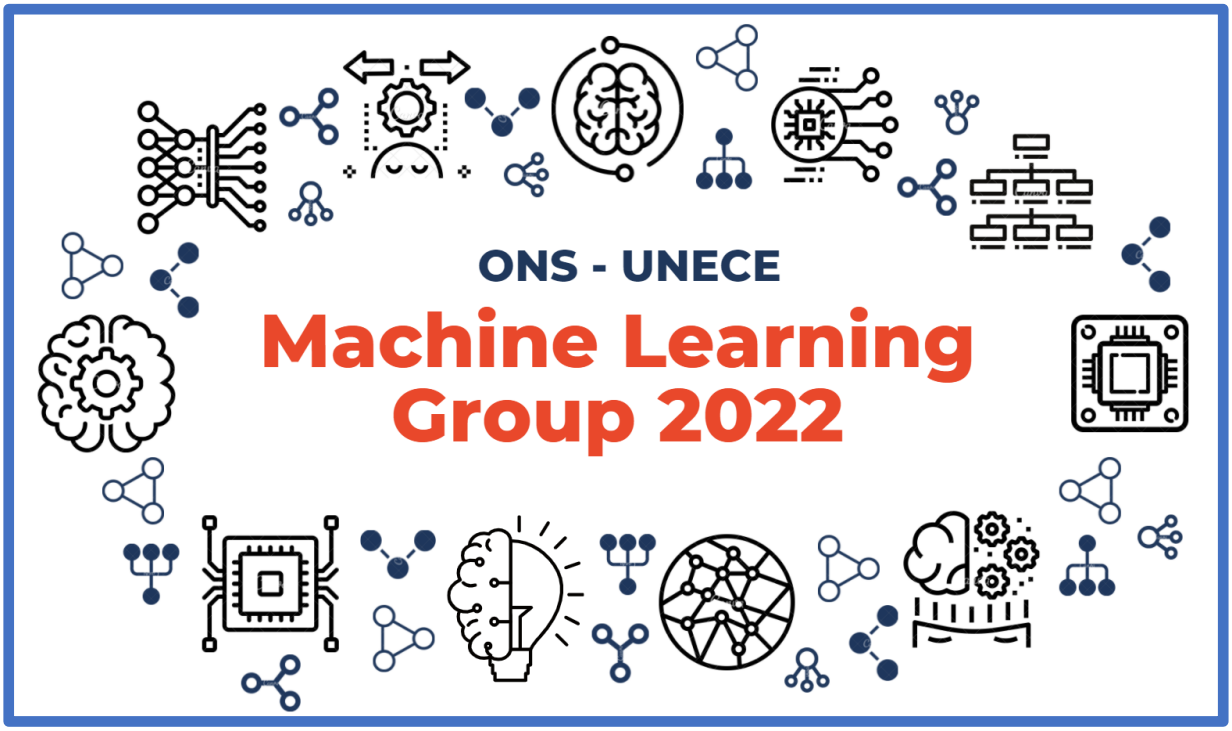
ONS-UNECE Machine Learning Group 2022

Final Report



Contents:

Authors:

InKyung Choi, Statistician, UNECE & ML Group 2022 Coordinator

Alison Baily, Programme Manager, ONS UK Data Science Campus & ML Group 2022 Coordinator

1. Background

Statistical organisations have made significant progress in recent years in exploring how machine learning (ML) can be used to increase the relevance and quality of official statistics. A rising number of organisations are moving from a purely experimental phase to adopting ML in various work areas. ML helps to produce new statistics based on new data sources (e.g., sentiment index based on twitter data), increase timeliness of existing statistics (e.g., crop yield estimation based on satellite data during non-survey periods) and assist humans to perform their tasks more efficiently (e.g., suggesting human coders most likely Standard Occupational Classification codes given a job description).

Interest in ML is spreading across the international statistical community. Statistical organisations in a wider range of countries and regions are dedicating resources to explore the potential of ML for their own statistical systems. International collaboration is important for those in the early stage of their data science journey. Drawing on existing use cases and organisational approaches, and sharing experiences can help accelerate development and help statistical organisations get up to speed with new technology and data sources in this fast-moving field.

