First Webinar of the Sprint on Artificial Intelligence and Data Science for Economic Statistics

7 November 2024, 7:00 am – 10:00 am (GMT -4)

The Sprint on Artificial Intelligence (AI) and Data Science for Economic Statistics is part of a collaborative effort between the UN Network of Economic Statisticians and the UN Committee of Experts on Big Data and Data Science for Official Statistics. The collaboration is part of the annual work plan to explore the potential of AI in transforming economic statistics. It aims to modernize the production of economic statistics and enhance user experiences by leveraging advancements in AI and data science. It is structured around three core objectives: creating a repository of impactful use cases in AI and data science to streamline data production and improve decision-making, exploring ways to enhance the dissemination of statistics through AI, and addressing critical challenges such as data privacy, ethical AI practices, and cross-domain integration. These objectives support the broader goal of ensuring that AI-driven statistics maintain high-quality outputs and integrity.

The Sprint is structured in the format of two webinars and an international seminar, to be held in Dubai in January 2025. Below is a summary of the discussions and key takeaways from the first webinar, which took place on 7 November 2024.

Opening session

The first webinar of the Sprint began with a focus on practical applications and capacity-building steps necessary for integrating AI into official statistics. Participants were encouraged to engage with key questions, including identifying the most relevant use cases for National Statistical Offices (NSOs), determining effective strategies for skills development, and exploring new partnerships to advance the use of AI.

The opening session also highlighted two major projects: the Data Science Leaders Network (DSLN) Playbook, which provides practical guidance on the use of data science and AI in official statistics, and Eurostat's initiative to develop resources on AI and machine learning.

The opening session concluded with a focus on the importance of collaboration in advancing the use of data science and AI in official statistics, as illustrated by the presentations on the DSLN Playbook and Eurostat's AI and machine learning project. Speakers emphasized the need for multidisciplinary teams within statistical offices and collective efforts across organizations to transition experimental tools into production-ready solutions. The role of collaborative networks, such as the Data Science Leaders Network, was highlighted as essential for knowledge sharing, capacity building, and driving innovation.

Update on the Data Science Playbook for Official Statistics

The UNSD provided an update on the development of the Data Science Leaders Network (DSLN) Playbook, which originated from strategic recommendations during the DSLN's first Sprint in early 2023 and was endorsed by its Bureau later that year. Virtual meetings toward the end of 2023 helped develop an outline of the Playbook's structure and content, which were further refined during a DSLN meeting held in January 2024 in Dubai. The Playbook is intended to be a practical, action-oriented resource, accessible to both traditional statisticians and emerging data science practitioners. It is organized into three main pillars—efficiency in operations, responsiveness to emerging needs, and full digital transformation—with cross-sectional themes addressing governance and strategic collaboration. The current drafting process is progressing toward an in-depth review at an in-person DSLN meeting the January 2025.

Eurostat's project on artificial intelligence (AI) and machine learning for official statistics

Francesca Kay, Chief Information Officer at Ireland's Central Statistical Office, provided an overview of Eurostat's project on artificial intelligence (AI) and machine learning for official statistics. This initiative, funded by an EU grant and launched in April 2023, aims to develop resources and frameworks that can support National Statistical Institutes (NSIs) in integrating AI and machine learning into statistical production. The project emphasizes transitioning experimental AI applications into standardized, reproducible, and production-ready solutions. Key outputs include a repository of use cases, training materials, and a sandbox environment for experimentation and model development. The initiative also focuses on creating a "platform in a box" to help countries establish AI environments and reduce duplication of effort. Collaborative work packages, led by 14 participating countries, cover areas such as Earth observation, imputation, error detection, text classification, and synthetic data generation. The project prioritizes transparency, ethical AI practices, and scalability while addressing challenges like resource constraints and the lack of decades-long experience with these techniques. Events and updates are planned through 2028 to ensure widespread engagement and dissemination of results.

After the opening session, participants transitioned to breakout groups designed to explore specific applications of data science and AI. The first breakout group, moderated by Bertrand Loison of Switzerland's Data Science Competence Center, focused on reproducible analytical pipelines with presentations from the UK, Switzerland, IMF, and Indonesia. The second breakout group, moderated by Marco Marini of the IMF, addressed applied data science for economic indicators with contributions from Mexico, the Netherlands, Austria, Indonesia, and the IMF.

