Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cjas20

Editorial to the special issue: modern streaming data analytics

Yajun Mei(Principal Guest Editor), Jay Bartroff (Co-Editor), Jie Chen (Co-Editor), Georgios Fellouris (Co-Editor) & Ruizhi Zhang (Co-Editor)

To cite this article: Yajun Mei(Principal Guest Editor), Jay Bartroff (Co-Editor), Jie Chen (Co-Editor), Georgios Fellouris (Co-Editor) & Ruizhi Zhang (Co-Editor) (2023) Editorial to the special issue: modern streaming data analytics, Journal of Applied Statistics, 50:14, 2857-2861, DOI: 10.1080/02664763.2023.2247646

To link to this article: https://doi.org/10.1080/02664763.2023.2247646

JOURNAL OF APPLIED STATISTICS
2023, VOL. 50, NO. 14, 2857–2861
https://doi.org/10.1080/02664763.2023.2247646

Taylor & Francis
Taylor & Francis Group

EDITORIAL

Editorial to the special issue: modern streaming data analytics

The special issue 'Modern Streaming Data Analytics' of the *Journal of Applied Statistics* (JAS), Taylor & Francis, features papers primarily presented in the 2021 ICSA Applied Statistics Symposium, held virtually from September 12 to September 15, 2021. Streaming data analytics has been an active topic and its interesting new applications have lately been seen in many fields such as biomedical sciences, epidemiology, engineering, finance and economics, network and data security, etc. With the support of the symposium's organizing committee, we organized five invited sessions, uniting twenty prominent researchers in the field. These sessions fostered in-depth discussions on cutting-edge ideas and the most pressing challenges surrounding streaming data analytics and its applications. Encouraged by Professor Jie Chen, the editor of JAS, and several distinguished speakers, the decision was made to curate a dedicated special issue on this compelling topic, soliciting additional paper submissions. Throughout the submission process, we received a total of twenty-four manuscripts. Our rigorous review process, carried out diligently by both referees and guest editors, culminated in the acceptance of eight outstanding papers. Despite the challenges posed by the COVID-19 pandemic, leading to delays and a smaller number of submissions, the selected papers exemplify high quality, offering readers the opportunity to access different modern topics and application areas of streaming data analytics. By providing a platform for the dissemination of valuable research and ideas, we hope this special issue will spark enthusiasm and curiosity among researchers and practitioners in developing more streaming data analytics, and will inspire future breakthroughs, contributing to the continuous growth and enrichment of streaming data analytics as indispensable tools across numerous disciplines.

Change-point problems, where the objective is to detect changes in the underlying data distributions over time, play an important role in streaming or sequential data analytics. When the data distribution permanently shifts to a new state, the change is "persistent", whereas when the data distribution temporarily deviates from the original state but eventually returns to it, the change is "transient". While there has been considerable research on the detection of persistent changes, investigations into transient changes have been relatively limited. Addressing this gap, Baron and Malov [1] made a significant contribution to the study of transient changes. Without assuming any a priori information, they developed very useful statistical methodologies for the detection of such changes and the estimation of the corresponding transient intervals. Specifically, given a fixed dataset, they first addressed the problem of testing for a transient change, where they also derived the asymptotic distribution of the MLE of the corresponding transient interval. Then, they explored the detection and estimation of multiple transient changes: first, in the case of a known number of transient changes and, then, in the even more challenging case of a completely unknown number of transient changes. The proposed statistical methods ensure simultaneous control of the familywise false alarm rate and the familywise false re-adjustment

rate. Moreover, they are illustrated with a number of insightful simulation studies. Overall, Baron and Malov's methodological and theoretical contributions in this work enhance the understanding of data that are characterized by temporary changes in distribution. Thus, they are particularly valuable in fields where such transient changes are common, such as finance, industrial process control, and environmental monitoring.

