UNECE High-level Group for the Modernisation of Official Statistics

Beyond Point Predictions: Ensuring Reliability in Official Statistics through Uncertainty Quantification

This business case was prepared by Mohammed Haddou, and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New activity	☐	Extension of existing activity
Proposed Modernisation Group(s) for Activity			
☒	Applying Data Science and Modern Methods	☐	Blue Skies Thinking
☐	Capabilities and Communication	☐	Supporting Standards
☐	*Other:*	*Unknown*	

Purpose

National statistical offices are increasingly using machine learning and deep learning models and algorithms to tackle a variety of complex tasks. To effectively apply these models to real-world prediction problems, **it is important to be able to quantify the uncertainty of the model's predictions**, typically through the use of prediction intervals/sets. This allows us to not only make a prediction, but also to understand the level of confidence we can have in that prediction. While these models may achieve high levels of predictive accuracy, quantifying their predictive uncertainty can be challenging. To accurately estimate predictive uncertainty, the prediction intervals/sets should cover the true prediction targets with a high probability and be able to discriminate between instances of high and low confidence predictions. Additionally, these models are known to be ***overconfident***, meaning they often attribute low uncertainty to their predictions even when uncertainty is actually high.

Uncertainty quantification (UQ) methods have been proposed in the literature as a potential answer to reduce the raw decision provided by the deep learning black box and thus increase the **interpretability** and **acceptability** of the result by the end user. This field has developed considerably over the last five years or so. Enhancing an automated prediction with an estimation of its confidence has many advantages. **First**, it allows the identification of uncertain samples that need *human reviewing*. **Second**, it enables the identification of the model's pitfalls. For example, unconfident predictions may indicate an incomplete training dataset. It provides insights into the knowledge captured by the model and can be used to extend the training set with supplementary data, if needed. High uncertainty can also reveal anomalies within the input data, which is **critical for *Quality Control*** (QC). Overall, **UQ increases trust in the algorithm**, and facilitates the interaction between the algorithm and the user and therefore is a **critical requirement for *trustworthy ML*.** The task team might want to place special emphasis on *conformal prediction techniques*.

Conformal prediction (CP) is a set of algorithms devised to assess the uncertainty of predictions produced by a ML model. It is a straightforward UQ technique to generate prediction sets **for any model**. It is a user-friendly paradigm for creating **statistically rigorous** uncertainty sets/intervals for the predictions of such models. Critically, the sets are valid in a *distribution-free* sense: they possess explicit, non-asymptotic guarantees even without distributional assumptions or model assumptions. One can use conformal prediction with *any pre-trained* model, such as a neural network, to produce sets that are guaranteed to contain the ground truth with a user-specified probability, such as 90%. It is easy to understand, easy to use, and general, applying naturally to problems arising in the fields of computer vision, natural language processing, deep reinforcement learning, and so on.

When the cost of making a wrong prediction is very high, it is important to understand the reliability of the algorithms we use. A very good way of understanding the reliability of the uncertainty of future predictions is to be able to return prediction intervals/sets. Assume we have training data $(X_1, Y_1), \dots, (X_n, Y_n)$ and a test point $(X_{n+1}, ?)$ for which we would like to predict the label Y_{n+1}. The data are assumed to be exchangeable (weaker assumption than i.i.d.) from some distribution P_{XY}. Conformal prediction allows us to construct a marginal distribution-free prediction interval $\Pr[Y_{n+1} \in C(X_{n+1})] \geq 1-\alpha$ for any unknown distribution P_{XY} and any sample size n.

Survey Sampling Context: Conformal methods are an active research topic in statistics and machine learning. However, it is only recently that they have been extended to non-exchangeable data. The task team might want to investigate how conformal prediction, for instance, can be applied to data from *sample survey designs* within a framework of design-based inference for a finite population.

Description of the activity and deliverable(s)

Objectives and Deliverables

- **Literature Review:** Conduct a comprehensive literature review of current research on Uncertainty Quantification (UQ) methods, with special attention given to Conformal Prediction, among other relevant topics.
- **Use Cases (optional):** Gather use cases demonstrating NSOs experiences in using UQ methods with real-world data.
- **Report and Recommendations:** Summarize the findings in a working document, providing clear and practical recommendations for ML practitioners. These recommendations should focus on how to effectively apply such methods in the context of official statistics and to ML algorithms developed at NSOs.
- **Concrete outcomes/deliverables.**
    - Report with practical recommendations and guidelines.
    - Python/R scripts showing how one can implement UQ methods
    - Knowledge sharing – Presentation to conferences

Alternatives considered

Traditional ML approaches lack reasonable uncertainty quantification guarantees. Class probabilities (scores), Bayesian posterior predictive intervals, or the bootstrap, for instance, provide heuristic notions of uncertainty but do not offer guaranteed coverage. Additionally, they rely on simplified distributional assumptions. In contrast, Conformal Prediction provides rigorous coverage guarantees, is model-agnostic, distribution-free, and enables the derivation of rigorous finite-sample prediction intervals/sets. Furthermore, Conformal Prediction is applicable to various tasks, including classification, regression, time series forecasting, segmentation, object detection, multi-label classification, hierarchical classification, nucleus sampling for large language model (LLM) text generation, and more.

How does it relate to the HLG-MOS vision and other activities under the Group or HLG-MOS?

Uncertainty quantification is a **critical requirement for trustworthy ML.** This essential activity seamlessly integrates with the overarching principles of responsible AI/ML and quality assurance/quality control (QA-QC). The presented business case strongly aligns with the vision and values of HLG-MOS, specifically aiming to instill trust in the accuracy of published official statistics.

Proposed start and end dates	
Start: January 2024	**End: December 2024 or June 2025**