A Review of Big Data and Machine Learning Operations in Official Statistics: MLOps and Feature Store Adoption

Carlos Eduardo Ramos Nunes

NOVA Information Management School (NOVA IMS), Campus de Campolide, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal carlosrnes@gmail.com

Afshin Ashofteh

NOVA Information Management School (NOVA IMS), Campus de Campolide, Universidade Nova de Lisboa, 1070-312 Lisboa, Portugal aashofteh@novaims.unl.pt

https://orcid.org/0000-0001-5075-9822

This is the final accepted version of the conference paper published by *IEEE*

Nunes, C. E. R., & Ashofteh, A. (2024). A Review of Big Data and Machine Learning Operations in Official Statistics: MLOps and Feature Store Adoption. In H. Shahriar, H. Ohsaki, M. Sharmin, D. Towey, AKM. J. A. Majumder, Y. Hori, J-J. Yang, M. Takemoto, N. Sakib, R. Banno, & S. I. Ahamed (Eds.), 2024 IEEE 48th Annual Computers, Software, and Applications Conference: COMPSAC 2024 (pp. 711-718). (Proceedings of the IEEE Annual Computer Software and Applications Conference). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/COMPSAC61105.2024.00101

A Review of Big Data and Machine Learning Operations in Official Statistics: MLOps and Feature Store Adoption

Carlos Eduardo Ramos Nunes
NOVA Information Management School (NOVA IMS), Nova University of Lisbon, Lisbon, Portugal
carlosrnes@gmail.com

Afshin Ashofteh
corresponding author and guarantor
NOVA Information Management School (NOVA IMS), Nova University of Lisbon, Lisbon, Portugal
aashofteh@novaims.unl.pt
https://orcid.org/0000-0001-5075-9822

***Abstract*— Integrating machine learning (ML) into the official statisticians' toolset is gaining popularity as National Statistical Offices (NSOs) strive to improve their methodologies. This trend poses new challenges and implications for incorporating innovative techniques that ensure the reliability of the official statistical production process. A comprehensive literature review was conducted using Scopus and Web of Science databases to explore the contemporary applications of data science in official statistics. A total of 178 research articles were identified, focusing on areas such as big data, machine learning, and data quality. While the literature review revealed extensive proposals on utilizing alternative data and applying machine learning techniques to support official statistics production, it also identified research gaps in the post-training steps of the machine learning process. Areas requiring further investigation include machine learning operations in a production environment, data quality assurance, and governance.**

Keywords*— *Feature store, Official statistics, Machine learning operations, Data science, Big data, Data quality

I. Introduction

Integrating big data and Machine learning (ML) applications into the official statisticians' toolset is becoming more popular as national statistical offices (NSOs) improve. New methodologies such as data science and ML are applied to official statistics, as discussed in a research paper that proposes a Data Science Model for Official Statistics (DSMOS)[1]. In another way, studies have emerged that solve specific official statistical problems, such as predicting individual and group behaviours based on mobile phone data [2] and establishing a relationship between personal expenditure patterns and socioeconomic indices of life quality using card transaction data [3].

All such systems consider a production environment that upholds compliance and adheres to some fundamental software development principles over an extended period. ML systems possess a unique capability to accumulate technical debt, which is described as the cumulative cost incurred when development teams consciously take actions to expedite the delivery of functionality or a project. These shortcuts, made for speedy development, later require remediation through refactoring. This is particularly pertinent for ML systems, as they face all the maintenance challenges of traditional code besides a distinct set of ML-specific issues. Where this debt arises at the system level rather than the code level, it may prove difficult to detect. The data can affect the behavior of ML systems and undermine or invalidate traditional abstractions and boundaries. To counter ML-specific technical debt at the system level, the conventional methods employed to reduce technical debt at the code level are inadequate [4].

Considering the aspects mentioned above, the Machine Learning Operations (MLOps) concept has acquired prominence. MLOps encompasses a range of practices aimed at deploying and maintaining ML applications in production, with both reliability and efficiency. Reduced model iteration time, increased model reliability, preserved compliance, and improved teamwork are the aims of MLOps. Key principles for MLOps include iterative-incremental development, automation, continuous deployment, versioning, testing, reproducibility, and monitoring. Moreover, the required roles and their intersections (Fig. 1) highlight that robust ML solutions involve more than an algorithm, as demonstrated by Kreuzberger (2022) [5].

In 2017, the inaugural feature store designed for industrial use was established. The objective was to equip engineers with the requisite tools to consume training data for feature curation, model training, and deployment, as well as the monitoring and maintenance of models post-deployment. The Feature Store (FS) was developed to tackle these challenges, offering a centralized repository of reusable features in ML pipelines, along with automatic management of that pipeline [6].

