RAF/AFCAS/23-E-105 **E**

December 2023

منظمة الأغذية والزراعة للأمم المتحدة | 联合国粮食及农业组织 | Food and Agriculture Organization of the United Nations | Organisation des Nations Unies pour l'alimentation et l'agriculture | Продовольственная и сельскохозяйственная организация Объединенных Наций | Organización de las Naciones Unidas para la Agricultura y la Alimentación

AFRICAN COMMISSION ON AGRICULTURAL STATISTICS

Twenty-eighth Session

Johannesburg, South Africa: 4 – 8 December 2023

AGENDA ITEM 10: **New developments in the use of alternative data sources for agricultural statistics**

Bridging Agricultural Data Gaps: Innovations in Geospatial and Non-Conventional Data Sources

Lorenzo De Simone, Office of the Chief Statistician
Christian Mongeau, Statistics Division

ORGANIZATION
Food and Agriculture Organization of the United Nations

SUMMARY

In decision-making processes that impact the economic, demographic, social, and environmental aspects of society, official statistics play a crucial role. They rely on the thorough and precise gathering of data. However, this process can be significantly hindered by data gaps. These gaps, which denote the lack of data or incomplete data sets, can greatly affect the trustworthiness, representativeness, and overall usefulness of these statistics. Addressing these data gaps is essential to ensure that the statistics we rely on for understanding various aspects of society are as robust, representative, and useful as possible. This endeavour not only enhances the quality of the data but also bolsters the integrity and efficacy of the decisions based on them.

This session is devoted at showcasing new developments in the use of alternative data sources for agricultural statistics in two specific domains. Firstly, there will be an emphasis on the field of geospatial data science by illustrating the EOSTAT project and the Farmer Registry project funded by the African Development Bank.

Secondly, we will share practical experiences in leveraging non-conventional data sources to gain timely insights by presenting some of the activities made by the Data Lab in the Food and Agriculture Organization of the United Nations (which will also be illustrated in this paper) and the African Development Bank's Data Innovation Lab.

Launched by the Office of the Chief Statistician and the Statistics Division (ESS) of FAO, the Earth Observation for STATistics (EOSTAT) project aims at building and strengthening technical capacity in National Statistics Offices (NSO) in the use of Earth Observations Big Data and AI as an alternative data source and methods to improve overall quality, completeness, granularity and timeliness of agricultural statistics with a focus on crop acreage and crop yield. This report provides an update on the status of work in Senegal, Mali, Rwanda, Lesotho, and Zimbabwe and describes upcoming related activities in the Region for the next biennium. Members are invited to express their views on the progress achieved and remaining challenges, and provide guidance as deemed appropriate.

The FAO Data Lab, established in late 2019, aims to modernize the Organization's statistics and expand data coverage using innovative methods and technologies for extracting data from unstructured sources. Its initial project involved gathering sub-national data from official statistical websites to address gaps and validate the national agricultural production dataset. This will be detailed in the following section. With the onset of the COVID-19 pandemic in 2020, the need for timely information to support decision-making became critical, particularly during emergencies. This led to a heightened demand for real-time information and automated analysis from non-conventional sources. To meet these emerging data needs, the Data Lab has been developing and managing various databases, including global newspaper articles, daily food prices information, and data on food loss and waste from scientific articles and other documents.

This document succinctly explores the issue of statistical data gaps and examines how geospatial data and non-conventional sources can help alleviate the challenges posed by a lack of data.

1. INTRODUCTION

Traditional statistical methodologies have historically served as the foundation for both data collection and subsequent analysis. These approaches, which span a spectrum from detailed surveys to hands-on field observations, have been instrumental in yielding critical insights across various sectors. Despite their value, traditional methods of collecting official statistics frequently encounter the challenge of "data gaps", meaning areas where information is either lacking or incomplete.

Data gaps can arise from a multitude of sources, including but not limited to surveys with low response rates, incomplete medical records, or the cessation of data collection methods due to unforeseen circumstances such as a global pandemic. These gaps can manifest during the initial data collection phase or may develop over time as data becomes outdated or irrelevant. The representativeness of data is particularly compromised when certain subgroups within the population are underrepresented due to missing data, leading to biased conclusions and potentially flawed policy-making. Incomplete data can significantly impact individual economic prospects and hinder social mobility, especially for marginalized groups, and can also lead to the formulation of policies that fail to adequately meet the requirements of these communities, consequently reinforcing systemic biases and maintaining social disparities.

The COVID-19 pandemic clearly underscored the susceptibility of statistical systems to unexpected events, with widespread interruptions in data collection activities worldwide being a prime example. The resilience and future progress of official statistics will depend on their capacity to navigate these challenges while maintaining data objectivity and reliability in a continuously changing social and technological environment.

In this scenario, data science methods have emerged as a valuable complement to traditional statistics, offering new perspectives and insights. Data science employs advanced analytical techniques and algorithms to analyse large and diverse datasets, including unconventional data sources. This approach allows for the exploration of hidden patterns and predictions, offering a more dynamic and comprehensive understanding of complex issues. Moreover, data science's ability to handle big data and apply machine learning models has opened new avenues for filling data gaps. For instance, predictive modelling and sentiment analysis can provide insights into public opinion and behaviour, areas often missed by traditional methods.

In the remainder of this document, we will explore some innovative methodologies across the fields of geospatial data science and non-conventional data sources to gather timely insights.