The ONS-UNECE Machine Learning Group is an international platform for research collaboration, knowledge exchange and capability building on machine learning for official statistics. It brings together statisticians, data scientists and academics from 45 different countries around the world to explore how ML can improve statistical output and be integrated successfully into production.

Through presentations, discussions and research projects, the group aims to demonstrate the added value of ML for producing more relevant, timely, accurate and trusted data in an efficient manner. It also aims to increase the capability of statistical organisations to use ML by building skills and knowledge of data science through peer-to-peer learning, and by identifying common challenges encountered when incorporating ML in production processes.

Open to all levels of ML experience from expert to novice, the group provides a space for statisticians and data scientists working in official statistics to connect with and learn from others working on ML.

Its objectives include:

- 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.

The group was formed in 2019 as a two-year modernisation project of the UNECE Higher Level Group for the Modernisation of Official Statistics (HLG-MOS). Due to the high level of interest in ML and the success of the initial project in 2019-20, the group continued in 2021, led by the UNECE and the UK Office for National Statistics Data Science Campus.

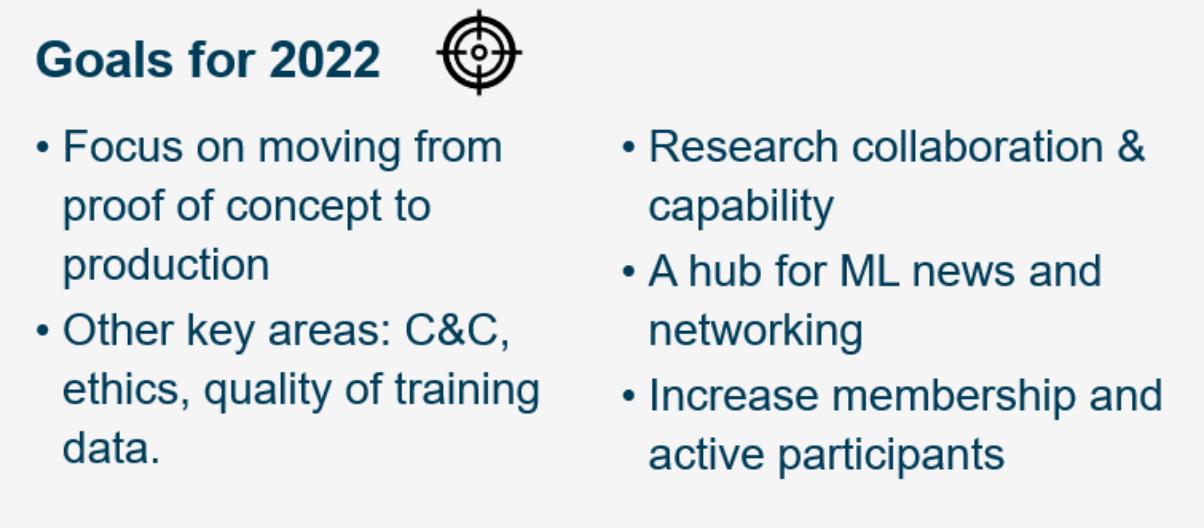
ONS and UNECE provide the central coordination while activities are proposed and run by members themselves. They include working together on exploratory research projects, participating in discussion forums to share technical knowledge and experience, and hosting regular expert presentations on ML innovations and implementation issues at statistical organisations. At the end of 2022, the group had grown from its original group of 12 members in 2019 to a community of over 400. It has produced a number of research reports, with highlights including guidance on ethical use of machine learning, a quality framework for statistical algorithms, model maintenance and infrastructure.

2. Programme Development and Objectives

In 2022 the ML Group aimed to respond to this growing interest in ML with an expanded programme. The objectives for this year were built around strategic priorities identified by the UNECE HLG-MOS Executive Board members and issues raised by group members.

