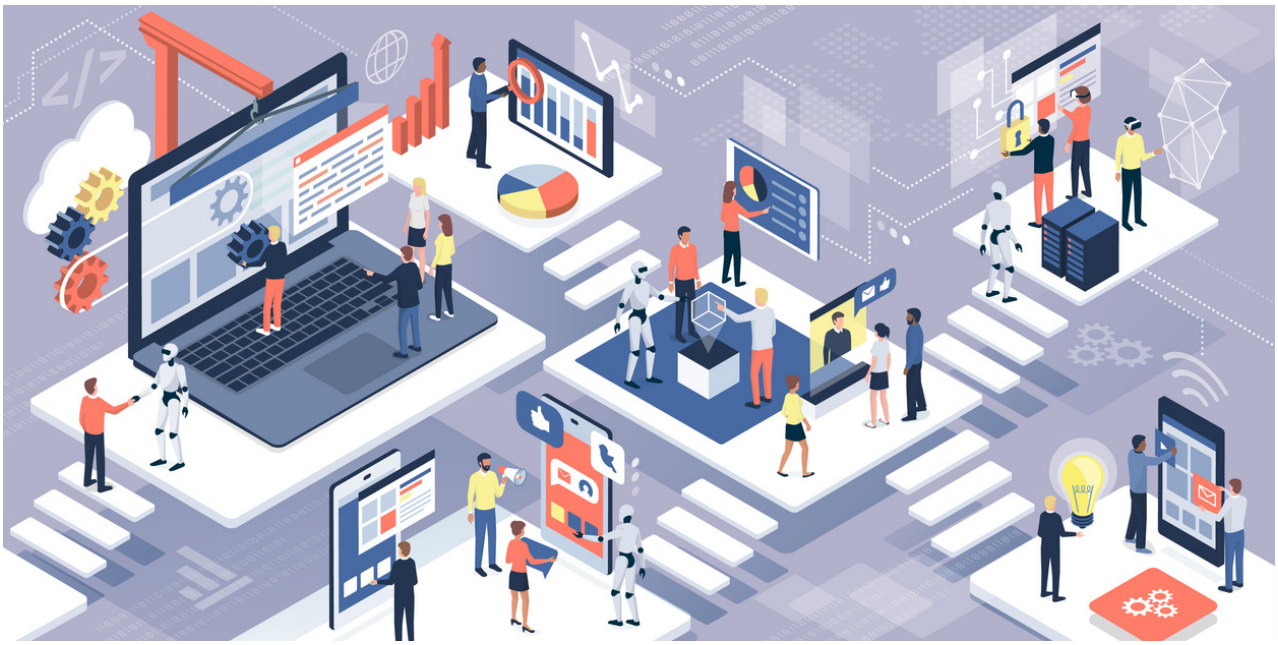
ONS-UNECE MACHINE LEARNING 2021 GROUP

By Alison Baily (UK ONS) and InKyung Choi (UNECE)



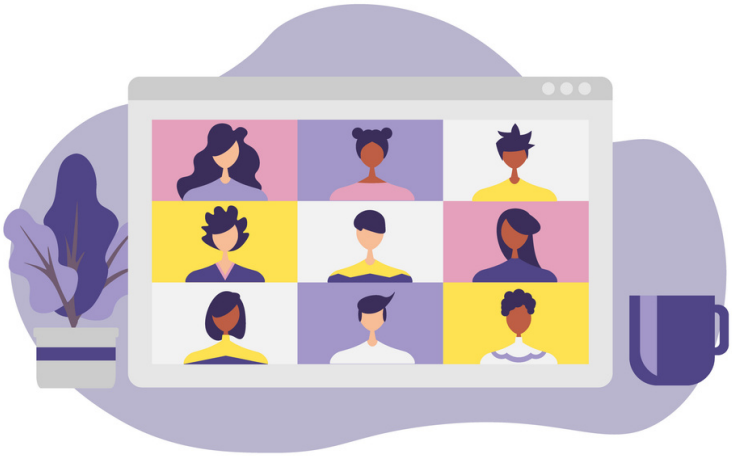
BACKGROUND

Machine Learning (ML) holds great potential for the production of official statistics. It can increase efficiency by automating certain tasks or by assisting manual processes. It also allows statistical organisations to use new types of data such as social media and imagery. Many national and international statistical organisations are exploring how ML can be used to increase the relevance and quality of official statistics in an environment of growing demands for trusted information, rapidly developing and accessible technologies, and numerous competitors. ML is still a relatively new method in official statistics and many statistical organisations are in the process of testing its feasibility for their data and environment. While local conditions vary, statistical organisations in different countries face similar types of challenges and they can benefit from sharing knowledge and experiences with each other, and from working together to develop common solutions.