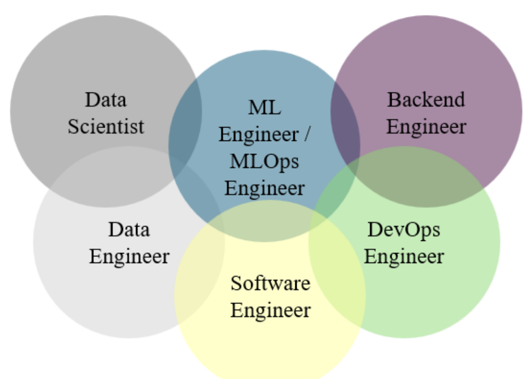

Fig. 1. Roles and their intersections contributing to the MLOps paradigm

The connection between the concepts of MLOps and FS is verified within the MLOps Architecture proposed by Mboweni (2022) [7]. As a technical component, a feature store system ensures that frequently used features in making predictions are centrally stored. It is set up with two databases, one of which is configured as an online feature store to serve features with minimal latency for production forecasts and the other as an offline feature store to serve features with typical latency for experimentation. Examples include Tecton.ai, Hopswork.ai, Amazon AWS Feature Store, and Google Feast. Most of the data used to train ML models will come from here. Additionally, data may originate straight from any type of data store [5].

II. RESEARCH FRAMEWORK AND METHODOLOGY

A. *Problem Statement and Justification*

As discussed in the introduction, MLOps and FS could integrate big data and ML into the official statistics. However, the principles to generate official statistics need to be guaranteed. The United Nations Economic Commission for Europe (UNECE) suggests that ML solutions should adhere to the Quality Framework for Statistical Algorithms (QF4FA), which encompasses five quality dimensions: accuracy, explainability, reproducibility, timeliness, and cost-effectiveness [8]. According to the QF4FA and due to the social impact, big data and ML solutions for official statistics require more rigorous controls compared to business solutions. This emphasizes the importance of studying which big data and ML solutions have been proposed for official statistics and if there are robust management solutions to achieve quality dimensions like explainability, reproducibility, timeliness, and cost-effectiveness.

B. *Research Questions*

In response to the previously outlined issues, we have articulated two research questions that will be the focus of our investigation in this systematic review:

1. What approaches have been employed to implement manageable ML solutions in a production environment applied to official statistics?
2. How can a combination of MLOps and feature store architecture contribute to managing and controlling ML processes in Official Statistics by considering necessary quality dimensions?

The two research questions at the core of our systematic review serve two main purposes. First, they seek to uncover successful strategies for implementing ML solutions in the context of official statistics. This could provide valuable insights for statisticians and organizations aiming to leverage ML in their statistical practices. Second, these questions aim to explore whether the combination of MLOps and feature store architecture can enhance the management and control of ML processes in Official Statistics.

C. *Methodology*

To answer these questions, on December 22, 2022, we first conducted a systematic review analysis of contemporary data science solutions for official statistics. Second, we complete it by focusing on big data, ML, and feature store.

We conducted a database search using the preferred reporting items for systematic reviews and meta-analysis methodology. The first step in the paper selection process involved screening the Scopus and WOS databases from 2018 to 2022, restricted to only papers written in English, using the follows keywords: “official statistics” AND (“feature store” OR “machine learning” OR “ML” OR “machine learning operation*” OR “MLOPS” OR “data science” OR “big data” OR “data quality”). To ensure the relevance of the results, we narrowed down the search to specific fields, including economics, econometrics, finance, business, administration, accounting, social sciences, decision science, engineering, mathematics, and computer science. The search results of scholarly articles from the two databases were then merged, and duplicates were removed.

III. RESULTS AND CONTENT ANALYSIS

A. *Results*

In the first search focused on solutions for official statistics to perform network analysis, we utilized VOSviewer to display the extracted data from the dataset. The process involved setting a minimum occurrence threshold for keywords, initially at one occurrence, resulting in a total of 822 keywords. Subsequently, we increased the minimum occurrence to two, yielding a total of 115 keywords. Figure 2 displays the application of this latter keyword set over time, where official statistics, big data, and machine learning were the top three keywords. Notably, some data sources were identified as keywords, such as Google Trends, mobile phone data, and social media.

In the second search focused on the ML feature store, Scopus and WOS reported 15 papers with the expression feature store inside the title, abstract, or keywords. However, analysing the content of papers, only five papers effectively discuss some aspects of the ML feature store.

B. *Content Analysis*

The first search yielded 208 articles, as discussed in the preceding section. After this, we excluded review articles and those not directly relevant to data science. In total, 30 articles were excluded.