At a strategic level, a need was identified to drive greater international collaboration, through joint research projects, stronger coordination with the UN Committee of Experts on Big Data and Data Science for Official Statistics (CEBD), and increased membership from different countries and regions not already represented in the group.

Another strategic priority was the need to focus on the process of moving from proof of concept to production, such as quality of training data and model retraining, as ML work developed in maturity at different statistical offices that had taken part in the earlier stages of the project. ML Group members also expressed a requirement for more training activities to build capacity, particularly for those with traditional statistical backgrounds who are new to data science and big data.



ML 2022 developed a programme that largely continued the successful formula of previous years - monthly presentations, informal, one-off training sessions, and regular news updates. It built upon this with a number of new initiatives that reflected the 2022 objectives. These included:

- Greater focus on knowledge exchange and best practice
- New research projects on Automatic Identification System (AIS) modelling and web scraping data
- In person sprint at ONS data science meeting
- Study group for earth observation self-study courses
- Stronger coordination and engagement of UN CEBD teams
- Establishing discussion forum for members to share ideas and ask questions
- In-person training workshops for UN Regional Hubs

The group launched its work at the UN Big Data Conference in Dubai in January 2022, with a series of training sessions, presentations, and workshops, attended by participants from national and regional statistical organisations in the Middle East, Africa, and Latin America.

The programme for 2022 focussed on the most important topics facing the use of ML in Official statistics, as identified by the group's members. Some 34 activity proposals were received from group members in response to an open call. As with previous years, text classification was the most popular topic. Members showed greater interest this year in modelling, model retraining and imagery.

The final programme was decided following individual discussions with the members who submitted the activity proposals. The Coordination Team also took into account the work already done in 2019-2021.

The following Theme Groups were formed:

- Text classification
- Imagery

- Modelling
- Web scraping data
- Quality of training data
- Model retraining
- IT infrastructure

The groups began work in March, with each comprising around 8-15 active members. Objectives, activities and ways of working were decided by group members themselves, based on interest and availability.

3. Activity and Results

The group carried out a range of activities covering research collaboration, knowledge exchange, capacity building and communication & outreach.

	Theme Groups	All Group **(and Public - in bold)**
Research Collaboration, Knowledge Exchange & Best Practice	AIS data modelling group Web scraping data group Newport sprint Model retraining Quality of Training Data IT Infrastructure Text Classification Earth Observation Research Group	13 x presentations at 7 Monthly Forum meetings 6 x presentations at 8 Text classification meetings
Capacity Building	Earth Observation Study Group	**2 x Coffee and Coding sessions** 4 x UN Regional Hub

		workshops
Communications	Members website Members newsletter	**Webinar** **Public website** Engagements with UN CEBD Task Teams

3.1. Research Collaboration & Knowledge Sharing

3.1.1 Theme Groups

A key ambition for this year was to invigorate international research collaboration within the group, with the establishment of a Global Data Squad. The Coordination Team invested significant time in exploring different proposals where colleagues from different countries could be part of this 'global squad' by working together on a single data set or research output.

Two separate research project groups were formed:

Web Scraping Data Theme Group. The web holds great potential to strengthen traditional statistical production as it contains an immense amount of information that is relevant for almost any policy domain. However, transforming web data to trustworthy statistics is not straightforward with numerous technical and methodological challenges.

In the Theme Group, three organisations, Statistics Flanders (lead, Michael Reusens), Statistics Poland and Turkish Statistical Institute, implemented experimental statistics using web scraped data in parallel for the production of identifying companies engaging in AI activity, R&D activity, corporate social responsibility activity respectively. The regular Theme Group meetings also provided a platform to discuss common problems and share use cases and experiences. More details about the collaboration can be found on the ML 2022 wiki.

AIS Modelling Theme Group. Automatic Identification System (AIS) is a tracking system for ships, originally developed for collision avoidance. In recent years, it has also been used as a data source for analysis in various fields such as trade and logistics[1]. To support statistical organisations in exploring the potential of AIS data for producing official statistics, the AIS Data Task Team of UN CEBD has developed algorithms as well as made AIS data available on the UN Global Platform .

