UNECE High-level Group for the Modernisation of Official Statistics

Business Case for Machine Learning

This business case was prepared by Barteld Braaksma (Statistics Netherlands) based on the position paper 'The use of machine learning in official statistics', and is submitted to the HLG-MOS for their approval.

Type of Activity			
☒	New project	☐	New activity
☐	Extension of existing project	☐	Extension of existing activity
Projects are undertaken by separate project teams. Projects are expected to produce a significant contribution to achieving the HLG-MOS vision		*Activities are undertaken by Modernisation Groups. These activities produce smaller, more detailed outputs to help achieve the HLG-MOS vision*	

Purpose

The interest in the use of Machine Learning (ML) for official statistics is rapidly growing. For the processing of some secondary data sources (including administrative sources, big data and Internet of Things) it seems essential to look into opportunities offered by modern ML techniques, while also for primary data ML techniques might offer added value, as illustrated in the ML position paper mentioned above. Although ML seems promising there is only limited experience with concrete applications in the UNECE statistical community, and some issues relating to e.g. quality and transparency of results obtained from ML still have to be solved.

Description of the project

The project is divided a priori into six work packages (WPs). If the project materialises the project manager and project team may decide that a slightly different layout is preferable, but the key elements will remain.

WP1. ML pilots. Based on mutual interest of the WP1 members, a number of concrete pilots (typically three to five) for ML in specific domains will be covered. Example domains, for which already interest has been expressed, are image recognition, websites analysis and automated coding; further areas can easily be identified. A pilot may either relate to common statistics production using e.g. survey data or alternative approaches using e.g. big data. In each pilot study at least two countries and/or international organisations collaborate. A pilot should preferably be comprehensive and practical, that is: develop the appropriate methodology, build CSPA-compliant tooling, test on realistic use cases and consider quality and transparency aspects.

WP2. Inventory of existing ML projects. An inventory of on-going and finished ML projects carried out in the statistical community will be developed, based on a survey addressing the international statistical community. The goal is to learn from each other, get inspiration and stimulate further joint work, either in methods and tools development or in practical applications. If time and interest allows, the inventory may be extended to cover selected relevant projects outside the statistical community, e.g. in the geospatial community, government, academia or private sector.

WP3. Design of a quality framework. As the ML position paper shows, a quality framework tailored to the use of ML techniques needs to be developed. Traditional statistical quality frameworks assume that the data-generating process and further data processing steps are explicitly known. When applying ML methods, especially to 'found' big data or in multisource statistics, these assumptions are usually not valid. To

guarantee quality, reproducibility and transparency, which are core values of official statistics, it is important to identify suitable quality indicators and performance metrics. The quality framework could either be a stand-alone effort or an extension of an existing quality framework.

WP4. Development of a 'ML for dummies' handbook. Based on experience with the concrete pilots, the ML inventory and the quality framework, a handbook with 'hands-on' practical guidelines, best practices and recommendations for reproducibility/transparency/causal inference etcetera of ML-based statistics will be developed.

WP5. Communication and dissemination. The task of this WP5 is twofold: on the one hand spread the results of the ML project itself, on the other hand consider how to communicate to users and other stakeholders on ML-based and ML-enhanced official statistics. The WP5 team will also propose an approach for maintaining, updating and disseminating project deliverables such as the inventory, quality framework and handbook after closure of the ML project.

WP6. Overall project management. This is where the usual project coordination, reporting and planning go.

In principle all WPs can run in parallel although there are some temporal interdependencies between them that require explicit attention. Most work can be done in a virtual setting but two or three physical sprints will be necessary.

The application of ML methods by nature implies interdisciplinary work. Modellers (methodologists), programmers (computer scientists) and subject matter specialists must work together. For each WP, teams should be formed that combine these different skills to achieve optimal results. It may be useful to engage interested project members from areas outside the official statistics community.

Due to the mostly exploratory nature of the project, an agile approach works best. It could be helpful to identify an explicit product owner for each WP (member of the Executive Board?) in order to decide on deliverables and next steps. A sandbox-like environment could be helpful as a common workspace.

The project is designed for a one-year duration but given the complexity of the ML topic an extension could be necessary; depending on progress and appetite for deeper results.

Alternatives considered

Instead of a full-fledged coherent project, it may be possible to isolate some of the WPs and allocate them to appropriate Modernisation Groups as separate activities. In particular WP2 and WP3 would seem suitable for this. It could also be possible to try and collaborate with the academic community in scientific programs like the European Horizon2020 program.

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

ML is a key modern technologies that the worldwide statistical community should consider and the methods, IT solutions and other related issues can be dealt with in a universal; manner. Since, at this moment in time, basically all NSOs are in the same pioneering phase this is an excellent opportunity for shared development and mutual collaboration. The ML proposal seamlessly fits the HLG-MOS mission, all four elements of its vision are covered and all five HLG-MOS values are addressed.

Proposed start and end dates

Start: *January 2019*	**End:** *December 2019*