Breakout Group 1

The first breakout group, moderated by Bertrand Loison, focused on practical applications of AI and data science in statistical production. Martin Ralphs from the UK Office for National Statistics (ONS) opened the session with a presentation on reproducible analytical pipelines (RAPs), emphasizing how they streamline statistical workflows by automating previously manual processes. Ralphs highlighted the substantial efficiency gains and quality improvements achieved through automation, open-source tools, and software engineering best practices. He also shared lessons learned, including the importance of starting with small, well-defined projects and building a culture of collaboration and continuous improvement.

The second presentation by Christopher Sulkowski from Switzerland's Federal Statistical Office showcased the implementation of a RAP for a major pension fund. The pipeline automated data aggregation and validation processes, reducing a five-day manual task to 15 minutes and enabling advanced analytical capabilities. Sulkowski emphasized the modularity of the pipeline, its focus on accessibility and replication, and the use of tools like Luigi for managing complex dependency graphs. He also described the collaboration as a training opportunity, ensuring the pension fund could maintain and expand the system independently.

On a third presentation, Mario Saraiva and Alessandra Sozzi from the IMF presented "Port Watch," showcased a platform leveraging AIS data to monitor global trade disruptions. The RAP processes massive datasets to track port activity, estimate trade volumes, and monitor rerouting during crises. The system, hosted on the UN Global Platform, emphasizes scalability, efficiency, and collaboration, demonstrating the cost-effectiveness and utility of RAPs in near-real-time data analysis.

Setia Pramana from Indonesia's Statistics Polytechnic followed with a presentation on using AIS data to estimate port activity and greenhouse gas (GHG) emissions in Indonesian waters. The study integrated multiple data sources, applied clustering algorithms, and analyzed emissions by fuel type and vessel activity. Key findings highlighted the dominance of heavy fuel oil in emissions and the significant role of major shipping routes like the Malacca Strait. Validation exercises showed alignment with official data, though refinements to port area definitions and ship classifications were identified as future steps.

The session concluded with a discussion on tools, implementation strategies, and the importance of balancing automation with human oversight. Questions addressed the choice of tools, ongoing relationships with stakeholders, and the broader applicability of RAPs for both national and international statistical systems, highlighting their role in enabling data-driven decision-making in complex, global contexts.

Breakout Group 2

The second breakout group, moderated by Marco Marini, focused on using artificial intelligence (AI) and data science to produce economic indicators, showcasing practical use cases from national statistical offices.

Abel Coronado Iruegas presented the supervised crop classification project by INEGI, Mexico, which uses satellite imagery and machine learning to monitor agricultural activity. Leveraging Sentinel and Landsat data, the project has made significant strides in crop identification and mapping within Mexico's agricultural frontier. A robust infrastructure, including a pre-production environment with cloud and local resources, facilitates collaborative work. The project has achieved promising results with a classification accuracy of 91% and plans to expand its geographic and crop coverage. Tutorials and open-source methodologies are being developed to share the findings widely.

Gert Buiten from Statistics Netherlands discussed efforts to estimate firm-level supply chain networks using AI/ML models. The project, part of a broader EU initiative, seeks to identify buyer-supplier relationships by leveraging administrative data from countries like Portugal. By incorporating factors such as geographic distance, firm size, and industry codes, the models aim to recreate realistic supply chain networks for countries lacking comprehensive data. The initiative emphasizes scalability and adaptability across EU countries and beyond, focusing on addressing rare event challenges in supply chain data. Future work involves refining models and creating synthetic datasets to ensure usability across various contexts.

Christian Stefan presented Statistics Austria's efforts to develop a methodology for detecting solar panels on rooftops using high-resolution aerial imagery and deep learning. The project leverages orthoimagery with a spatial resolution of 0.2m, updated in a three-year cycle, to enhance geospatial energy statistics. They tested Esri's pre-trained Mask R-CNN model for object detection and semantic segmentation but found limited success due to differences in resolution and land-use patterns between U.S.-trained data and Austrian landscapes. Despite initial setbacks, the team identified potential for retraining the model using higher-resolution imagery and integrating additional data such as building footprints and near-infrared (NIR) imaging. This initiative aims to improve the granularity and accuracy of renewable energy statistics, ultimately enabling a large-scale assessment of solar panel distribution in Austria.

Hadi Susanto outlined how AI and machine learning can enhance the estimation of economic indicators, focusing on time series forecasting, nowcasting, and big data processing. Highlighting examples like GDP growth, inflation, and unemployment forecasting, he demonstrated how AI techniques have improved accuracy, reduced errors, and enabled timelier insights. For example, using graph neural networks and hybrid models, institutions like the U.S. Bureau of Labor Statistics and WTO have achieved significant efficiency gains, such as processing 30 million

monthly online job postings. However, Susanto emphasized challenges, including data bias, lack of transparency in AI models, privacy concerns, and limited technical expertise. He called for stronger collaboration among statistics producers, standardized AI practices, and ethical considerations to ensure accountability and reliability in AI-driven economic statistics.