Led by colleagues from CSO Ireland (Justin McGurk) and the Norwegian School of Economics (Gabriel Fuentes), the Theme Group explored methods to identify berth areas in ports using AIS data. Due to the vast size of the data, the raw data was filtered through the H3 index .

[1] https://unstats.un.org/wiki/display/AIS/Introduction

During the regular meetings, various geospatial objects handling methods were introduced. As the Theme Group was only able to start work in July, it will continue work and deliver its results in the first quarter of 2023. The aim is to further test the methods at a larger scale and produce a "cookbook" that could help statistical organisations with AIS data.

The other Theme Groups focussed their work on exchanging knowledge and experience, reviewing research papers, and developing new thinking on best practice. They also learned together through discussions and presentations of projects carried out by different statistical organisations.

Text classification Theme Group: Traditionally, text classification used to be done manually or through a complex rule-based system, both of which are resource-intensive. ML can help statistical organisations conduct this text classification task in a more efficient way. The Theme Group provided a knowledge exchange platform for those working on text classification in statistical organisations to share their works, receive feedback from peers and discuss common challenges. The Theme Group had a series of meeting over the year consisting of presentations from the group members. Experiences and lessons learned observed through the series of presentations are summarised in its report available on the ML 2022 wiki.

Imagery Theme Group: This group focussed on the use of machine learning for earth observation data. It provided a forum for presentations of use cases and research, as well as running separate sub-groups for research discussion and capability building. The Research Group was aimed at members who already had strong experience of working with Earth observation (EO) data and ML. The Group discussed a number of research papers on architectures of convolutional neural networks (CNNs), class imbalance and explainable artificial intelligence (XAI). The Study Group was formed for those seeking to build foundational skills in using EO and ML. The group worked together on self-study courses, with regular check in meetings to share advice and resources, and to agree study goals. A report of these group's activities can be found on the ML 2022 wiki.

Quality of Training Data Theme Group: Although a growing number of statistical organisations are implementing ML for various work areas, little attention is paid to training data quality. The issue is usually discovered after significant resources have already been invested when the ML model isn't performing as expected. The Theme Group, led by Statistics Netherlands (Marco Puts), explored the quality issue of the training data sets and identified processes of how different types of error affects the quality of the model. Two main sources of error were examined: human annotation process and representativity. The group's report is available on the ML 2022 wiki.

Model Re-training Theme Group: ML models built based on patterns learned from past data start to decay as they encounter new data that they have not been exposed to before. This happens due to changes of the data on which the model needs to make predictions (e.g., new products in market, new type of jobs) or change of relationship between input features

and output (e.g., update of statistical classification system). It is important to have a governance plan in place before the deployment so that the model can be continuously monitored and re-trained when needed. The Model Re-training Theme Group examined the key concepts around the drifts (e.g., drifts in data, drifts in model), methods for monitoring and detecting those drifts (e.g., performance based-approach, distribution-based approach). It also discussed the implications for statistical organisations (pros, cons), as well as the factors that enable the monitoring and re-training. Full report is available on the ML 2022 wiki.

Infrastructure Theme Group: Integrating machine learning capability into an organisation requires more than just ML skills. As with all new technology and methods, after initial experimentation, it needs a supporting environment and infrastructure to scale beyond specific small group of experts. The Infrastructure Theme Group, led by Statistics Sweden (Jakob Engdahl) explored various cross-cutting and horizontal issues that arise from ML such as linking ML processes with traditional production process, generic pattern for ML deployment and serving, roles, capabilities matrix. The full report of this group's activities will be available soon on the ML 2022 wiki.