2. MAIN BODY OF THE DOCUMENT

Earth Observation Big Data and the EOSTAT project

Earth Observation Big Data constitute a valuable alternative data source that can support the work of National Statistics Offices in the production of agricultural statistics and reporting on SDG indicators, reducing the overall burden through increased efficiency. Nowadays the conditions for adopting EO data for agricultural statistics are very favourable: the unprecedented and generous availability of free and open high and medium-resolution satellite data resulting from the democratization of earth observation data, the growing amount of expertise, methods and tools to use them are important enablers. It is now possible to access Petabytes of satellite images from a variety of data sources (Landsat, Modis, Sentinel) allowing for building dense time series to detect and discriminate land cover classes, crop phenology and spectral traits. This in turn, permits the accurate mapping of land cover, crop type and for deriving accurate agricultural statistics. The development of cloud storage and computing technology, and the rise of machine learning and artificial intelligence have further widened the horizon of options for deploying low-cost infrastructures and automation.

Despite all, the actual uptake of EO data for operational use in NSOs is still very low globally, and especially in developing countries, due to a series of technical, financial, and administrative barriers. The most relevant challenge though is certainly the lack of in-situ data of adequate quality.

In-situ data, or more simply georeferenced crop data, is required in large amounts for the training, testing, and validation of supervised classifiers, which are then used to "transform" satellite images into categorical maps, hence crop type maps.

In this context, FAO, through the EOSTAT project, is supporting countries in building technical capacity to use EO data for official agricultural statistics by working together on the key issues in collaboration with the NSOs, with a strategy articulated over three strategic actions:

- Action 1 focuses on reviewing the existing national field data collection protocols (Agricultural Surveys and Agricultural Censuses) and ensuring that these are optimized for EO applications in both the sampling distribution across classes and geography and that best practices for georeferencing are implemented.
- Action 2 focuses on the development and testing of classification methods applied to time series of satellite images, which are more data frugal compared to the Random Forest, and on supporting the research in the field of transfer learning.
- Action 3 articulates the co-design and co-creation of the national crop monitoring system to demonstrate the feasibility and then to support their operationalization.

Since 2019, EOSTAT has been implemented in 21 countries globally, building technical capacity in the following key technical domains:

- Optimization of field survey design and integration of best practices in georeferencing within the Annual Agricultural Surveys and Agricultural Census
- Land Cover mapping
- Crop type mapping
- Crop yield modelling
- Field parcel mapping
- Computation of SDG indicators using EO data

Within the scope of AFCAS 2023, FAO is providing an update on the work carried out in Rwanda, Lesotho, Senegal, Mali, and Zimbabwe

Land Cover mapping

In Rwanda, in collaboration with the National Institute of Statistics Rwanda (NISR), the Rwanda Agricultural Board (RAB), the National Land Authority (NLA) and the Rwanda Space Agency a first national land cover (LC) prototype has been developed using an adaptation of the automated procedure developed by De Simone L. et

al[1], in the absence of in situ data, yielding an overall accuracy of 75%. An optimized field survey has been designed to be implemented at national level in December 2023/January 2024, to develop the first national LC baseline map for the reference year 2023, with an expected accuracy above 95%. In 2024, in addition to the LC mapping, the crop type mapping and the crop yield activities will be implemented as well. The NISR is collecting georeferenced in-situ data through the Annual Agricultural Survey operation, georeferencing both crop field boundaries and parcel centroid and is an exceptional role model in this domain.

SDG monitoring and reporting

In collaboration with the Bureau of Statistics (BOS) Lesotho, FAO has developed a national SDG monitoring solution which allows to compute the SDG indicator 15.4.2, the Mountain Green Cover Index (MGCI) on an annual basis, in an automatic fashion from EO data. The solution is based on the MGCI methodology developed by De Simone at al., 2021 at global level, and optimized to Lesotho using national land cover data in conjunction with the JAXA global digital elevation model. Such solution has allowed the BOS to report on the MGCI statistics disaggregated at elevation zone level from the 2017 through 2022. As the new LC baseline is updated in the system, the MGCI indicator baseline is also updated with a cascade effect.

Crop type mapping, crop acreage

In Mali, EOSTAT is being implemented in collaboration with the Institut d'Economie Rurale (IER) from the Ministère du Développement Rural, the Cellule de Planification et de Statistique (CPS), and the Compagnie Malienne pour le Développement des Textiles (CMDT). An optimized field ad-hoc campaign has been jointly designed and tested to collect crop in situ data in 2023. The methodology conists in the following steps:

- Stratification based on cropping intensity (0% - 30% ; 30% - 60% ; 60% - 100%) calculated from the ESA WorldCover land cover map
- Random selection of 300 segments (500m X 600m) within the different intensity zones.
- Manual digitizing (on-screen) of homogenous crop block/parcel using Google Earth /Bing imagery for each segment
- Integration of the 4igitized map with an app called MapMe, used for the teams navigation (driving to the place of each segment);
- Survey based on ODK Collect, to collect field data (answering a form about crop type and crop aera);
- Qfield app, to confirm the crop block/parcel boundaries and edit them in case of errors

The work has focused on cotton, sorghum, maize, groundnut, and rice and the Area of Interest (AOI) was defined in the Malian cotton belt and more precisely in the Diola District. EO data was used to design the agricultural survey. In 2024, the data collected in the field will be used to produce a national crop type map and crop yield map, results will be compared with the estimates from the official annual agricultural survey.

