mainly data security)
  - All projects must under go ethical vetting (informed consent or not)

Old tradition in Swedish social sciences

- Register studies older than e.g., survey methods
  - Birth records and national registration used early for demographic research (e.g., Commission on emigration 1913)
- Principle to keep acts on individuals born on the 5th, 15th and 25th each month
- 15th-born register
  - Created from 1950 census
  - Sampling frame for surveys
  - Linked to tax authority income data (1951-)

Old tradition (cont'd)

- Gunnar Boalt (1947) merged pupils' school records to administrative registers
  - The (Swedish/Nordic) principle of public access to official records
- Torsten Husén (1950) combined a school survey with military enlistment data
- Gösta Carlsson's (1958) study of social mobility combined 15th born register with census data + birth records
- Project Metropolitan (1964-): school survey with extensive follow up through registers (until 1986)
  - Data security and integrity (Data inspection board conflict)
- Swedish Level of Living Survey (LNU) based on 15th born register (1968-), survey matched to taxed income data

Key developments at Statistics Sweden

- Packages, pre-matched datasets for researchers (with documentation)
  - LINDA database (1994)
    - 3 % of population merged with tax income records back to 1968
  - LOUISE (later LISA) created in mid 1990s
    - Full population 1990 and onwards

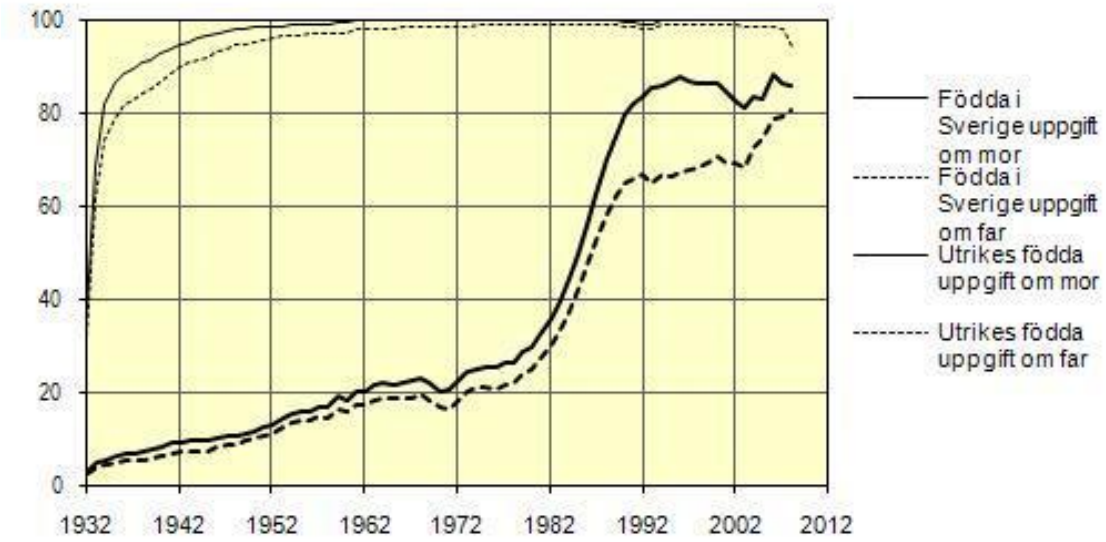
Core registers

- Background data
  - Date of birth, county, gender, birth country
- National registration (1968-)
  - Does not identify apartments, families identified by multigenerational links
- Income and taxation register, IoT (1968-)
  - Increasing level of detail over time, many components
- Multigeneration register (2000)
- Centennial censuses (1960-1990)
- Education register (1985-)

Multigeneration registret (2000)

Parents linkages for children born 1932 and onwards

- Complete for children born 1968 and onwards
- Biological and adoptive links
- Mother's and father's birth country



Censuses (1960-1990)

- Apartment ID = valid household definition
  - Family type
  - Region
- Occupation consistently coded (social class etc.)
- Education
- Environmental indicators (overcrowding, apartment size)

School registers

- Compulsory school (1988-)
  - GPA, school
- Upper-secondary school
  - Applicants (1971-): chosen track, admission GPA
  - Graduates (1973-): GPA, track, school
- Tertiary education
  - Enrolees (1977-), type of program, university
  - Graduates (1962-), type of degree, university
  - Applicants (1995-) , type of degree, university
- Adult secondary education (1990s-)

Education register (1985-)

- Aim: create information on highest achieved education for every individual
  - Hotchpotch of data sources
  - Primarily school registers
  - Survey to immigrants
- Updated annually (data is not corrected backwards)
- Data sources continually added
  - Major update in 2000
- Caveat: looks neater than it is…

Earnings structure register (1970s-)

- Official wage statistics collected by Statistics Sweden via trade unions and employer federation since 1950s
  - Digitized from 70s and onwards
  - Included in EU's Structure of Earnings
- Total sample for public sector and large private firms
  - Private firms <500 employees are sampled
  - 2.2 million cases annually
  - Monthly work hour standardized wage + occupation code

Matched employer-employee data

- LISA (1990) + Employment register (1985-89)
- Identifiers of organization and workplace
  - Status in november (1985-)
  - Largest contritubor to individual wage (1990-)
    - Basis for aggregation: % female in workplace, contextual effects
- Industry and sector, type of ownership
- Characteristics of private firms
  - Value added, profits, capital