3.1.2 ML monthly meeting

The Monthly Forum provided a main meeting point for the whole ML Group to come together to share experiences, build connections and keep up to date with the new developments. Regular progress updates were shared from the Theme Groups. High quality and relevant expert presentations were scheduled from both the academic and applied contexts. These included using advertising data to model poverty (Qatar), estimating linkage errors without training data (Canada); CNN for extracting image features from EO data (Netherlands). These meetings regularly attracted around 100 participants, reflecting their usefulness in transmitting knowledge of new applications and approaches to production issues in statistical organisations. The list of presentations and presentation slides are available on the ML 2022 wiki.

3.1.3. ML Sprint

Members of the ML Group 2022 met at the UK Data Science Campus in Newport, UK, 12-14 July for an in-person sprint. The sprint was part of a wider meeting organised by the UK Office for National Statistics entitled: Advancing International Collaboration in Data Science and Big Data for Official Statistics.

The Machine Learning Group 2022 sprint involved three of this year's Theme Groups:

- The web-scraping group aims to develop statistical indicators using web-scraped data and ML methods.
- The model re-training group explores various issues related to model re-training, such as how to detect model drift and how to decide when to re-train.
- The quality of training data investigates the concept of data quality in the context of machine learning and its effect on the ML model.

The participants of the work sessions found in-person discussions very productive, helping expedite the work. The meeting also offered a great opportunity to cross-fertilize among three Theme Groups as well as receive valuable feedback from outside the Theme Groups. A full report on their work is available on the ML website here.

3.2. Capacity Building

Capability building was another key priority for the ML2022 programme this year. The main way in which the ML Group aims to build capacity is through peer-to-peer learning in our research projects and knowledge exchange activities. To reinforce this, we arranged a few more targeted training activities, based on the priority needs identified by the community.

We ran three Coffee and Coding events, our one-off interactive and informal training sessions aimed at building foundational skills required for machine learning.

- The first session in January was delivered at the UN Big Data Conference in Dubai to an in-person audience of UAE data scientists from government, enterprise and universities. In the sessions, UK Data Science Campus colleagues Alex Noyvirt and Claus Sthamer demonstrated practical code examples and trained participants to carry out the development processes by themselves (available on youtube). There were also two other sessions looking at intermediate and advanced level applications for participants attending in person.

- The second session in April, delivered by Tom Wise from the UK Office for National Statistics, discussed Machine Learning foundations and focused on the theory behind these techniques. The course was designed to be accessible to non-programmers and incorporated examples relevant for official statistics (available on youtube).
- The third session, delivered in November, delivered by Tabitha Williams and Brittny Vongdara from Statistics Canada, provided an introduction to Git and GitHub. Topics covered included forking a repository, making a commit, and collaboration. It also covered the theory on the difference between GitHub and Git, what a Git project looks like normally, and best practices (available on youtube).

The sessions were open to non-members as well as members, and attracted high levels of interest. For the final Coffee and Coding session with Statistics Canada some 213 colleagues from over 40 different organisations attended. Participants were very engaged in the session with many questions asked. Some 87% of survey respondents rated the session as good or very good.

The Theme Groups have also played a significant capacity building role. The Imagery Group set up a Study Group focussed on gaining skills in EO data, where participants worked together on self-guided study courses. The AIS group's main aim was to give participants hands-on experience of dealing with AIS data on the UN Global Platform. This kind of learning through peer-to-peer exchange and hands-on experience is a good complement to traditional tutor-delivered training courses. Given the level of online learning resources available, these groups provide a structure that enable members to make fast progress on their learning goals.

Alongside technical skills, the Group has also been active in training on business strategy for ML. The ONS team delivered three workshops on building ML capability at Statistical Offices for the UN Regional Hubs for Big Data. The UN Regional Hubs provide NSOs with opportunities for training and collaborative hands-on project activities to help them adopt big data and data science into their statistical production systems. The workshops involved representatives of the Hubs for Asia, the Middle East, Africa and Latin America. In the sessions, the Regional Hubs learned more about the value of ML for statistical production, as well as approaches and techniques to strategy, planning and business delivery. The interactive nature of the

workshops enabled them to make progress on developing approaches tailored to issues faced by statistical organisations in their regions.