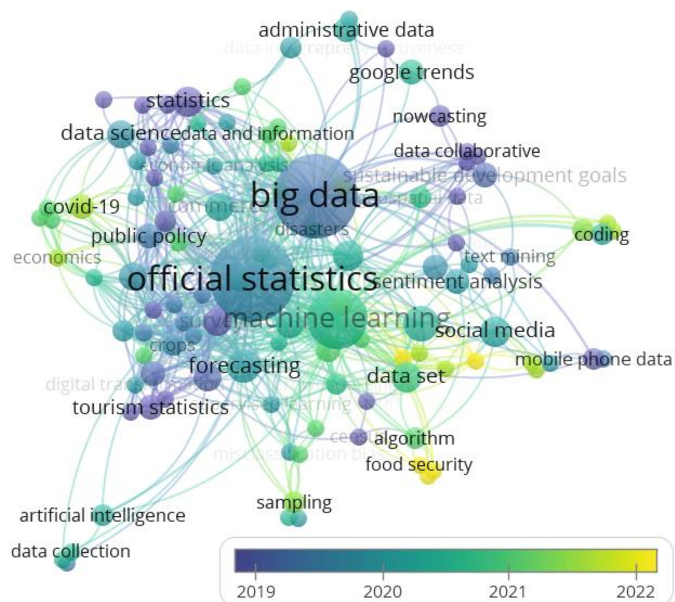

Fig. 2. Application of keywords over the years for the search for official statistics

The 178 articles that remained after the exclusion process were subjected to analysis and classified into four categories, namely country, main content, data source, and method. The country category pertains to the nation where the data was gathered and studied. The main content category refers to the focus of the study concerning data science, including machine learning, big data, data quality, and machine learning operations. The data source category identifies the origin of the variables utilized as inputs, such as mobile phones and social media. Finally, the method category indicates the ML method employed when applicable. This classification was carried out to accentuate the key features of the literature that dealt with the research subject, and the final dataset is available in the data.menedley [9].

As per Table I below, an analysis of the found articles with respect to their main content (feature store, machine learning, data science, big data, data quality, or MLOps) reveals that most of the articles, constituting 24.72%, focus on the utilization of Big Data as an alternative data source. ML is the primary subject in 22.47% of the articles, while 21.35% delve into the challenges and implications of Big Data.

TABLE I. CLASSIFICATION OF THE REVIEWED ARTICLES BY MAIN CONTENT OF RESEARCH

Main content	Number	Percentage
Big Data as an Alternative Data Source	44	24.72%
Machine Learning	40	22.47%
Big Data Challenges and Implications	38	21.35%
Big Data and Machine Learning	27	15.17%
Data Quality	18	10.11%
Big Data and Data Quality	5	2.81%
Data Quality and Governance	5	2.81%
Big Data Infrastructure	1	0.56%

It is evident that nearly 60% of the studies specify the country whose data was analysed. Approximately 10% of the studies used North American data, and another 10% used data from Italy. Surprisingly, despite China's significant academic productivity, only one article utilizes Chinese data. Additionally, approximately 10% of the studies use data from the European Union.

Regarding the input data used in the studies, *mobile phone data* was the most popular, used in about 12% of the 109 studies proposing data applications, followed by *social media data*, which was used in 11% of the studies. Only one study explicitly used both. A few studies applied *geospatial data* in their proposals.

Of the 42 articles that used prediction methods, about 19% utilized a *regression model*, and a similar percentage used *ensemble-based methods*. Other models, such as support vector machines, Bayesian approaches, or time series models, were less commonly used. In addition, studies utilizing *classification* or *regression techniques* other than *neural networks* and *random forests* accounted for roughly 53% of the total articles reviewed.

Finally, it is worth noting that the articles' proposals varied depending on their main content. Generally, alternative data sources were proposed to support official statistics. However, some studies focused on highlighting big data challenges and implications, while others applied ML techniques to improve data quality. Notably, there were no papers available on machine learning operations or feature stores.

1) Most-cited articles

Table 2 lists the most cited articles identified by the literature review conducted, and studies around COVID-19 were more frequent in the last 5 years. Ashofteh & Bravo (2020) [10] conducted a comparative analysis of datasets used in modeling respiratory diseases at an international level. Their findings highlight the significance of public health information systems as institutionally based repositories for electronic data collection and dissemination, along with tools for data quality assessment. Their conclusion underscores the low quality of official data on the COVID-19 pandemic in its primary stage and the growing importance of utilizing information from both traditional and unconventional sources with web scraping and ML to obtain high-quality data during a crisis to have better official statistics and policy decisions. From QF4FA point of view, this article covers accuracy, explainability, and timeliness dimensions.

Ghaffarzadegan & Rahmandad, (2020) [11] offered a simulated approach to estimate the epidemic magnitude. They used a dynamic model in this study based on the official statistics of show the COVID-19 pandemic. According to them, using information from multiple sources, they hypothesize that the official statistics available were at least a factor below the true extent of the pandemic. Even under the most hopeful scenarios, the disease's toll was significant and lingered for many months. Authors recongnised accuracy, explainability, and timeliness as aspects related to the QF4FA for this article.

The article of Wang et al. (2022) [12] re-calculated the pandemic's impact on American oil consumption by comparing consumption in 2020 with simulated values. Their hybrid simulation approaches, combining simulation methodology and ML techniques, produced low Mean Absolute Percentage Errors (MAPE), indicating reliable forecasting of U.S. oil consumption statistics. The paper focuses on accuracy and reproducibility and indirectly touches on explainability, making it relevant to the QF4FA.