A subtle yet pivotal advancement in the subfield of hot-spot detection and localization involves actively sampling incomplete data under sampling control, as opposed to relying on complete raw data. A compelling example of this scenario is monitoring the number of COVID-19 confirmed cases across various spatial locations over time. In practice, testing each and every individual regularly to determine their infection status is impractical due to limited testing resources and individual preferences, such as not getting tested or failing to report self-testing results at home. As a result, statistical inferences must be drawn from incomplete data, subject to the constraints of sampling control. In a notable contribution, Hu *et al.* [2] proposed a methodology to address this challenge. They approached the problem by modeling count data as Binomial distributions, and then they detected hot-spots by combining the upper confidence bound algorithm, commonly used in the context of multi-armed bandit problems, with the Cumulative Sum (CUSUM) statistics, typically utilized in sequential change-point detection. This advancement opens up new research avenues for actively sampling or learning from spatio-temporal streaming data, with potential applications beyond the domain of COVID-19 monitoring. For instance, this methodology may prove valuable in efficiently monitoring and detecting crimes or accidents while adhering to staff or manpower constraints.

In many real-world scenarios, specifying the nonlinear mean structures of spatio-temporal data can be challenging and impractical. To address this issue, Lue and Tzeng [3] have developed a data-driven approach that enables the estimation of mean structures through effective dimension reduction. Their methodology relies on the use of empirical orthogonal function (EOF) analysis, which is a low-rank method based on principal component analysis. In the context of spatio-temporal data, the proposed approach considers the linear combination of inner products of temporal and spatial functions. By incorporating pairwise directions estimation with kriging, the proposed approach is able to explain how temperature cycles vary with location and/or with time when modeling spatio-temporal data. By providing a data-driven and interpretable approach to spatio-temporal data analysis, Lue and Tzeng's work opens up new possibilities for uncovering hidden patterns and trends, ultimately leading to a deeper understanding of the dynamic interplay between spatial and temporal factors in diverse real-world scenarios.

Measuring association or correlation between two time series is a crucial aspect of spatio-temporal data analysis. However, the presence of autocorrelation in time series poses challenges for applying classical approaches designed for independent and identically distributed (i.i.d.) data without modifications. In a significant contribution, Lun *et al.* [4] addressed this issue by developing the asymptotic distribution of nonparametric measures such as Spearman's Rho and Kendall's Tau, specially tailored for correlations between two autocorrelated time series exhibiting short-range dependence. The applicability of Lun et al.'s results extends to a wide array of time series models, including popular ones such as ARMA (AutoRegressive Moving Average), GARCH (Generalized Autoregressive Conditional Heteroskedasticity), and certain copula-based models. Their methodological

developments enable researchers and practitioners to account for autocorrelation while accurately quantifying the correlation between two time series. The authors demonstrated the efficacy of their proposed methods by analyzing observed climatological time series data, specifically flood discharges and temperature data in Europe. As spatio-temporal data analysis becomes increasingly vital across various domains, the methodology presented by Lun et al. equips researchers with the tools to draw more accurate and meaningful conclusions from their data, fostering advancements in fields such as climatology, hydrology, economics, and more.

Active feature selection in streaming data analytics is a fascinating topic that aids in making reliable and cost-effective decisions. From a machine learning or statistical perspective, the goal is to identify informative features from historical or training data that lead to high prediction accuracy, and the conventional approach is to collect as many informative features as possible to enhance model performance. However, in certain real-world scenarios, such as clinical or medical studies, this approach may be impractical when timely decisions are crucial. Indeed, in the medical domain, doctors often need to make quick decisions to diagnose and treat patients effectively. They may begin with simpler and more convenient laboratory tests to gain initial insights into a patient's condition. If necessary, they may proceed to collect more informative yet expensive features to refine their diagnosis and treatment plan. Tian *et al.* [5] proposed a multistage sequential decision-making model designed to support the medical doctors in their decision-making process. The key idea is to actively collect necessary information from each subject in a sequential manner. This dynamic approach allows doctors to iteratively assess the patient's conditions, adding more informative features as needed, thereby making well-informed decisions in a timely manner. The proposed model and method is applied to a case study of common bile duct (ABD) stone evaluation for pediatric patients. This research presents a promising direction for actively selecting features in streaming data analytics, catering to the specific needs of medical professionals and supporting them in facilitating informed decision-making that is quick, reliable, and cost-effective.