3.3. Communications and Outreach

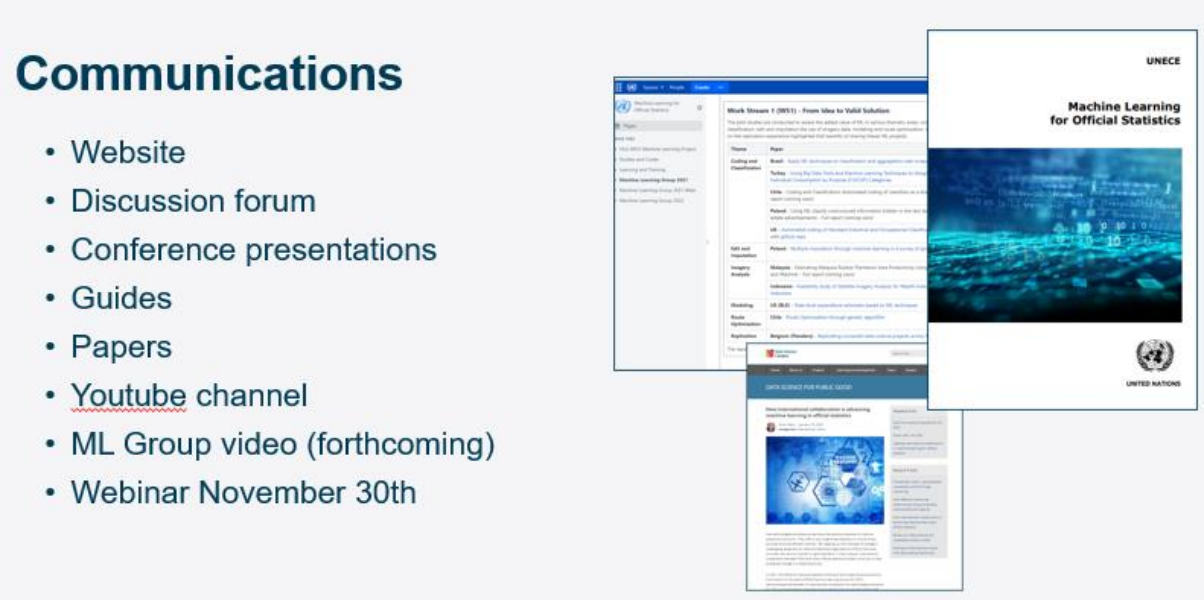
Effective communication is vital to the group's aims of sharing knowledge and good practice, and of offering easily accessible resources to the statistical community.

The main communications channel is the UNECE wiki website, which provides information on the group's activities to both members and non-members:

- The public page provides information about the group, how to join, upcoming public events and presentations from the Monthly Forums.
- The members page (non-public) provides full information about the Theme Group activities, recordings from the Monthly Forums, and news of training opportunities, events, research papers and technical guides from the ML field.

In addition, members receive regular updates and an occasional newsletter via email.

A new initiative for 2022 was a discussion forum for group members on the platform Slack. This provided a space for colleagues to connect, share quick updates and ask questions from other members. This complements the existing Yammer channel which acts more as a noticeboard for information to the wider statistical community.