In Senegal, EOSTAT was first launched in 2019 in collaboration with the Direction de l'Analyse, de la Prévision et des Statistiques Agricoles (DAPSA), the Agence Nationale de la Statistique et de la Démographie (ANSD) and the Centre de Suivi Ecologique (CSE), Sénégal. During the first phase of the project, crop data from the Enquete Annuelle Agricole (AAS) for the year 2018 was used to produce the first national crop mask and crop type map at the national scale for major and minor crops using a random forest classification. The maps' accuracies were 90% and 76% respectively for the crop mask for the crop type map. In 2022, an experimental optimized field survey was implemented jointly with DAPSA to demonstrate the added value of digiting parcel boundaries and centroids. Such in-situ data was then used to produce a crop mask and a crop type map with overall accuracy of 91.8% and 90.2% respectively. Finally, the crop acreage was estimated first directly from the in-situ data and then from the combination of in situ and EO data, hence from the crop type map. The exercise showed that by the integration of EO and in situ would reduce the coefficient of variation of the acreage estimation by 50% for millet and groundnut. This proves the added value of EO data in increasing accuracy, using free and open data. Achieving similar results through intensification of field sampling would have incurred higher costs. Since 2023, the AAS has officially introduced the georeferencing of crop parcel boundaries and centroids.

In Zimbabwe, EOSTAT is being implemented in collaboration with the Ministry of Agriculture, the Zimbabwe Space Agency (ZINGZA) and the Zimbabwe Statistics (Zim-Stat) under the framework of the Zimbabwe

[1] De Simone, L.; Ouellette, W.; Gennari, P. Operational Use of EO Data for National Land Cover Official Statistics in Lesotho. *Remote Sens.* **2022**, *14*, 3294. https://doi.org/10.3390/rs14143294

Emergency Food Production Project (ZEFPP) with financial support from the African Development Bank (Africa Emergency Food Production Facility). In 2023, an optimized field survey was designed implemented, collecting 600 circa in-situ data points, and a first national map of Winter Wheat has been developed 2023 with an overall accuracy of 98%.

Crop yield forecasting

Crop yield forecasting using EO data was piloted in Senegal and in Camerron, using a statistical approach and a process-based approach respectively. In Senegal, geo-located reference yield measurements were collected from hundreds of crop plots in the Nioro landscape. Depending on the crop, the size of the measurement square varied between 5 and 25 m². A regression model was developed using EO data as explanatory variable of the yield. Results showed a poor correlation, however the sampling of the crop squares appeared to be insufficient and likely main cause of the poor results. In Cameroon, a process-based model, the System Approach to Land Use Sustainability (SALUS). As in-situ data was not available the SALUS model was calibrated for five main commodities (Maize, Rice, Sorghum, and Cassava) using yield statistics from FAOSTAT. The model was used to produce crop yield prediction at national level, at 10 meters resolution for each commodity for every year between 2011 through 2019.

Data innovation for Closing Data Gaps and Achieving Timely Insights

Data, an essential intangible asset, often faces availability challenges due to factors such as low statistical capacities, insufficient funding, and poor data dissemination and usage culture. To address these data gaps, it is vital to explore unstructured web data and combine it with innovative methods for information generation. National and international organizations are recognizing the need to adapt to new data sources and methods in light of the crisis facing traditional data collection systems. In response, the Food and Agriculture Organization of the United Nations (FAO) established the "Data Lab for Statistical Innovation" in 2019 (see Fabi et al., 2022). This initiative aims to modernize the Organization's statistical processes by enhancing the timeliness and granularity of data collection, enabling automated analysis, and facilitating the detection of early warning signals. The Data Lab employs advanced technologies like web scraping, text mining, geo-spatial data analysis, and artificial intelligence, along with nonconventional data sources such as social media and online newspapers.

While some domains experience an abundance of data (e.g., social networks, smart devices), obtaining information in others, like agricultural production losses, can be challenging and costly. The process requires designing surveys, compiling information, and standardizing results, necessitating coordinated efforts, financial resources, and human and technical capacities. Often, these resources are lacking in some countries, leading to infrequent or unavailable data. The situation is exacerbated by factors like decreased statistical capacities, reduced budgets, and weak regulatory frameworks.

The COVID-19 pandemic in 2020 further highlighted the importance of nonconventional data sources and innovative methods for improving coverage and providing quick insights. The pandemic accelerated the transformation towards these new approaches, emphasizing their urgency. A good overview of how organisations at the international level are employing these new methodologies and sources is reported in the latest United Nations Activities on Artificial Intelligence report (ITU, 2023), which showcases 281 innovative projects by 40 entities.

The FAO Data Lab's efforts have been instrumental in increasing the Organization's statistical coverage by filling data gaps. The following subsections aim to detail some of the steps taken by the Data Lab.

Obtaining Insights from News Articles

As the COVID-19 pandemic spread across the globe in 2020, the critical need for timely, actionable information became strikingly clear, particularly for supporting crucial decision-making processes in a rapidly evolving crisis. The pandemic highlighted the pressing demand for real-time or near-real-time information, with a specific emphasis on non-traditional data sources and the necessity of automated analysis of this data. To address these emerging requirements for data, the Data Lab for Statistical Innovation has been proactively involved in developing and managing various databases. One significant initiative initiated in January 2020 is the creation of a comprehensive database that includes tweets from over 500 global newspaper accounts. As of the end of April 2023 it contains around 28 million tweets and articles. This database covers multiple languages, such as

Arabic, English, French, Italian, Portuguese, Russian, and Spanish, and uniquely amalgamates the text of tweets with related news articles when both are available.

