Fourth Workshop on Methodologies for Official Statistics

Discussion of papers in Session 2 : AI/ML for Official Statistics

Brunero Liseo

December 1st, 2025

Sapienza Università di Roma
brunero.liseo@uniroma1.it

Official Statistics is dramatically changing

The last generation of AI/ML Learning algorithms allows

- huge availability of data, sometimes in real time
- great computational capabilities
- speed in synergistically integrating information from different sources

These new potentials are inevitably re-shaping the figure of the (Official) Statistician in terms of

- Methodologies
- Targets
- Perspectives

Language Models for Automated Coding in Official Statistics: Risks and Opportunities

Bruno, Di Zio, Ortame

- The use of LLM to help in classifying economic activities into the ATECO categories (1,200 more or less..)
- Benchmark: CIRCE Thesaurus, manually updated every year.

The goal is to consider a natural language query and provide a (as small as possible) set of potential ATECO categories

Several techniques are used

- `Cosine-based retrieval`
- `Linear Discriminant Analysis`

Conformal Prediction techniques are then used to guarantee exact coverage of the produced subsets

Some comments

- The advantage of a dynamically updating classifier over a clerical maintained thesaurus (CIRCE) is obvious.
- Why

$$\text{score}(D_j) = \max\left(\text{sim}(\pi(x_i, e_{1j}), \ldots, \pi(x_i, e_{Mj})\right)?$$

  - What happens if some of the sim are very low?
  - Maybe better to average?

Conformal Prediction

- The *Adaptive Prediction Sets* (APS) works better than the standard Split-CP. This is also called the Cumulative Likelihood Approach, that is, one greedily includes classes until the true label is reached (better UQ!)

In classification problems, it might be useful to consider two variants of CP

- Selective Classification

  This is useful when we only want to show a model's predictions when it is confident.

  For example, we may only want to make medical diagnoses when the model will be 95% accurate, and otherwise to say *Sorry, I do not know . . .*

  More formally, we want to guarantee a conditional coverage, restricted to those cases where the evidence of a specific classification is strong

- Conformal Risk Control See for example Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2022). Conformal risk control. arXiv preprint arXiv:2208.02814.

Conformal Risk Control

CP provides a guarantee against miscoverage, that is

$$\mathbb{P}(Y^* \notin \mathcal{C}(\mathcal{X}^*)) \leq \alpha$$

- This is equivalent to controlling the expected value of a particular loss function

$$\mathbb{E}\left[\mathbb{I}\left(Y^* \notin \mathcal{C}(X^*)\right)\right]$$

- One can choose a different loss function and minimise, more generally

$$\mathbb{E}\left[\ell(\mathcal{C}(X^*), Y^*)\right]$$

For example, one might want to minimise the False Negative rate

$$\ell(\mathcal{C}(X^*), Y^*) = 1 - \frac{|C(X^*) \cap Y^*|}{|Y^*|}$$

- On the use of mobile network (MNO) data to anticipate Tourism estimates
- Results: MNO data generally overestimates the number of overnight stays.
- However, MNO data underestimates in the seaside area.
- This behaviour might be motivated by the different types of non-resident presences in the two sub-regions.

Summary

- MNO data are debiased
- three approaches are compared
  - Augmented Learning - AL (using MNO data as a proxy)
  - Quasi-Transfer Learning - QTL
  - Traditional Approach, based on RF
- Final message: adding an extra step, which combines AL and QTL, seems the best solution
- Uncertainty Quantification via Conformal Prediction, using a sort of jackknife approach

Conformal Prediction Intervals

- An alternative, which takes into account the temporal dimension of the dataset, is to Adapting Conformal Prediction to distributional shift (Gibbs & Candes, 2022).

  Instead of fixing α, we learn a sequence of α_t, by online updating, based on

  $$\alpha_{t+1} = \alpha_t + \gamma(\alpha - \mathsf{err}_t),$$

  where err_t acts as an unbiased estimate of the current miscoverage probability

  $$\mathsf{err}_t = \begin{cases} 1 & y_t \notin \widehat{C}_t \\ 0 & y_t \in \widehat{C}_t \end{cases}$$

  and γ is a sort of learning rate.

  This procedure adapts to the observed frequency of miscoverage, increasing or decreasing α_t in response to over-covering or under-covering events. The update increases α_t if the prediction set is too conservative and vice versa.

  Under no assumptions, one can prove that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathsf{err}_t \overset{a.s.}{=} \alpha$

- A thorough investigation into how Generative AI affects – and will affect in the future – official statistics.
- I have a fundamental agreement with the point of view developed in the work.
- So I will only emphasise the most significant aspects.

- *Statistical organisations report particularly high levels of concern regarding confidentiality, data security, and the factual accuracy and consistency of AI outputs. Further concerns relate to potential misuse, ethical issues, dependency on external AI providers, and negative public perception.*

  This is a severe problem, especially for NIS, which should always maintain an image of impartiality and scientific authority, providing better and more robust answers than others!
- *Integrating generative AI systems with existing data pipelines and production frameworks while maintaining reproducibility, traceability, and quality.*

  This is a more technical issue, but still relevant. AI systems are sometimes **black boxes**, full of acronyms, even for experts in that field, and the results obtained are often inadequately argued, with potentially critical assumptions not being made sufficiently explicit.

- *As a result, reproducibility in AI-assisted statistical production cannot be limited to re-running a single model [...]. It requires systematic capture of model versions, configuration settings, prompt templates, retrieval mechanisms, and any post-processing logic, as well as clear recording of when and how humans interacted with the system.*

This is the key issue, which opens the way to the paper's main thesis: moving from a model-centric to a system-centric philosophy.

In this sense, the Swedish experience of the public-private joint project MLOps is very innovative and important for NIS

What is MLOps?

Machine Learning Operations (MLOps) are a set of methods, tools, and processes that aim to make machine learning development more robust and scalable, with a higher degree of automation and collaboration across different departments and disciplines.

From their website:

Challenges

The technical challenges encompass changes in both data and models. Key questions include:

- Will new data interact with a current model in the same way as existing data?
- Will an updated model interpret data consistently with its predecessor?
- How can models be effectively retrained with new data?

From an organizational perspective, key considerations include defining necessary roles and responsibilities, establishing processes to guide AI toward appropriate objectives, and determining prerequisites for new projects, data utilization, and technology adoption.

Do LLMs understand language?

There is no clear answer. Wittengstein said:

Comprehension is measured by evaluating how language is used. If I use words correctly, if I am understood, if I can navigate everyday situations, it means that I am using language well, and that is the test for understanding that I understand

Next challenge:

- Causality
- Semantic

The human mind learns through multisensory experiences, deep insights, and complex mental models, going far beyond mere data analysis.

The new frontiers of AI are inspired by the ability to explore abstract concepts and causal reasoning mechanisms for creating intelligent and adaptable machines.