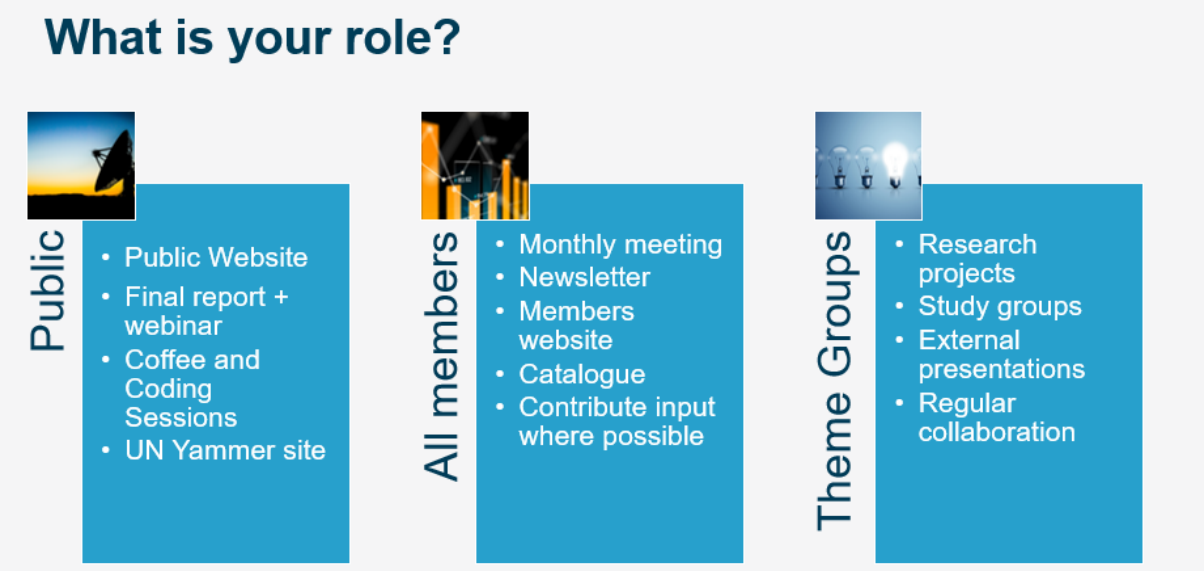


As part of its objectives to expand membership to new countries, the ML Group promoted its work at several international statistical meetings which included: a presentation on the ML Group's work at the UN Big Data Conference in Dubai in January, where the group promoted the value of machine learning for national statistical offices to new audiences in the Middle East. The group's work was also featured in a new publication, "Machine Learning for Official

Statistics", from the UNECE. This publication draws together the lessons learned from the ML Group's work since 2019, presenting the practical applications of machine learning in text classification, edit & imputation and imagery analysis) and discusses their value added, challenges and lessons learned.

3.4 Membership and Organisation

Over the past year, membership has grown from 270 members (end 2021) to almost 430 by the end of 2022. Members come from national statistical organisations in 45 different countries as well as the UN and other international organizations. The group has been run on a community-driven model - members were encouraged to initiate, lead, assist and follow the group's numerous activities.



Group members have been enthusiastic and committed as participants. The meetings receive regular positive feedback. However, the majority of members are passive - out of the 430 on the mailing list around 100 attended the main monthly meeting, plus roughly another 60 regularly engaged in the small group activity. A major issue has been time pressure on participants who have not been able to take time out of their normal duties to commit to regular participation or more in-depth collaboration. Many people sign up to activities at the start of the year, but then drop out due to other work demands.

This lack of volunteers means that responsibility for leading and developing the different activities this year has fallen mostly on the Coordination Team. A core group of members are able to ensure the activities continue by stepping forward to do presentations, and to lead smaller group activities. However, sustaining the group relies heavily on the Coordination Team of two staff tasked with running the group as part of their official duties by ONS and UNECE.

4. Conclusion

4.1. Impact

Since it was established in 2019, the Group has had significant impact on the development and application of ML in many statistical organisations. Statistical offices around the world are at different stages in their ML journey. Therefore, one of the group's most valuable outcomes has been the sharing of knowledge and experiences within the official statistics community.

Firstly, the knowledge sharing from those organisations with some experience of developing ML models with those who are planning to has played an important role for organisations in the early stages of adopting ML. Feedback from the group's members shows that the sharing of tried-and-tested approaches of more experienced statistical organisations has helped other NSOs to use their ML resources in a more targeted way and accelerate their ML journey.

The Group is also highly valuable for those more experienced organisations who are looking to drive further innovation and upscale machine learning applications. The group has become a well-established focal point for statisticians and data scientists to come together to explore and test ideas, to share tools and methods, discuss the latest academic research papers and to receive valuable feedback for addressing a range of common production challenges. The challenges around production, which are more prominent issues once you are beyond the initial phase of ML maturity, are discussed as well.