ACTIVITIES

The ONS-UNECE Machine Learning 2021 Group is a platform for international research collaboration, knowledge exchange, resource sharing and capacity building in the use of ML for official statistics. Coordinated by the UK's Office for National Statistics (ONS) Data Science Campus and the UNECE High-Level Group on Modernisation of Official Statistics (HLG-MOS), the Group's objectives are to:

- Facilitate the creation, development and implementation of research projects and skill-building activities that meet the global statistical community's needs.
- Build and engage a strong machine learning community by sharing resources and good practice, exchanging ideas and experiences, and keeping abreast of developments in the field.
- Offer open, shareable, and easily accessible resources to the community; and
- Facilitate machine learning capacity building for official statistics.

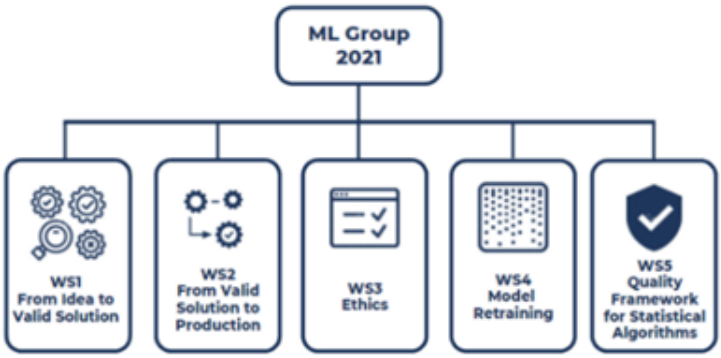


Following a HLG-MOS ML Project in 2019-20, the Group expanded its activities and reach in 2021. Membership doubled, from 120 members from 23 countries to 248 members from over 30 countries. Research workstreams were expanded from three to five, to include new issues such as ethics and model retraining that arise following the deployment of ML models. Members worked together on issues across the ML production cycle, from idea through to operationalisation and quality assurance. Highlights include the publication of ethics guidelines, the piloting of a data lake to support ML models development and modelling for estimating consumer spending of US states.

Throughout the year, the group has provided a regular schedule of high-quality expert presentations and training sessions to enable knowledge exchange, build skills and foster discussion on key issues such as model retraining and data privacy preservation. The community shares its outputs for the benefit of the wider statistical community through its website and a public webinar in November which was attended by 279 people.

RESULTS

Research collaboration



Research activities built on the results of the pilot, demonstrating the added value of machine learning in coding and classification, editing and imputation, and also exploring approaches to identify and address challenges of ML solutions in production.

WS1. From Idea to Valid Solution. Coding and classification was the most popular application area in this year's research of ML applications. New application areas investigated included modelling using ML and route optimisation. One study highlighted the benefit that NSOs gain from replicating ML projects from other NSOs.

WS2. From Valid Solution to Production. The workstream explored issues related to how to make the operationalisation of ML solutions smooth and efficient such as how to develop a user-friendly interface and how to build a data lake that data scientists can efficiently draw data from. It also produced a paper outlining typical steps that statistical organisations take from ML experiment to deployment and challenges during the journey.

WS3. Ethics. The workstream produced high-level guidance on ethical considerations that arise in ML projects to support analysts, researchers, data scientists, and statisticians. It has been published by the UK Statistics Authority.

WS4 Model Retraining. The workstream carried out a simulation study exploring how to identify the circumstances under which an ML model should be retrained in order to maintain the predictive power and quality of the model.

WS5 Quality Framework for Statistical Algorithms. In the 2019-20 project, a quality framework was developed to compare different methods including ML. This year the framework was tested using a real NSO use case. It reaffirmed the importance of having a holistic view, with quality dimensions having different priority for different stakeholders at different stages of the production cycle.