Levin et al. (2018) [13] introduced an innovative method for monitoring conflict and refugee crises by combining data from various sources, including night lights and social media, to measure conflict levels in Arab nations post-Arab Spring. This approach has potential applications in environmental monitoring and geopolitical analysis on Earth. The authors considered accuracy and reproducibility as the direct quality dimensions from QF4FA for this article, and indirectly explainability and timeliness.

Fatehkia et al. (2018) [14] is a highly cited article in the database, with 60 total citations and an annual reference rate of 12. The study focused on utilizing Facebook data to estimate gender disparities in internet and mobile phone access across over 150 countries. The authors successfully established a robust correlation between Facebook data on users by age and gender and official data on digital gender inequalities. Their method highlighted the potential of leveraging web data to enhance a critical development indicator, particularly benefiting low-income nations. For QF4FA dimensions, the authors evaluated the paper primarily related to accuracy dimension and indirectly touches on reproducibility, explainability, and timeliness.

The study by Saluveer et al. (2020) [15], explained how to extract tourism statistics from mobile phone roaming records and demonstrated a strong correlation with accommodation official statistics. They compared the statistics collected from mobile positioning data and statistics about accommodation, concluding that locating data enables the creation of precise tourism statistics. The authors evaluated the paper's content support accuracy, cost-effectiveness, and timeliness of the QF4FA.

Another applicability for a mobile phone dataset analyzed by Vanhoof et al. (2018) [16] was the performance of five home detection algorithms (HDAs) applied to a large French Call Detailed Record (CDR) dataset, promoting a set of recommendations. According to them, mobile phone data presented fascinating new challenges and potential for official statistics, particularly in recognizing house locations, which is likely the most significant issue impeding the use of mobile phone data in official statistics. For the QF4FA, the authors found accuracy, explainability, and timeliness dimensions relevant.

Regarding the natural language process (NLP), Loureiro & Allo (2020) [17] examined the perceptions and emotions toward climate change in the United Kingdom and Spain in early 2019 using NLP. They explored sentiments in tweets, advocating social media for uncovering views not addressed by official statistics. Accuracy and timeliness considered as the relevant dimensions of the QF4FA.

Zhang et al. (2019) [18] proposed an ML system predicting agricultural planting maps, achieving 88% congruence with the Cropland Data Layer (CDL) for future seasons. Their study emphasized the importance of accuracy and timeliness in QF4FA, aligning with previous findings.

Finally, Batista e Silva et al. (2020) [19] proposed a new population grid dataset at one km² resolution for the European Union that accurately captures the population, day and night, for each cell. They introduced a multi-layered dasymetric methodology that created and validated a population grid dataset for the entire European Union by fusing geospatial data from new sources with official statistics. Accuracy was candidate as the direct dimension of the QF4FA for this paper.

To sum up, the top 10 most cited articles listed in Table 2 demonstrate that the literature about official statistics using ML and big data concerns different data sources for different applications to compound a better perspective about official statistics. Using multi-sources applying data quality methods often generates better predictions supported by ML operations and feature stores applied for official statistics was found.

TABLE II. THE 10 ARTICLES MOST CITED IN THE COMPILED DATABASE

References	Title	Cit.	Cit. per year	ML Algorithms	Features Used	QF4FA[a]
Wang et al. (2022)	Impact of COVID-19 pandemic on oil consumption in the United States: A new estimation approach	25	25.00	Regression models (e.g., linear regression)	Oil consumption data, pandemic-related variables	Accu. Repr.
Fatehkia et al. (2018)	Using Facebook ad data to track the global digital gender gap	60	12.00	Classification algorithms (e.g., logistic regression, decision trees)	Facebook ad data (user demographics, ad impressions)	Accu.
Ghaffarzadegan & Rahmandad (2020)	Simulation-based estimation of the early spread of COVID-19 in Iran: actual versus confirmed cases	35	11.67	Agent-based simulation models	COVID-19 case data, mobility patterns	Accu. Expl. Time.
Saluveer et al. (2020)	Methodological framework for producing national tourism statistics from mobile positioning data	32	10.67	Clustering algorithms (e.g., k-means)	Mobile positioning data (GPS traces)	Accu. Cost. Time.
Loureiro & Allo (2020)	Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the UK and Spain	31	10.33	Natural language processing (NLP) models (e.g., sentiment analysis)	Social media posts (Twitter, Facebook)	Accu. Time.
Ashofteh & Bravo (2020)	A study on the quality of novel coronavirus (COVID-19) official datasets	26	8.67	Descriptive statistics, data profiling	COVID-19 official datasets (case counts, testing rates)	Accu. Expl. Time.
Levin et al. (2018)	Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study	35	7.00	Remote sensing algorithms (e.g., vegetation indices)	Satellite imagery, conflict event data	Accu. Repr.
Batista e Silva et al. (2020)	Uncovering temporal changes in Europe's population density patterns using a data fusion approach	17	5.67	Data fusion techniques (e.g., weighted averaging)	Population density data (census, remote sensing)	Accu.
Zhang et al. (2019)	Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps	21	5.25	Machine learning models (e.g., random forests)	Historical crop planting maps, climate data	Accu. Time.
Vanhoof et al. (2018)	Assessing the quality of home detection from mobile phone data for official statistics	25	5.00	Not specified	Home detection data (mobile phone records)	Accu. Expl. Time.