Alongside the building of technical capability, the group has also had a noticeable impact at the strategic level. During the course of the group's activity, it has helped 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.

4.2. Lessons Learned

This year's activities have helped deepen understanding of the principles and approaches that statistical organisations should consider for harnessing the potential of machine learning for statistical production. A few highlights include:

Invest for continuous knowledge expansion: Machine learning is a fast-changing field. New methods and algorithms are becoming quickly established as mainstream approaches. For example, in the ML Group 2022, the use of transformer, embedding-based methods became more prominent compared to the past few years. The fast pace in this field implies that statistical organisations should continuously invest in expanding their knowledge base as well as supporting staff in acquiring new skills.

Continue exploring use cases for statistical organisations: In 2022, text classification and imagery analysis continued to be the most popular use cases in statistical organisations, as was the case in the group in previous years. ML is seen again as an essential tool for big data sources such as web-scraped data and satellite data which statistical organisations are increasingly looking to as new data sources. The ML2022 Group also saw use cases outside these usual application areas. For example, machine learning was used to select survey respondents to follow (Australian Bureau of Statistics), efficiently sort customer inquiries (Statistics Canada), assist sampling by selecting companies that are likely to engage in a particular industry (TurkStat), and to verify statistical statements made by politicians (Non-Governmental Organisation - Full Fact, UK). Statistical organisations could explore the full range of ML application areas in order to harness its potential while also deepening the knowledge and experience in the areas where ML has been shown to be effective.

Make the best use out of existing statistical expertise to improve quality: As is often the case with the introduction of new technologies, especially disruptive ones, ML seems to have created a cultural division in statistical organisations. Although there are many overlaps, traditional statistics and ML have different principles, approaches and languages (even when referring to the same concept). However, the tension between the two fields could impede constructive collaboration. The expertise, know-how and best practices established in statistical organisations can be used for ML. For example, the rigorous approach for quality analysis and thorough quality control procedure that statistical organisations have long used can be used to analyse the quality of ML training data and to maintain the quality of ML predictions.

Create a supporting environment for ML: While the awareness of ML and its potential has spread significantly in the official statistics community over the last few years, it is still considered as a niche technology. Putting ML models into production still remains one of the most challenging aspects in using ML in the statistical organisations. Therefore, they need to establish an environment (culture, skills, process, data management, IT infrastructure, organisational setting) that can support this process of productionisation alongside innovative experimentation. Enabling these two different processes is critical for ensuring that statistical organisations can use ML in the long term in an effective way. Further investigation on what elements are required in this environment and how to move forward is needed.

Make international collaboration a core part of ML adoption strategies: International knowledge sharing and research collaboration is an important way to help statistical organisations keep up with the scale and pace of change required for adopting ML. There is high demand from the statistical community for the ML Group's capacity building activities. However, most members lack time to participate fully and lead activities. This restricts the level of collaboration and makes progress reliant on the availability of a full-time

Coordination Team of two people (technical lead and project manager). To achieve the full potential, participation in international collaboration activities should be part of an organisation's modernisation strategy, with staff allocated protected time within their official duties to take part in these projects.

4.3. Next Steps

The ONS-UNECE Machine Learning Group will close its activities at the end of this year and will not be continuing in 2023. The group has achieved its original intention of investigating the value and requirements of machine learning for official statistics, and has achieved much more besides. Its work has had a significant impact on supporting machine learning's development through its early stages in statistical organisations. Now that the community is at a greater stage of maturity, the focus needs to move from exploration to production and other issues.

It is expected that some functions of the ML Group will be carried out by other new international and regional initiatives (e.g., the UN CEBD Data Science Leader's Network and Regional Hubs, and the ESSnet Center of excellence). The UNECE wiki space for Machine Learning for Official Statistics and Slack discussion forum will remain open. The UNECE Expert Meeting on Machine Learning for Official Statistics in June 2023 will provide a platform for statistical organisations to share new developments in the field and discuss issues and challenges.