This database is subjected to a series of advanced Natural Language Processing (NLP) algorithms to enrich the data with valuable insights. A main component of this analysis is the identification and categorization of relevant topics. Tweets and articles are classified into categories relevant to the Sustainable Development Goals (SDGs), including "food value chain," "food prices," "climate change," "food security," and others. For this classification, we employ a machine learning classification model, specifically XGBoost, which has been trained on a dataset prepared and validated by human expertise across various topics and languages. The categorization of these topics is instrumental in facilitating focused analysis and generating insights into specific areas of concern. This approach enables the formulation of informed policies and strategic responses to crises.

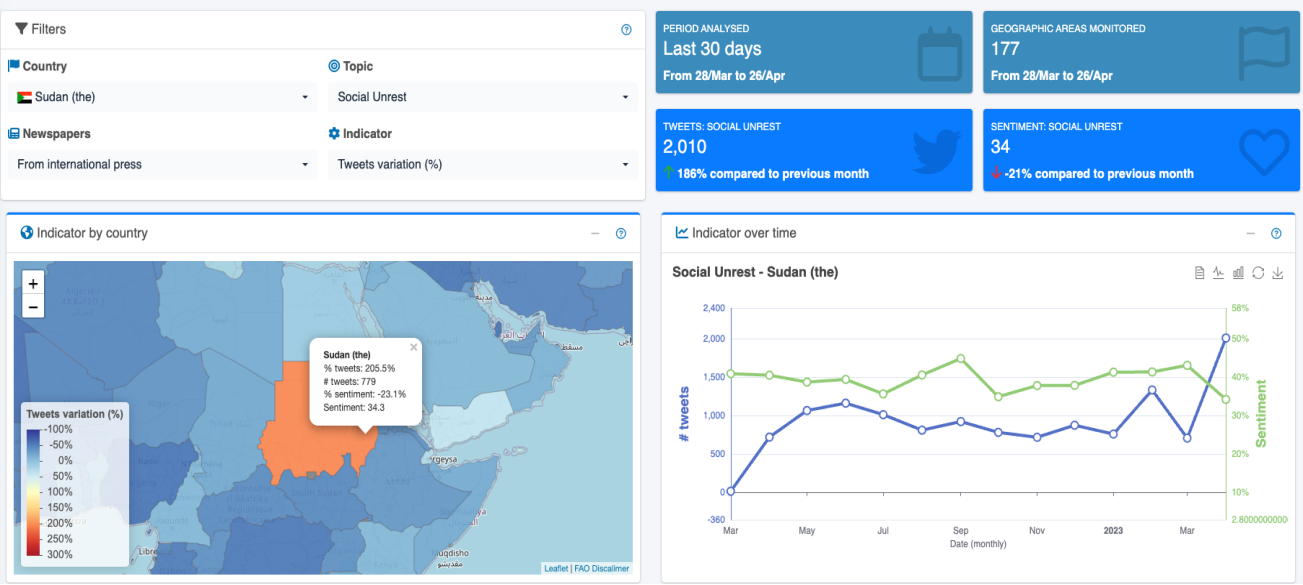
In addition to categorizing the articles into various topics, we also apply methods to extract mentions of countries within the texts and calculate a sentiment index. Identifying country mentions is particularly insightful as it reveals how different countries perceive or discuss various topics in relation to other nations. This analysis provides a unique view of international perspectives, showcasing how nations comment on and understand issues happening beyond their borders. Moreover, the sentiment analysis plays an important role in understanding the tone and nature of the coverage of various topics. By assessing the sentiment, we gain insights into how specific issues are portrayed and the prevailing mood or public opinion in different countries regarding these issues. This sentiment analysis is conducted using a dictionary-based approach, where words in the text are compared against a predefined lexicon consisting of positive and negative words. The formula applied is:

$$S = \frac{\#w^+}{\#w^+ + \#w^-} \times 100$$

where $\#w^+$ is the number of positive words, and $\#w^-$ is the number of negative words found in the text. The index goes from 0 to 100, where 0 represents a highly negative sentiment, 100 a highly positive sentiment, and 50 being a neutral text.

The information on topics and sentiments for all countries of the world are made available in a user friendly dashboard available at https://www.fao.org/datalab/early-warnings/topics-explorer/

The figure displayed below serves as a clear demonstration of the effectiveness of our analytical tool. It illustrates the way in which the topic of "social unrest" in Sudan captured significant attention from the international press. More importantly, it demonstrates a noticeable, consistent decline in sentiment as conflicts in the country escalated. This decline in sentiment effectively mirrors the increasing gravity and intensity of the situation in Sudan, as reported by media outlets worldwide. Such a visualization not only highlights the tool's capability in tracking global media focus on specific events but also its proficiency in capturing the evolving emotional tone of the coverage, which serves as early warning of possible issues. This analysis provides a dynamic view of how international perceptions change in response to developing situations in different countries, offering valuable insights into the global reaction to significant events.