[a] Five quality dimensions of QF4FA: Accuracy, Explainability, Reproducibility, Timeliness, and Cost-effectiveness.

2) Most recent articles

The following paragraphs provide a review of the most recent articles examined in this bibliometric study. Table 3 lists the ten most recent articles that summarize recent discussions on applying data science to official statistics. The topic of COVID-19 is still present, but less frequently, giving way to topics such as nowcasting. In this last, Hopp (2022) [20] demonstrated the usefulness of extended *short-term memory networks (LSTMs)*, an artificial neural network (ANN) type, for forecasting economic time-series, as compared to the dynamic factor model. They claim that one of LSTM's additional benefits is its capacity to handle many input features at various temporal frequencies. The stochastic nature of outputs, which applies to all ANNs, is a drawback.

Corona et al. (2022) [21] introduced a novel nowcasting method for the Mexican Global Economic Activity Indicator (IGAE) using dynamic factor models (DFMs) and non-traditional variables like Google Trends. They found this approach effective, especially for short-term estimates during the COVID-19 pandemic.

For estimating small area averages, also present in Table 3, the article of Krennmair & Schmid (2022) [22] offered a cogent framework based on mixed effects random forests and suggested a non-parametric bootstrap estimator to measure the accuracy of the estimates.

The study by Abraham (2022) [23], also displayed in Table 3, discussed statistical challenges related to new forms of data beyond surveys and administrative sources. The study explores statistical hurdles with non-traditional data. They highlight integrating big data into official economic statistics amid pandemic-driven shifts. Existing data systems struggled to meet policymakers' evolving needs promptly.

Complementing the study above, Tillé et al. (2022) [24], in their article, examined the value of big data in official statistics. They emphasized the need to enhance techniques for integrating various data sources as they become more accessible. The study highlighted that official statistics share similarities with other statistics in terms of applying probability theory and statistical inference. Additionally, it stressed the importance of carefully selecting data sources and procedures for effective integration.

Meertens et al. (2022) [25] introduced a strategy for ML algorithms in official statistics, addressing concept drift. They proposed using bias correction techniques to minimize prejudice when updating models due to the availability of new data. The authors compared two techniques, the misclassification estimator and the calibration estimator, focusing on predicting category proportions. Their findings highlighted the bias and variance of each approach, establishing a boundary for their relative effectiveness.

Still concerning the concept drift, the article of Van Delden et al. (2022) [26] addressed techniques for quantifying the bias and variance of statistical approximations of statistical indicator ratios that were influenced by classification errors. The techniques were used in two very different case studies, one on categorizing errors in registers and the other on categorizing errors made by a ML classifier.

Also present in Table 3, the article of Perepi et al. (2022) [27] suggested that many more cases of COVID-19 than those included in the official statistics may have existed in various countries. They advocate statistical modeling and simulations to forecast the pandemic's global spread, considering factors like distance and travel restrictions. Linear regression analyzes the virus's spatiotemporal pattern.

TABLE III. THE 10 MOST RECENT ARTICLES IN THE COMPILED DATABASE

References	Title	Source	ML Algorithms	Features Used
Krennmair & Schmid (2022)	Flexible domain prediction using mixed effects random forests	Journal Of The Royal Statistical Society Series C-Applied Statistics	Mixed effects random forests	Domain-specific features
Abraham (2022)	Big Data and Official Statistics	Review Of Income And Wealth	Not specified	Not specified
Perepi et al. (2022)	Predictive analysis of COVID 19 disease based on mathematical modelling and machine learning techniques	Computer Methods In Biomechanics And Biomedical Engineering-Imaging And Visualization	Machine learning techniques (not specified)	COVID-19-related data (e.g., case counts, demographics)
Corona et al. (2022)	Timely Estimates of the Monthly Mexican Economic Activity	Journal Of Official Statistics	Not specified	Economic indicators (e.g., GDP, employment)
Hopp (2022)	Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM)	Journal Of Official Statistics	Long Short-Term Memory (LSTM) neural networks	Economic time series data (e.g., GDP, inflation)
Meertens et al. (2022)	Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms	Journal Of Official Statistics	Machine learning algorithms (not specified)	Official statistical data (e.g., census data)
Van Delden et al. (2022)	Accuracy of Estimated Ratios as Affected by Dynamic Classification Errors	Journal Of Survey Statistics And Methodology	Not specified	Ratios based on classified data
Bandam et al. (2022)	Classification of Building Types in Germany: A Data-Driven Modeling Approach	Data	Data-driven modeling (not specified)	Building-related features (e.g., architectural characteristics)
Heikinheimo et al. (2022)	Detecting country of residence from social media data: a comparison of methods	International Journal of Geographical Information Science	Not specified	Social media posts, user profiles
Tillé et al. (2022)	Some Thoughts on Official Statistics and its Future (with discussion)	Journal Of Official Statistics	Not applicable (discussion paper)	Not applicable (discussion paper)

