Session 3: Methodologies for big data

Overview of new approaches on the topic and discussion

Piet J.H. Daas Statistics Netherlands / Eindhoven University of Technology

6 December 2022

Comment on papers

2 papers that deal with:

-Use of texts for official statistics

i. Enterprise characteristics

ii. Social Mood on Economy index

- Use of images for official statistics

iii. Land use and maps (LandCover project)

- Under development

iv. TERRA experimental statistics

i. Enterprise characteristics

- Nice overview of total process and various processing steps & options
  - Aim produce statistics on website related activities of enterprises
  - Often a ML classifier is used to produce BD-based results
  - Comparing BD-results with survey data based findings are usually very similar
  - Sometimes a correction is needed, how?
    - Via auxiliary variables included in both datasets
    - More complex approaches are investigated (next slide)

i. Enterprise characteristics: SN perspective

- Webpages are a great source of ‘data’
  - But have much more potential than only providing Enterprise characteristics
  - For instance: to create ‘enriched’ subpopulations (inovation, platform, ..)
- Bias in ML: ML-classifiers as an example (for inspiration)
  - Such classifiers are biased as they are affected by the ratio of Positive and Negative examples used in training set
    - Ratio used in Training(test)set vs. ratio occurring in real world data
      - Correction method available: Puts and Daas 2020*, BayesCCal.py on Github (mputs)
- Missing validation results
  - How well is the coupling between url and enterprises? And for the different units in Business Register? (in other referred to papers?)
  - Actual accuracies of survey- and web-based results? And more..

*https://arxiv.org/abs/2102.08659

ii. Social mood on the Economy index

- Based on daily samples of Tweets (~47,000)
  - Keyword based selection (what are the words?; is there also a black list of words?)
  - Highly dependent on Twitter (what if Twitter stops?)
  - Daily mood on the economy, published Quarterly (Ehm?)
- Unsupervised lexicon based sentiment classification
  - Positive, Negative, Neutral (percentages?)
  - Daily index is derived from the Pos. and Neg. clusters obtained (intriguing, how exactly?)
- Anomaly detection approach (to deal with off-topic issues)
  - To deal with unforeseen circumstances (I assume Internet hype related stuff, Great!)
- Various new developments:
  - Evaluate quality of filtering (WordEmb, it might work, try and see!)
  - Improve interpretation (but the index will remain ‘volatile’, aggregation is the easiest way to deal with that, alternative is applying a median filter)
  - Topic analysis (ehm I’m not sure, Wordclouds do provide info)
  - Create a more focussed/specific indicator (concept related, next slide)

ii. Social mood on the Economy index: SN persp.

- Challenging to produce high quality social media based statistics/indicators
  - Highly depended on filter words used and population dynamics
  - Usually social media indices are message based (not population based)
- **What is the concept measured?**
  - Need to check that. Validation is required to make sure that the indicator is measuring what it is supposed to be measuring?
  - The index is certainly affected by COVID (check which events occur as the index changes)
  - An extensive study has been performed for the Social Unrest Indicator at Statistics Netherlands (in Dutch)
- Beware that an experimental statistics is not an official product (yet)!
  - For the latter quality demands are higher! (concept, population related)
  - Could be a potential problem for this index
    - Are there any plans to make the Economy index official?
    - Beware of pitfalls and the bad experiences at Stat Netherlands when dealing with these kind of indicators

iii. Land use and maps (LandCover project)

- Automatic land cover/use *estimation* system
  - Reduce human interaction
- Highway and water affect the analysis in a negative way
- Very nice work on dealing with data in various resolutions and combining sources (very ‘tech’ oriented)
- End solution is great but *very* computational intensive
  - Extremely detailed, the work is clearly technology driven (keep main objective in mind; I thought it was about estimating land use?)
  - Nice map of Pisa area is obtained after 3 days of calculations (and what about the whole country?)
  - Validation? (diff. Resolutions, diff. Classifications)
    - There must be a single Ground Truth
      - Good idea to use admin data as input here
      - Project would have benefitted from this early on!

iii. Land use and maps: SN perspective

- DL models have an extremely high accuracy
  - But results on new data they may not be that high
  - Are you sure the approach works on other (unseen) parts of Italy?
  - Is there a computational less intensive approach that could be used as an alternative? (what is the current accuracy?)
    - Has this been determined?

- ‘Black box’ approaches and official statistics?
  - Could be an issue (transparency; see next point)
  - Really need to validate DL-based findings
    - In this case, there is a Ground Truth!
    - Any validation checks done?
    - Somehow need to relate/unify classifications used in sources used

iv. TERRA experimental statistics

- New cloud-based application based on open data (Eurostat)
  - Still under development (end of 2022)
  - Focussed on velocity, getting data in as quickly as possible
- Dashboard, to observe shocks in transport/trade relations
  - Includes a combination of R and Python (why?)
- Input is Eurostat COMEXT data put on website
  - Enables detailed comparisons over time
  - Updated every month
  - Relation with Google mobility data (this will soon be lost!)
- Nice example of BD-based dashboards!!
  - Will assists users and annalists
  - What is the initial reason for developing the application?

• iv. TERRA experimental statistics: SN persp.

- Corona initiated a need for more dashboards at SN
- During corona a number of statistics were produced at:
  - A higher frequency and were made available must faster
  - Mainly achieved by just speeding up the internal process, not by using more rapidly available (Big)data (What's the urge for TERRA?)
- Certainly helps in providing quick insights
  - Users and annalists will be very happy with that
  - Examples of its use? (user stories?)
- What about using this kind of approach for (combinations of) ISTAT data produced every month or quarter?
  - Support users of ISTAT data by enabling a better grip on (part of) the output). Such as economic data, …

General remark

- The papers describe very interesting and relevant work but lack detailed info on the validity of the BD-based findings
  - This is a major topic at SN!
    1. What is the *concept* measured by BD-based statistics?
       - Is it still the concept that was initially intended to be measured?
    2. *Population perspective* on BD-based statistics
       - Comparing/combining with representative survey/admin data
       - Get grip on population in BD-source (very challenging)
    3. *Comparability over time*
       - How reproducible are the BD-based findings?

Validation is essential for any new official statistics produced (and for old statistics)

- Required for an official (new) BD-based statistics

Methodology of Big Data for official statistics

- Important topics regarding Machine Learning (AI) and Official Statistics are described in a paper in the Survey Statistician*
  - Methodology concerning the human annotation of data
  - Sampling the population to obtain representative training sets
  - Using stratification in the context of Machine Learning
  - Data structure engineering and selection to increase the transparency of models
  - Reducing spurious correlations
  - Methodology for studying causation
  - Correcting the bias caused by ML models
  - Dealing with concept drift (comparability over time)

*Puts, M., Daas, P. (2021) Machine Learning from the Perspective of Official Statistics. The Survey Statistician 84. pp. 12-17.

Questions?