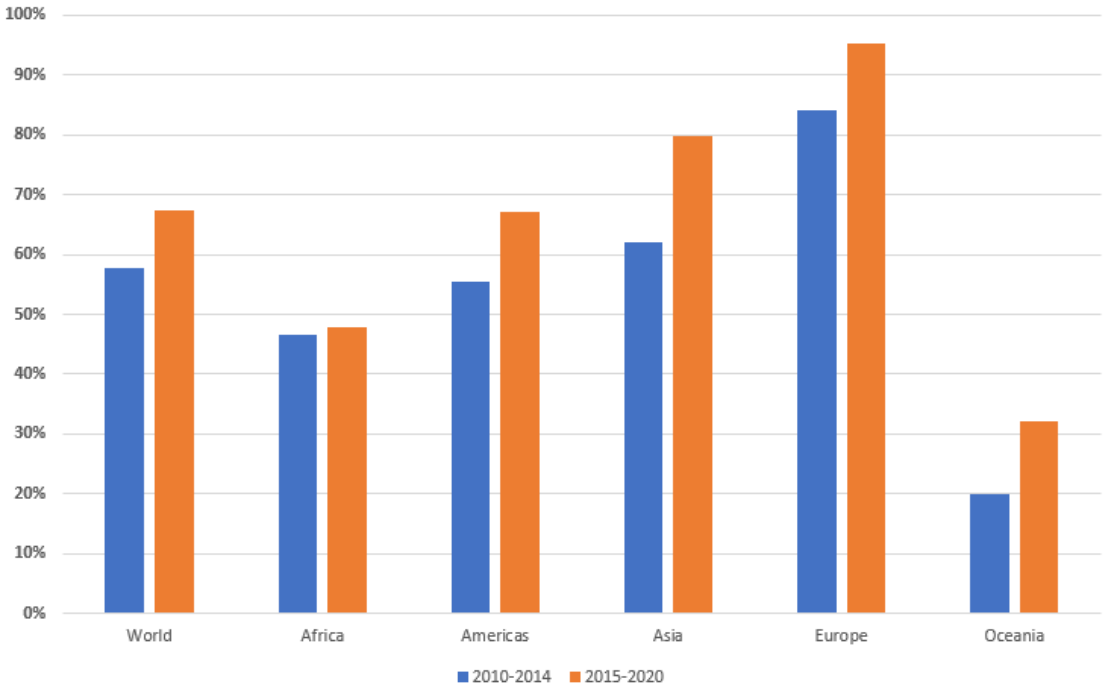


Filling Data Gaps in official production and area statistics

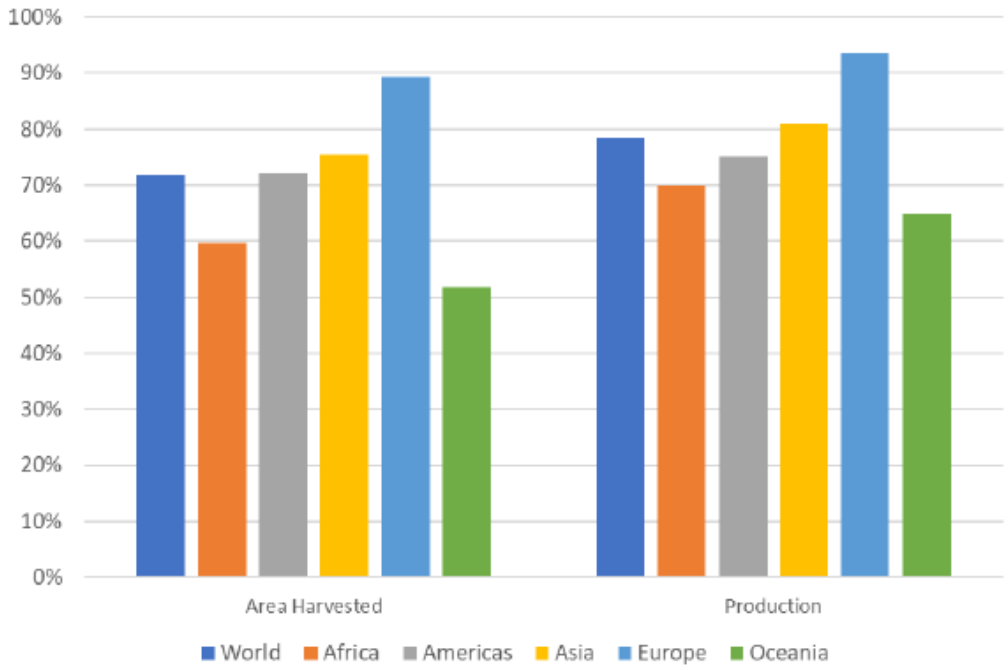
The Hand-in-Hand (HiH) initiative by the FAO focuses on accelerating agricultural transformation and sustainable rural development to tackle poverty and hunger, aligning with Sustainable Development Goals. It targets countries with limited capacities or facing crises, utilizing advanced geo-spatial modelling and analytics to enhance incomes and reduce vulnerabilities in rural populations. The initiative seeks to improve policy interventions, finance, investment, and institutional reforms.

Central to HiH is the Data Lab's role in agricultural data collection, particularly in countries where data are scarce. This is achieved through artificial intelligence techniques, focusing on sub-national data to fill gaps and validate existing data, thereby enhancing data quality and supporting government improvements.

The figure below shows the average response rates Agricultural Production Questionnaires (APQs), i.e., percentage of questionnaires received to questionnaires sent out. While the response rate to APQs has shown an overall increase during recent years, there are some areas for which it is well below 50% (Oceania) or consistently slightly below it (Africa).



A comparable view emerges when we calculate an APQ "completeness indicator," which is the proportion of official data points relative to the total potential combinations of country/commodity for which at least one official data point has been provided in an APQ. The results of this calculation are presented in the figure below. Together with the previous figure, it becomes evident that there is ample scope for enhancing the extent of our statistical coverage.



To bridge these gaps, the Data Lab uses web scraping to gather data from online sources like National Statistical Offices or Ministries. This process includes extracting data from various formats, standardizing commodity and administrative names, and resolving language barriers.

The web scraping process involves challenges, such as handling diverse document formats and translating non-English documents. The standardization of commodity and administrative names employs a fuzzy matching approach, with expert consultation for unmatched cases. As of October 2023, the Data Lab has collected a substantial number of data points on various agricultural commodities. This data is used for cross-checking and supplementing FAOSTAT's agricultural production data, improving the overall dataset's quality and comprehensiveness.