Social media keeps being researched to support the traditional official statistics process in the articles listed in Table 3. Heikinheimo et al. (2022) [28] used Instagram data to pinpoint Kruger National Park visitors' home countries. They found social media metrics like travel frequency and posting habits more revealing than aggregate data. Their study adds to discussions on biases in mobile big data. Finally, the article by Bandam et al. (2022) [29] introduced building classification models for predicting building types based on features and models created by combining the OpenStreetMap buildings' data and other datasets and applying ML algorithms. The data science literature for official statistics continues to expand on a range of topics, as evidenced by the articles listed in Table 3. Despite studies on challenges and implications, new themes such as nowcasting emerge in the most recent publications.

3) *Machine learning feature store*

The content analysis focused in the second search around feature store presented the 5 papers cited below. The first article by Ormenisan et al. (2020) [30] proposed a bottom-up provenance framework for capturing ML provenance that informs potential users where information comes from, demonstrating how the information has changed and why it has changed meaningfully to other components in a distributed ML workflow. Patel (2020) [31] proposed ML Feature unification to mitigate problems with inconsistent ML training and testing. This paper presented details of methods to achieve ML feature unification. Another article by Patel (2020) [32] introduced the concept of democratization of ML features, one of the emerging themes associated with ML application development. ML Feature store to democratize various features in ML development. The article by Orr et al. (2021) [6] addressed the difficulties and solutions associated with handling embedding-centric pipelines in feature store systems. Lastly, Quon & Gaudiot (2022) [33] proposed using recently emerged feature stores to handle concept drift on real world applications that allow data models to process multiple, ever-changing inputs.

To summarize, it was impossible to explore the concept of feature store in Scopus and WOS more deeply, and the other documents do not address ML, although they contain the term feature store. As a result, ML and feature store adoption for official statistics is recognised as a research gap for future studies by authors.

IV. Conclusion

Regarding the first research question that intend to understand what has already been produced in terms of ML solutions applied to official statistics. This research demonstrated the application of big data and ML methods in the domain of official statistics. The study reveals extensive proposals for alternative data sources and ML to enhance the production of official statistics. These developments in data sources offer research opportunities for further improving official statistics.

The second research question, related to MLOps and feature store applied in Official Statistics ML solutions, reveals that while ML techniques are widely employed, more research is needed to address the operational aspects of ML, data quality, governance, and frameworks such as QF4FA. Nonetheless, the feature store component of MLOps was not studied for ML in official statistics and was identified as a research gap for future studies. In leveraging FS and MLOps within the domain of official statistics, a systematic approach to feature engineering and management is essential. This begins with identifying relevant features crucial for statistical ML solutions, which could range from demographic data to economic indicators. Establishing a centralized repository with FS allows for efficient storage and management of these features, supporting online and offline storage to accommodate varying latency requirements. Versioning and lineage tracking mechanisms within the FS ensure a clear record of feature evolution and usage, while automation capabilities streamline the feature pipeline by automating extraction, transformation, and ingestion processes.

Furthermore, effective model lifecycle management through MLOps is imperative for successful implementation when the production becomes continuous. Adopting an iterative-incremental approach to model development enables continuous training and evaluation using the latest data. Implementing CI/CD pipelines automates model deployment, ensuring seamless integration of updated models into production environments. Version control systems track changes to models, ensuring reproducibility and facilitating rollback in case of issues. Robust monitoring and alerting systems track model performance metrics, allowing for proactive identification and resolution of issues while establishing a feedback loop between model performance monitoring and feature engineering, closing the loop between data ingestion, feature engineering, model training, and deployment.

Finally, addressing the QF4FA dimensions is paramount to ensure the reliability and validity of ML solutions. Ensuring accuracy through continuous evaluation and validation of model performance against ground truth data is crucial. Employing techniques for model explainability enhances the transparency of model predictions, especially in domains where interpretability is critical. Maintaining reproducibility by versioning both models and features within the FS facilitates collaboration and result reproduction. Optimizing model training and deployment pipelines minimizes latency, ensuring timely delivery of insights and predictions, while implementing cost-effective solutions optimizes resource utilization, thus adhering to quality frameworks such as QF4FA and enhancing the overall effectiveness of ML applications in official statistics.