Ike Maduako showcased the IMF's innovative approach to nowcasting GDP using satellite data and machine learning. He discussed the IMF's challenge in providing timely economic assessments for countries lacking high-frequency macroeconomic indicators. Leveraging nontraditional data sources like nighttime lights, NO2 emissions, and vegetation indices (NDVI/EVI), the IMF developed machine learning models capable of producing reliable GDP nowcasts. For instance, NO2 emissions emerged as a strong proxy for economic activity in manufacturing sectors, while nighttime lights proved valuable for tracking urban development. Maduako shared examples of model performance in countries like Uganda and Afghanistan, achieving high accuracy even in data-sparse regions. He emphasized the importance of explainable AI for stakeholder trust, demonstrating how partial dependency plots can clarify the impact of each variable on GDP estimates. Future plans include expanding a global nowcasting tool accessible to IMF economists.

The breakout group concluded with a discussion focused on challenges and opportunities in adopting AI and machine learning for official statistics. One key question addressed resistance to change among staff when introducing modern methodologies. Presenters acknowledged this challenge, emphasizing the importance of capacity-building, clear communication about the value of AI tools, and fostering a culture of innovation. The potential of AI to complement human expertise rather than replace it was highlighted, as well as the need for transparency, standardized practices, and privacy-preserving techniques in AI-driven processes. There was widespread agreement on the transformative potential of AI, particularly in resource-constrained settings, but the consensus was that collaboration and rigorous evaluation are essential for sustainable implementation.

Conclusions and key sprint takeaways

Webinar participants reconvened in a plenary session which synthesized insights from breakout discussions, highlighting the future potential of big data, data science, and AI in official statistics. Moderators and panelists emphasized key themes, including promising applications, the need for capacity building, the balance between innovation and quality, and the importance of collaboration in advancing statistical practices.

Promising use cases

Discussions underscored the transformative potential of Earth observation data, particularly satellite imagery, in enhancing statistical outputs. Applications such as forestry monitoring, crop identification, displacement tracking, and macroeconomic forecasting were highlighted.

Advances in satellite technology, such as higher-resolution sensors and broader spectral coverage, have expanded the scope of possible applications. Similarly, reproducible analytical pipelines (RAPs) were identified as crucial tools for improving efficiency, reallocating resources, and enhancing data quality across both economic and social statistics. These pipelines also offer opportunities for collaboration, particularly when combined with platforms like the UN Global Platform, facilitating shared development and standardized approaches.

Capacity building and skills development

Participants agreed on the critical need for integrating theoretical training with practical, on-the-job applications to foster data science capabilities. Hands-on project work was identified as an effective method for enabling statisticians to develop and apply new skills, supported by cross-agency collaboration and shared guidance. Innovation projects and experimental statistics were presented as important precursors to official statistics, providing a structured pathway for integrating novel approaches. Regional collaboration was also emphasized, with examples of capacity-building workshops fostering shared resources, standardized methodologies, and joint efforts to tackle common challenges. Such initiatives were framed as essential for building a robust global community of data science practitioners within statistical systems.

Balancing innovation and quality

The session explored the challenge of integrating innovative methods while maintaining the high-quality standards expected of official statistics. Panelists emphasized that while fundamental quality assurance principles remain unchanged, new technologies, particularly AI and generative models, introduce unique challenges such as algorithm transparency and interpretability. Strategies to address these challenges include adhering to established quality frameworks, ensuring human oversight, and maintaining public trust by clearly explaining methodologies and assumptions. The complementary use of traditional and new data sources was seen as a way to enhance the granularity and timeliness of outputs without compromising comprehensiveness or reliability.

Collaboration and future directions

The first webinar concluded with a call for increased collaboration among statistical agencies and stakeholders to advance the use of big data and AI in official statistics. Joint projects and resource sharing were identified as essential mechanisms to leverage collective expertise while avoiding duplication of effort. The development of a data science playbook was highlighted as a practical next step, offering accessible guidance, reusable code, and real-world case studies to support broader adoption. Looking ahead to the upcoming Dubai symposium, participants proposed a focus on replicable use cases, frameworks for cross-agency collaboration, and strategies for integrating innovative methods into official statistics. These efforts will ensure the continuous improvement of data quality, adherence to ethical standards, and alignment with the evolving needs of statistical users.