Machine Learning for Economic Demography: Data and Challenges

Bernhard Rengs[1]

[1] Vienna Institute of Demography, Austrian Academy of Sciences, 1010 Vienna, Austria

Make it (Net)Work! - MLA2S Networking Seminar #10
Vienna, 09.06.2026

Data in Demography

- Data = Datasets of individuals, households, firms
  - Legal aspects of data security often an issue / a topic
  - Each datapoint is a dataset consisting mostly of these data types
    - Categorial (sex, level of education, SRH, etc.)
    - Numerical (age, income, etc.)
    - Sometimes textual (“open”) -> to be categorized later
  - Different research questions will use different subsets of the data
- Different datasets cover different aspects
  - In addition to basic demographic aspects
  - Labor market or Family / Fertility or Health related, …
  - Often the are the result of surveys
    - participation is time consuming – number of questions limited (trade off)
  - Often these are not linkable Labor market <> health, though some are

Dataset sizes in Demography

- Very different dataset sizes
  - Sometimes aggregate data
  - Often rather small samples of data (n~500) available
  - Often larger samples are available
    - Eurostat Share (n~160.000 – 28+ countries)
    - German SOEP (n~120.000 panel)
    - E.g. Austrian micro census (n~32.000 p.a.)
  - Sometimes full demographic data (n=N)
    - E.g. Austrian Microdata Center (residential records, employment records, ..)
    - Denmark, Sweden, Norway NSI's microdata -> Pioneers
  - Simulated data from agent-based methods, microsimulations, ABM
    - May be huge -> alternate realities (stochastic + parametric)

Data gathering and accuracy

- Data from questionnaires -> self-reported information
  - Possible self-reporting biases / cultural biases
  - (Individual) errors while entering data / mistyping, unreadable (PAPI)
- Data from (real / old) censuses
  - Partial self-reporting
- Data from current censuses
  - Register counts – i.e. information extracted from official registers
  - Potentially less information, but more accurate
- Data from full demographic data
  - Which data is available / allowed to be linked
  - I.e. Linking basic register data with social insurance data allowed?
  - Is really all data as accurate as expected – or is some data imputed?
    - Imputed data often sufficient to analyze the population as a whole
    - But – is it still usable for multivariate analyses of individuals? (multidimensional vs. marginal distributions)

Issues with questionnaire data

- Survey mode (PAPI, CAPI, CATI, CAWI) may have an impact on answers
  - Potential interviewer bias (personal interviews)
  - Unsupervised surveys
    - Higher item non-response (missing attributes)
    - Lower cooperation rate (less returned / completed questionnaires)
- Often survey participants truly anonymous or data anonymized
  - Multiple participations of same individual? (only for anonymous)
  - Can‘t directly link participants data with other data sources
  - Not straightforward to link participants data with responses in other waves (only probabilistic)
- Representativity of sample?
  - Do the participants follow the required multidimensional distributions?
  - Is the general multidimensional distribution of population known?
  - I.e. can we calculate weights

Temporal aspect

- Many research questions require multiple datapoints in time
  - For the same individuals / households, etc.
  - For all the important attributes of these individuals at the same time
- Some surveys gathering special data have only one wave
- Panel surveys (are planned to) have multiple waves
  - Are there enough panelists or too many refreshers?
  - Are the relevant questions in every wave / since when?
- For register data
  - When did the register start (often after y2000)
  - Are all relevant data available since the start?
- Ofc. no ex-post gathering of data not recorded in the first place
  - Though digitalization of older diverse sources sometimes possible

Access to data

- Specific surveys – data sometimes not available only results
- Many surveys / Most panel surveys
  - Data access for accredited research project and or payment
  - Data will likely be able to be analyzed directly (on-premises) (GREAT)
- Full datasets (e.g. AMDC, etc.)
  - Most of the time data may not leave the servers of the data provider
  - Access only to virtual desktops through some RD via encrypted vpn
  - Very often computational capabilities on these ridiculously low
    - E.g. AMDC 16GB Ram, 50GB Storage, 2(!!!) CPU cores, no GPU
    - Light upgrades only with additional non-negligible monthly costs
  - No in/out data transfer (including scripts, programs) without manual checks
  - Quite challenging to employ modern methods

Employing AI and ML methods in demography

- In many samples, minorities and vulnerable groups are under-represented, but may be especially relevant
  - A problem for every analytical technique
- Many samples are too small or have too many dimensions
- Black-box AI methods do not produce the required explanations of individual behavior (causal relationships)
  - Might still be used when/if only aggregate forecasts are relevant
- Big empirical datasets are only available in environments that don‘t support calculation intensive methods
  - Though simulated datasets (following empirical distributions) may be simulated and analyzed on on-premise servers and clusters
- Still a number of ML methods and AI support of many steps of the research process are relevant for demographic research

Wittgenstein Centre
FOR DEMOGRAPHY AND GLOBAL HUMAN CAPITAL

VIENNA INSTITUTE OF DEMOGRAPHY

Thank you!

Make it (Net)Work! - MLA2S Networking Seminar #10
Vienna, 09.06.2026