The data collected not only helps in enhancing the existing dataset but also plays a crucial role in identifying and rectifying discrepancies in the reported data, thus contributing significantly to the overall accuracy and reliability of agricultural statistics.

To give an idea of the results of the data gap resolution by web scraping and text mining as illustrated above, we report in the next table the results of the 2023 exercise on the agricultural production dataset. It shows the number of products that were web-scraped, and the number of products that were integrated (or accepted) in the agricultural production dataset. It also shows the actual number of data points accepted for each product, and the percentage of products that were integrated on the total of web-scraped products.

Country	Products scraped	Accepted		
		Products	Data points	% products
Afghanistan	20	0	0	0%
Bangladesh	85	4	15	5%
Bhutan	63	11	35	17%

Burkina Faso	13	0	0	0%
Burundi	13	5	5	38%
Cambodia	44	9	121	20%
DR Congo	15	6	10	40%
Dominica	39	29	200	74%
Ecuador	32	6	141	19%
Gabon	20	9	39	45%
Gambia	15	0	0	0%
Haiti	14	2	3	14%
Indonesia	22	0	0	0%
Laos	12	0	0	0%
Lesotho	7	4	13	57%
Mali	6	0	0	0%
Mozambique	20	7	11	35%
Nepal	9	0	0	0%
Niger	17	2	13	12%
Pakistan	33	0	0	0%
Rwanda	65	6	12	9%
Senegal	11	4	43	36%
Zambia	27	12	50	44%

The data presented in the previous table allows to obtain a significant result: on average, 20% of the products acquired through web-scraping have been successfully integrated into the agricultural production dataset. This incorporation highlights the efficacy of employing web-scraping and text-mining methodologies in enhancing data availability. Such methods not only broaden the scope of the dataset but also contribute to its richness and accuracy.

Moreover, the successful integration of a substantial proportion of web-scraped data underscores the potential of these innovative techniques in filling data gaps and updating existing datasets with more current and comprehensive information. The process of web-scraping, in particular, enables the extraction of a wide array of data from diverse online sources, thus providing a more comprehensive view of the agricultural sector. Similarly, text-mining facilitates the extraction and analysis of relevant information from large volumes of unstructured text, ensuring that the dataset is not only extensive but also insightful.

Food Loss Data Extraction via Text-Mining

SDG Target 12.3 aims to halve global food waste per capita by 2030 and reduce food losses in production and supply chains. To monitor this, the Food Loss Index and the Food Waste Index have been developed, managed by FAO and the UN Environmental Program respectively. The Food Loss Index tracks agricultural products lost before reaching the retail stage, often at production, post-harvest, and distribution stages. In contrast, the Food Waste Index focuses on food waste at the retail and consumer levels, involving good quality food that is discarded.

However, data on food loss and waste are scarce. As of May 2023, only 7% of food loss data was officially reported, mostly through questionnaires by FAO to countries. While there is a substantial amount of literature providing estimates of food loss, official statistics remain limited. To address this gap, FAO has developed a semi-automated text-mining process to extract food loss data from various documents, including scientific articles, working papers, and grey literature. This process involves three main steps: automated document collection and preprocessing, a statistical model assessing the relevance of documents, and guided data extraction. The process starts by querying for documents using loss-related keywords (e.g., "post-harvest loss") in scholar search engines, followed by extracting metadata like authors, titles, and publication dates. A text summarization routine and keyword identification are also performed.

The relevance of the downloaded documents for actual loss numbers is determined using a machine learning classifiers, specifically a Random Forests model. Documents are then sorted by their relevance for data extraction, i.e., only document that are likely to have food loss percentages according to the model are kept while the others are discarded. The final step involves manual validation by analysts, who assemble and validate the information by using a web-tool that highlights in the text key elements such as product names (e.g., "maize", "rice", etc.) along with country names, percentages, where they are placed in a neighbourhood of words related to loss ("loss", "losses", "lost", etc.). This is a "semi-automated" process as there is a final step of validation by a human expert that can accept or reject the text found by the algorithms.

As of May 2023, this method has yielded a dataset with nearly 30,000 data points covering 127 countries and 147 commodities, primarily focusing on data from 2000 onwards. This comprehensive dataset is publicly available on the FAO website at the following link: https://www.fao.org/platform-food-loss-waste/flw-data/

The FLW dataset serves two main purposes. Firstly, it aids researchers and analysts in studying food loss and waste in a domain where data is sparse. Secondly, it is used by FAO to estimate missing data on primary product losses for all countries. To impute losses where data is missing, FAO has developed an imputation model that takes as input the official data and the data scraped as obtained for the FLW dataset. In particular, it is a random effects model that takes into account various factors like electricity and oil prices, weather, and agricultural investment. This model helps provide yearly loss estimates for country-commodity combinations where data is unavailable. By aggregating official and imputed data, the global level of food loss is estimated to be around 14% in 2020.

3. CONCLUSIONS AND RECOMMENDATIONS

Data gaps pose a significant challenge to the integrity of official statistics and, by extension, the efficacy of policy-making. The underrepresentation of certain population subgroups, the risk of biased conclusions, and the potential for perpetuating social inequalities are among the critical concerns associated with incomplete data. As the world grapples with unprecedented challenges such as the COVID-19 pandemic, the role of official statistics has never been more important. It is imperative that stakeholders in the statistical system advocate for changes that enhance data-sharing legislation, secure flexible and dependable funding, and modernize the statistical system to ensure its resilience and relevance in the face of future challenges.