Knowledge exchange and Resource Sharing

Monthly meetings bring all members together to make connections, share experiences, and keep up to date with news from the workstreams and wider field. High quality presentations of projects are the main feature. Topics reflect the needs and interests of the official statistics community. Recent items include input privacy preservation, matching records when data is missing, MLOps and WordGraph2Vec algorithm for creating sentence embedding and a panel discussion on NSO model retraining. A newsletter is also shared with members summarising news, resources and upcoming events and opportunities.

The group shares its main outputs publicly to benefit the wider statistical community. All presentations, project reports and news items are shared, along with any open code and data, on the project website. Its final webinar attracted 279 attendees from 48 different countries for presentations of the group's research highlights and discussion with national data science leads of the way forward for ML in official statistics.

CAPACITY BUILDING

The Group has responded to high demand for training activities with a new series, Coffee and Coding. These are informal interactive sessions with expert data scientists who run a tutorial on a technical issue. Attendees have the chance to ask questions, pick up practical advice and learn new techniques and methods. At our last session in November over 100 attended from 20 countries. The Group also has a dedicated page for training resources and regularly shares new resources through its newsletter and main wiki page.

"The project has created the opportunity to spread the knowledge of ML among other members of the institution, benefiting the institution as a whole."

- ML 2021 Evaluation Survey

"The webinar provided me with new knowledge on how machine learning is being adopted in other countries. This also provided me with some insight in the areas where our NSO is currently lacking"

- ML 2021 Webinar Evaluation Survey

From Workstreams to Production

UK ONS - Household Financial Survey

ONS participation in the ML group led to exploration of ML to combine three separate surveys into one new one, the Household Financial Survey. Close work with Germany and Italy in particular.
Result: Time saved on editing and imputation, larger sample sizes and greater accuracy. ML application now being built into HFS production pipeline for rollout out by end of this financial year.

IMPACT

The Group has had significant impact on the development and application of ML in many statistical organisations. One of its most valuable outcomes has been the sharing of knowledge of those organisations with some experience of developing ML models with those who are planning to. Its meetings have provided a place to explore and test ideas, to share tools and methods, and to receive valuable feedback for addressing a range of common production challenges. By understanding the tried-and-tested approaches of other organisations, NSOs can use their ML resources in a more targeted way and accelerate their ML journey.

The Group has also had significant impact at the strategic level. It has changed the profile of the potential of ML among NSO decision-makers from a niche experiment to a credible technology for modernising statistical production. Seeing examples of successful ML projects from other NSOs helps persuade senior leads to invest in ML development for the first time.

LESSONS LEARNED

Machine learning is a fast-evolving field. This year, the ML Group has seen new methods discussed and experimented on in the group for the first time. This has emphasised the importance of having a solid quality framework to compare different methods (traditional and ML), as well as the need for continuous knowledge sharing within the global statistical community. The workstreams have also made progress on several new and under-explored production issues such as how to obtain high-quality training data set, how to monitor model decay once deployed, and how to develop an interface for users.

Many statistical organisations are exploring different ways of collecting data, with the use of big data such as web-scraping and satellite data becoming more and more common. The combination of big data and machine learning can greatly help modernise the production process, hence close collaboration between the two fields will be important to integrate them into production in an efficient manner. Another key lesson learned was that concerns on privacy and ethics will grow as public awareness about AI and call for putting the relelvant legislation in place increases. Statistical organisations will need to establish systems to protect privacy and address ethical concerns proactively.

CONCLUSION AND NEXT STEPS

In 2021, the ML Group has made significant progress in exploring new methods and tools, and in understanding how to address a wider range of production challenges. The Group's expansion has turned it into a valuable and well-known platform for the global statistical community. It has also demonstrated the high demand for international knowledge sharing and capacity building activities on ML and the importance of building on its work in 2022. Areas to focus on next year include more international research collaboration, recruiting more data scientists to participate actively in the workstreams, more systematic input from strategic data science leads at NSOs, and further investigation of production issues.

If you wish to join the ML Group 2022, please contact us (ML2022@ons.gov.uk)