We aim to make a positive contribution by initiating a productive discussion about operational aspects that could create limitations to the broad use of ML in Official Statistics.

References

[1] A. Ashofteh and J. M. Bravo, “Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems,” Stat. J. IAOS, vol. 37, no. 3, pp. 771–789, 2021, doi: 10.3233/SJI-210841.

[2] E. Jahani, P. Sundsøy, J. Bjelland, L. Bengtsson, A. S. Pentland, and Y.-A. de Montjoye, “Improving official statistics in emerging markets using machine learning and mobile phone data,” EPJ Data Sci., vol. 6, no. 1, 2017, doi: 10.1140/epjds/s13688-017-0099-3.

[3] S. Sobolevsky, E. Massaro, I. Bojic, J. M. Arias, and C. Ratti, “Predicting regional economic indices using big data of individual bank card transactions,” in Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, T. M. Nie J.-Y. Obradovic Z. ,. Suzumura T. ,. Ghosh R. ,. Nambiar R. ,. Wang C. ,. Zang H. ,. Baeza-Yates R. ,. Baeza-

Yates R. ,. Hu X. ,. Kepner J. ,. Cuzzocrea A. ,. Tang J., Ed., Institute of Electrical and Electronics Engineers Inc., 2017, pp. 1313–1318. doi: 10.1109/BigData.2017.8258061.

[4] D. Sculley et al., “Hidden technical debt in machine learning systems,” presented at the Advances in Neural Information Processing Systems, 2015, pp. 2503–2511.

[5] D. Kreuzberger, N. Kühl, and S. Hirschl, “Machine Learning Operations (MLOps): Overview, Definition, and Architecture.” arXiv, May 14, 2022. doi: 10.48550/arXiv.2205.02302.

[6] L. Orr, A. Sanyal, X. Ling, K. Goel, and M. Leszczynski, “Managing ml pipelines: Feature stores and the coming wave of embedding ecosystems,” Proc. VLDB Endow., vol. 14, no. 12, pp. 3178–3181, 2021, doi: 10.14778/3476311.3476402.

[7] T. Mboweni, T. Masombuka, and C. Dongmo, “A Systematic Review of Machine Learning DevOps,” in 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), Jul. 2022, pp. 1–6. doi: 10.1109/ICECET55527.2022.9872968.

[8] “Machine Learning for Official Statistics | UNECE.” Accessed: Nov. 15, 2022. [Online]. Available: https://unece.org/statistics/publications/machine-learning-official-statistics

[9] C. Nunes and A. Ashofteh, “A Review of Big Data and Machine Learning Operations in Official Statistics: MLOps and Feature Store Adoption.” https://data.mendeley.com/, Nov. 10, 2023. doi: 10.17632/96mxz7jvkr.1.

[10] A. Ashofteh and J. M. Bravo, “A study on the quality of novel coronavirus (COVID-19) official datasets,” Stat. J. IAOS, vol. 36, no. 2, pp. 291–301, 2020, doi: 10.3233/SJI-200674.

[11] N. Ghaffarzadegan and H. Rahmandad, “Simulation-based estimation of the early spread of COVID-19 in Iran: actual versus confirmed cases,” Syst. Dyn. Rev., vol. 36, no. 1, pp. 101–129, 2020, doi: 10.1002/sdr.1655.

[12] Q. Wang, S. Li, M. Zhang, and R. Li, “Impact of COVID-19 pandemic on oil consumption in the United States: A new estimation approach,” Energy, vol. 239, 2022, doi: 10.1016/j.energy.2021.122280.

[13] N. Levin, S. Ali, and D. Crandall, “Utilizing remote sensing and big data to quantify conflict intensity: The Arab Spring as a case study,” Appl. Geogr., vol. 94, pp. 1–17, 2018, doi: 10.1016/j.apgeog.2018.03.001.

[14] M. Fatehkia, R. Kashyap, and I. Weber, “Using Facebook ad data to track the global digital gender gap,” World Dev., vol. 107, pp. 189–209, 2018, doi: 10.1016/j.worlddev.2018.03.007.

[15] E. Saluveer et al., “Methodological framework for producing national tourism statistics from mobile positioning data,” ANNALS OF TOURISM RESEARCH, vol. 81. PERGAMON-ELSEVIER SCIENCE LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, ENGLAND, Mar. 2020. doi: 10.1016/j.annals.2020.102895.

[16] M. Vanhoof, F. Reis, T. Ploetz, and Z. Smoreda, “Assessing the quality of home detection from mobile phone data for official statistics,” J. Off. Stat., vol. 34, no. 4, pp. 935–960, 2018, doi: 10.2478/jos-2018-0046.

[17] M. L. Loureiro and M. Allo, “Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the UK and Spain,” ENERGY POLICY, vol. 143. ELSEVIER SCI LTD, THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND, Aug. 2020. doi: 10.1016/j.enpol.2020.111490.