In this context, we have highlighted how data science methods contribute significantly to the advancement of knowledge in geospatial information fields, enhance official statistics, and assist in acquiring insights in a timely manner.

EO data use is proving to be a relevant, innovative and cost-effective statistical tool that once successfully fine-tuned within an operational system can help NSOs in countries to improve the capacity to produce high quality, highly disaggregated and timely agricultural statistics that are usually costly and that provide estimates after the harvest, such as the traditional agricultural surveys.

Nevertheless, the uptake of EO data use is still problematic due to general lack of training data (in-situ data) of adequate quality. Existing Annual Agricultural Surveys (ASS) and periodic Agricultural Census (AC) implemented in countries generally do not yield data that is readily compatible for integration with EO data. In fact, they generally do include the georeferencing of the parcels centroids and boundaries. Furthermore, the distribution of samples across crop classes is suboptimal for the training of machine learning and deep learning algorithms. The sampling is unbalanced with more samples for major classes and few samples for minor classes, this leads to commission and omission errors by the classification algorithm. Furthermore, the field samples are seldom stratified by agroecological zones, and therefore they miss to represent the different growing conditions of crops in the country.

Under such conditions ad hoc surveys are necessary to demonstrate the use of EO data for agricultural statistics, however, this is not a sustainable solution and should be only adopted as a short-term solution.

- AAS and AC represent established mechanisms in countries for field data collection and are therefore the best potential source for in situ data that is sustainable. In this context, the most urgent recommendation for NSOs is to review the existing protocols for field data collection of in-situ data in the AAS and the AC. Adjustments should be made to ensure full compatibility with EO based crop mapping analysis requirements. In particular, NSOs should ensure that in AAS and AS:
    i) crop parcels centroid and the crop parcel boundaries are georeferenced.
    ii) crop cuts are georeferenced;
    iii) professional GPSs are used in the field and that georeferencing does not rely on mobile's internal GPS as these have very low positional accuracy the crops all sampled across the different growing conditions (e.g. agroecological zones AEZ);
    v) minor crops are sufficiently sampled, no less than 50 data points per crop per AEZ;Agroecological zones are considered for stratification.

- NSOs that are already georeferencing crop data within the AAS and are already using EO data to produce land cover maps and crop type maps could benefit from testing different analytical solutions, in order to maximize map accuracy and minimise the bias in area estimation.

- In line with this it is recommended:
    - for countries with high cloud frequency, to use radar data (Sentinel-1) in alternative or in addition to the optical data (Sentinel-2);
    - to experiment different combination of bands and vegetation, soil and water indexes derived from EO data to best discriminate between different crop class, especially in countries with mixed complex agricultural systems; consider using of mixed crop classes to map mixed cropping systems, or to switch to commercial satellite imagery at very high spatial resolution (from 5 meters to below meter resolution) to discriminate the mixed crop types individually.
- It is finally recommended that collaboration is established and or reinforced between NSOs and the relevant national space agencies and high-performance computing centres. The sustainability of EO based solutions deepens in fact also on the available infrastructure that can allow for efficient creation and analysis of dense satellite image time series. Such collaboration would be essential for the establishment of a national Satellite Image Data Cube. This will allow for dense EO time series analysis from multiple sensors, both optical and passive: the intensification of the image time series allows to better describe the phenology of each crop class and hence to better discriminate these. Countries should also consider the possibility to create their own national data cubes, in order to build their national crop monitoring systems. This would in turn address another key issue which is the confidentiality of the in situ-data, which countries may not share with the public. The use of common EO platforms such as Google Earth Engine, AWS, etc, imply the upload of in situ data in such systems, and this may constitute an issue. The establishment of national data cubes may solve such problem.

Next Steps

In the next years to come FAO will be continuing providing support to NSO's in countries in building capacity in the use of alternative big data sources such as Earth Observation dense time series to produce agricultural official statistics, crop yield and acreage improving accuracy, timeliness, and granularity as well as also introducing disaster dimensions as flood and drought have dramatic impacts on the yield.
FAO has in the pipeline a series of project for 2024/2025 in several countries in Africa. In Rwanda, Angola, Kenya, Mali, Ethiopia, Senegal, Uganda, Cameroon, FAO will support the analysis of satellite image time series, applying a machine learning algorithms as well as deep learning algorithms and the integration with field data from optimized and/or adjusted AAS.
Such endeavour of FAO will be carried out in collaboration and in alignment with the broader scope of the work of the joint UN Committee of Experts on Agriculture and the UN Committee of Experts on Big Data. Such joint Committee has the intent of supporting the modernization of national statistical systems globally using alternative big data sources and innovation. Such Committee is participated by NSOs, by international organizations, by Academia. Under such committee a specific Task Team on EO for Agricultural statistics has been established and is co-chaired by FAO, WB, and INEGI with a focus on remote sensing and EO big data.
The programme of work 2024-27 of the UN-CEAG/UN-CEBD include strengthening specific areas in order to address challenges that were identified by the TT during 2022/2023 as per the last report to the UN Statistical Commission.