Geographic data

- National records
  - County/Municiplatiy/Parish
  - Migration history (year + month)
- Geography register (1990-)
  - All property geocoded
  - SAMS (Small Area Market Statistics) = neighborhoods (~ 1,000 inhabitants)
  - 100 × 100 m squares
  - Special needs
- Local labour markets (clustering of municipalities)

Important auxiliary data

- Emigration and immigration
- Internal migration
- Civil status changes
- Deceased
  - All register events (year and month)

Recents advancements

1. Occupation register (2001-)
   - Various sources, mainly employer reports
2. Apartment register (2011?-)
   - Avoids use of multigeneration register to define households (cohabitants without children)

= Register based census

External registers

- National council for crime prevention
  - Suspicions via police register (1991-), events
  - Convictions via legal system (1973-), events
- Swedish Public Employment Service
  - Unemployment spells w/ detailed information (1992-)
- Department of applied educational science, Umeå
  - SweSAT (Swedish Scholastic Aptitude Test)
- Board members of private firms

Enlistment data (1968-)

- Mandatory enlistment for men until recently
  - Length, weight
  - Physical capacity
  - Cognitive ability test (IQ): Reasoning, Verbal comprehension, Spatial ability, Technical understanding
  - Non-cognitive skills rated by psychologist (see Mood et al 2011)
    - Social maturity
    - Intensity
    - Psychological energy
    - Emotional stability

Medicine and epidemiology

- National registers (The National Board of Health and Welfare)
  - Cause of death, cancer, hospitalization, medical birth records, drug prescriptions, social service, etc. http://www.socialstyrelsen.se/register
- Vast array of local 'registers'
  - 89 quality registers (treatment for specific diseases) http://www.kvalitetsregister.se
  - Biobanks (research projects)
  - Twin register
  - Primary care registers

Swedish register based research example: Almond, Edlund & Palme

- Prenatal exposure to Chernobyl fallout
  - Cluster municipalities based on fallout, recode birth dates to reflect 8-25 weeks gestation
  - Differences-in-differences methodology
- Worse performance in secondary school, mathematics in particular
- No corresponding damage to health outcomes
- Cognitive ability is compromised at radiation doses currently considered harmless

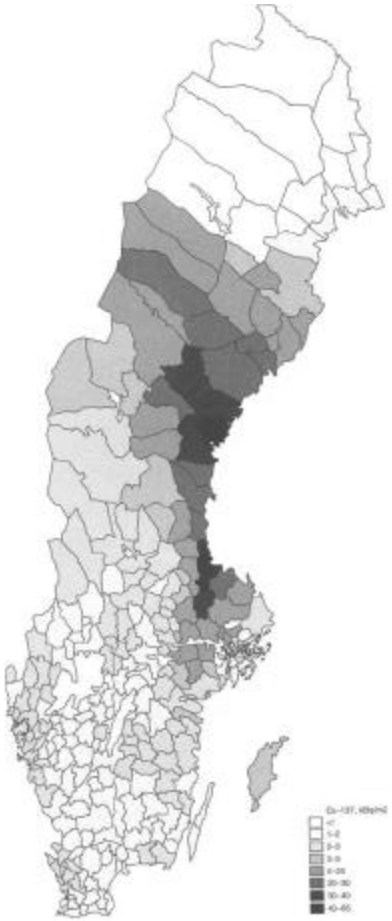


FIGURE II
Cesium-137 Ground Deposition in kBq/m^2 by Municipality

Almond, Douglas, Lena Edlund, and Mårten Palme. 2009. "Chernobyl's Subclinical Legacy: Prenatal Exposure to Radioactive Fallout and School Outcomes in Sweden." *The Quarterly Journal of Economics 124:1729-1772.*

Experience of the MONA-system

- Easy to use standard software – plug and play
  - Batch server for heavier workloads
- Connected to SCBs data warehouse
  - Little data administration
- Still some teething problems (after 5 years)
  - Sensitive to high workloads (conference deadlines…)
- Special software not always compatible
- Conflict with other data providers/owners (National Board of Health and Welfare)
- Conflict with `weaving' – integrating analysis code and text into one document for reproducible research

Problems with register based research

- Many registers *not* designed for research
  - Researchers passive users, data generated by bureaucrats
  - No established ways of feedback, cooperation or influence
    - Large potential for improvements
- Documentation = detective work
  - Good quality for *packages* such as LISA, else generally low quality
  - Sometimes documentation is completely missing (archive digging or assumptions…)

Problems (cont'd)

- Data retrieval only possible for specific research projects
  - What is a project?
- Statistics Sweden charges for every project
  - Data base build up cost-efficient but not allowed
  - Statistics Sweden is a monopolist
- Increasing administrative burden of ethical vetting

Pros and cons for the future

- Cost-efficiency of register based social science research may move resources away from social surveys
  - STAR data (1.5-2 million SEK) vs. LNU 2010 (20 million SEK)
  - Many aspects of human life cannot be followed through administrative registers
- Follow up through registers less costly
  - Project Lifegene (biobank + survey, 500,000 individuals) uses informed consent to follow individuals through registers *prospectively*