[18] C. Zhang, L. Di, L. Lin, and L. Guo, “Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps,” Comput. Electron. Agric., vol. 166, 2019, doi: 10.1016/j.compag.2019.104989.

[19] F. Batista e Silva et al., “Uncovering temporal changes in Europe’s population density patterns using a data fusion approach,” NATURE COMMUNICATIONS, vol. 11, no. 1. NATURE PORTFOLIO, HEIDELBERGER PLATZ 3, BERLIN, 14197, GERMANY, Sep. 15, 2020. doi: 10.1038/s41467-020-18344-5.

[20] D. Hopp, “Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (LSTM),” JOURNAL OF OFFICIAL STATISTICS, vol. 38, no. 3. SCIENDO, BOGUMILA ZUGA 32A, WARSAW, MAZOVIA, POLAND, pp. 847–873, Sep. 01, 2022. doi: 10.2478/jos-2022-0037.

[21] F. Corona, G. Gonzalez-Farias, and J. Lopez-Perez, “Timely Estimates of the Monthly Mexican Economic Activity,” JOURNAL OF OFFICIAL STATISTICS, vol. 38, no. 3. SCIENDO, BOGUMILA ZUGA 32A, WARSAW, MAZOVIA, POLAND, pp. 733–765, Sep. 01, 2022. doi: 10.2478/jos-2022-0033.

[22] P. Krennmair and T. Schmid, “Flexible domain prediction using mixed effects random forests,” J. R. Stat. Soc. Ser. C Appl. Stat., vol. 71, no. 5, pp. 1865–1894, 2022, doi: 10.1111/rssc.12600.

[23] K. G. Abraham, “Big Data and Official Statistics†,” Rev. Income Wealth, vol. 68, no. 4, pp. 835–861, 2022, doi: 10.1111/roiw.12617.

[24] Y. Tillé et al., “Some Thoughts on Official Statistics and its Future (with discussion),” J. Off. Stat., vol. 38, no. 2, pp. 557–598, 2022, doi: 10.2478/jos-2022-0026.

[25] Q. A. Meertens, C. G. H. Diks, H. J. Van Den Herik, and F. W. Takes, “Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms,” J. Off. Stat., vol. 38, no. 2, pp. 485–508, 2022, doi: 10.2478/jos-2022-0023.

[26] A. Van Delden, S. Scholtus, J. Burger, and Q. Meertens, “Accuracy of Estimated Ratios as Affected by Dynamic Classification Errors,” JOURNAL OF SURVEY STATISTICS AND METHODOLOGY. OXFORD UNIV PRESS INC, JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, 2022. doi: 10.1093/jssam/smac015.

[27] R. Perepi, K. Santhi, R. Saraswathi, and O. A. Bég, “Predictive analysis of COVID 19 disease based on mathematical modelling and machine learning techniques,” Comput. Methods Biomech. Biomed. Eng. Imaging Vis., 2022, doi: 10.1080/21681163.2022.2120829.

[28] V. Heikinheimo, O. Järv, H. Tenkanen, T. Hiippala, and T. Toivonen, “Detecting country of residence from social media data: a comparison of methods,” Int. J. Geogr. Inf. Sci., vol. 36, no. 10, pp. 1931–1952, 2022, doi: 10.1080/13658816.2022.2044484.

[29] A. Bandam, E. Busari, C. Syranidou, J. Linssen, and D. Stolten, “Classification of Building Types in Germany: A Data-Driven Modeling Approach,” Data, vol. 7, no. 4, 2022, doi: 10.3390/data7040045.

[30] A. A. Ormenisan, M. Meister, F. Buso, R. Andersson, S. Haridi, and J. Dowling, “Time travel and provenance for machine learning pipelines,” in OpML 2020 - 2020 USENIX Conference on Operational Machine Learning, USENIX Association, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091923536&partnerID=40&md5=094a2347bdc88ce4eb24f9b4168a7242

[31] J. Patel, “Unification of Machine Learning Features,” in Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020, A. S. I. Chan W.K. Claycomb B. ,. Takakura H. ,. Yang J. J. ,. Teranishi Y. ,. Towey D. ,. Segura S. ,. Shahriar H. ,. Reisman S., Ed., Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1201–1205. doi: 10.1109/COMPSAC48688.2020.00-93.

[32] J. Patel, “The Democratization of Machine Learning Features,” in Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 136–141. doi: 10.1109/IRI49571.2020.00027.

[33] B. A. Quon and J.-L. Gaudiot, “Concept drift detection for distributed multi-model machine learning systems,” in Proceedings - 2022 IEEE 46th Annual Computers, Software, and Applications Conference, COMPSAC 2022, S. H. Va Leong H. Sarvestani S. S. ,. Teranishi Y. ,. Cuzzocrea A. ,. Kashiwazaki H. ,. Towey D. ,. Yang J. J., Ed., Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1076–1080. doi: 10.1109/COMPSAC54236.2022.00168.

Appendix – Graphical Summary of Search Process