Such areas include:

- The integrated use of Radar and Optical satellite images to crop type mapping and disaster monitoring.
- EO and drone-based surveys
- Enhancement of satellite time series analysis and development of Satellite Image Data Cubes
- Confidentiality of information and in situ data sharing.
- NSO access to data, tool and information generated by TT.
- Improvement and updating of the EO training app

In this context, it is worth noting that the UN Regional Big Data Hub for China has transitioned in 2023 from Regional to a Global Hub and has inaugurated a Remote Sensing Lab. The Global HUB has appointed international experts from FAO, INEGI and the National Institute for Space Research advisors to the Lab. This will greatly facilitate international cooperation with the Chinese Academy of Science and the Space Research Institute and allow cross-fertilization of projects through sharing of innovative methods and solution and access to very high spatial and temporal resolution images from the Chinese satellite GAOFEN.

In conclusion, the work of the joint UN-CEAG/CEBD on E for Agriculture is hereby recalled for its relevance as an ideal forum to collectively advance on key issues found in the adoption of EO data for official statistics.

In this paper we have showcased the experience in FAO with novel methods to fill data gaps and obtain insights. The FAO Data Lab for Statistical Innovation exemplifies the significant impact of data science methodologies and non-conventional data sources in transforming the field of statistical analysis and data collection. The Lab's integration of advanced technologies such as web scraping, text mining, and geo-spatial data analysis highlights the immense potential of data science to enhance the quality and scope of data. By tapping into diverse, unstructured data from sources like social media and online news outlets, scientific articles, reports, and other type of literature, the Lab successfully addresses and fills gaps in traditional datasets and providing new insights, thereby widening the range and enriching the depth and timeliness of data analysis.

Furthermore, the utilization of data science in processing and analyzing data introduces unprecedented levels of efficiency and precision. With the ability to navigate through vast data volumes, discern patterns, and extract pertinent information rapidly, these methodologies prove invaluable in a data-rich, complex world. Their offer indispensable insights into diverse global challenges, including food security, climate change, economic development, and public health. The capacity to analyze extensive datasets and derive meaningful insights is critical in tackling these multifaceted issues in our interconnected world.

The confluence of data science with non-conventional data sources is revolutionizing our approach to problem-solving and informed decision-making. By bridging data gaps and delivering timely insights, these innovative practices empower a more dynamic and enlightened approach across various fields. As technology evolves and data abundance grows, the possibilities for uncovering latent insights through these methods will undoubtedly increase, marking an exciting era for researchers, organizations, and policymakers.

The ongoing evolution of data science methodologies, particularly with the progressive application of Artificial Intelligence (AI), is set to continuously enhance data processing and interpretation. AI brings a transformative flair to the field, enabling more intricate analyses and the effective harnessing of extensive datasets. In an era increasingly reliant on data-driven insights for policy making and societal advancement, the amalgamation of AI with conventional statistical methods is vital. This integration promises to overcome previous limitations and unlock new avenues in data analysis, ensuring that official statistics remain relevant, reliable, and reflective of our dynamic world.

4. QUESTIONS AND INVITATIONS TO AFCAS MEMBER COUNTRIES

Members are invited to:

- Take note of the report and provide feedback as deemed necessary;
- Express their views on the relevance of the EOSTAT programme for the production of statistics in their respective countries;
- Share their experience and challenges on the EO data for land cover mapping, SDG indicator monitoring and reporting, including the Mountain Green Cover Index (MGCI), crop type mapping, crop acreage and yield estimates, and express their most pressing methodological and/or capacity development needs;
- Take note of the UN-CEAG/CEBD proposed areas of work for 2024-27, share recommendations and suggestions for the finalization of this programme of work and express their interest in becoming members of the task force;
- Share and discuss the primary challenges encountered in adopting data science techniques for statistical analysis, as well as explore exploration and identify the type of support needed to effectively overcome these challenges;
- Evaluate and share perspectives on the role of non-conventional data sources, such as social media and online news, in augmenting traditional statistical data. This discussion aims to understand if and how these sources can contribute to statistical systems development;
- Identify and discuss opportunities for public-private partnerships in the realm of data collection and analysis. This includes a focus on how non-conventional data sources can be integrated and leveraged in these partnerships;
- Articulate a vision for the future role of Artificial Intelligence (AI) in enhancing national data systems. If any activity is currently being adopted, elaborate on the steps being taken towards achieving this vision, fostering a collaborative discussion on the strategic integration of AI in data practices;
- Explore the opportunity to collaborate with the FAO Big Data Lab, in alignment with the funding obtained for a project on anticipatory action in emergencies and for policy making. This invitation is extended to countries seeking to strengthen their analytical capabilities in emergency response and policy formulation through advanced data science and AI methodologies. We encourage members to collaborate with the FAO Big Data Lab to enhance their data-driven strategies and decision-making processes.

5. REFERENCES

De Simone, L.; Navarro, D.; Gennari, P.; Pekkarinen, A.; de Lamo, J. Using Standardized Time Series Land Cover Maps to Monitor the SDG Indicator "Mountain Green Cover Index" and Assess Its Sensitivity to Vegetation Dynamics. ISPRS Int. J. Geo-Inf. **2021**, 10, 427. https://doi.org/10.3390/ijgi10070427

De Simone, L.; Ouellette, W.; Gennari, P. Operational Use of EO Data for National Land Cover Official Statistics in Lesotho. *Remote Sens.* **2022**, *14*, 3294. https://doi.org/10.3390/rs14143294

Fabi, C.; Rosero Moncayo; J., Mongeau Ospina; C.A., Silva e Silva, L.G. The FAO Data Lab on statistical innovation and the use of big data for the production of international statistics, Statistical Journal of the IAOS, **2022**, 38(3)

ITU (International Telecommunication Union), *United Nations Activities on Artificial Intelligence (AI) 2022*, **2023**. https://www.itu.int/pub/S-GEN-UNACT-2022