The problem of identifying or localizing affected local streams after quick detection is a long standing problem in the high-dimensional sequential change-point detection or statistical process control. Tsang *et al.* [6] have proposed an innovative approach to this problem by applying modern Knockoff filtering from the false discovery rate (FDR) literature. They showed that their proposed approach allows for the identification of out-of-control (OCC) local streams while controlling the FDR when monitoring high-dimensional data streams. It is important to highlight the challenges of sequential statistical procedures: even if one controls the FDR at each time point, this does not mean that the FDR is controlled at the stopping time when signals are detected! Moreover, the tradeoff between using as few post-change observations as possible for quickest rapid detection and employing more post-change observations for enhanced localization accuracy presents an intriguing and challenging research direction. Striking the right balance between these two objectives is essential for optimizing the performance of the change-point detection and localization method in streaming data. The research carried out by Tsang, Tsung, and Xu has practical implications across numerous domains where timely detection and precise localization of changes are crucial for effective decision-making and process improvement.

The application of monitoring real-time imaging data in high-throughput plant phenotyping (HTPP) as presented in Zhan *et al.* [7] is both fascinating and practical. HTPP has emerged as a valuable technique for studying plant traits due to its numerous advantages including speed, accuracy, non-destructiveness, and labor-saving properties. It finds wide applications in plant breeding and crop management. However, the implementation of HTPP comes with a challenge: the generation of massive image data, which can be overwhelming to analyze efficiently and accurately. To tackle this issue, Zhan et al. proposed a two-step image-based online detection framework for monitoring the individual plant leaf area via real-time imaging data. First, they developed an efficient supervised learning algorithm to extract plant leaf area from the multiview RGB images. Then, they proposed to use the adaptive CUSUM procedure to monitor the standardized relative change of the predicted plant leaf area for the quickest change detection. The authors illustrated the detection efficacy of their proposed framework by real data analysis of multiview RGB images of soybean plants, which were collected at the Nebraska Innovation Campus Greenhouse, High-Throughput Plant Phenotyping Core Facilities. The research by Zhan et al. showcases the potential of real-time imaging data monitoring in HTPP and its relevance in advancing agricultural research and practices. By addressing the challenges posed by massive image data, this framework opens up new avenues for efficient and informed decision-making in the field of plant phenotyping and crop management.

Streaming count data are prevalent in biosurveillance and healthcare applications, where one tracks the occurrence of new patients with various types of infectious diseases across different cities, counties, or states repeatedly over time. Detecting and localizing hot-spots, characterized by unusually high infectious rates, is a crucial task to enable timely and appropriate responses. In their pioneering work, Zhao *et al.* [8] addressed this challenge by leveraging Poisson distributions and tensors to model count data. They introduced a novel method called Poisson assisted Smooth Sparse Tensor Decomposition (PoSSTenD) specially tailored for the spatio-temporal count data. The strength of PoSSTEND lies in its ability not only to detect the occurrence of hot-spots but also to pinpoint their locations. The proposed method is applied to a real dataset that includes the annual number of ten different infectious diseases from 1993 to 2018 for 49 mainland states in the United States. The results showcased the method's potential to uncover and localize hot-spots, shedding light on regions with unusually high disease incidence rates over time.

Acknowledgments

We would like to thank the authors for their excellent contributions as well as the dedicated peer reviewers for their careful and constructive comments which greatly improved the quality of the papers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Y.M. is supported in part by NSF grant DMS-2015405. R.Z. is supported in part by NSF grant ECCS-2236565.

References

[1] M. Baron and S.V. Malov, *Detection and estimation of multiple transient changes*, J. Appl. Stat. 50 (2023), pp. 2862–2888.
[2] J. Hu, Y. Mei, S. Holte, and H. Yan, *Adaptive resources allocation CUSUM for binomial count data monitoring with application to COVID-19 hotspot detection*, J. Appl. Stat. 50 (2023), pp. 2889–2913.
[3] H.-H. Lue and S. Tzeng, *Interpretable, predictive spatio-temporal models via enhanced pairwise directions estimation*, J. Appl. Stat. 50 (2023), pp. 2914–2933.
[4] D. Lun, S. Fischer, A. Viglione, and G. Blöschl, *Significance testing of rank cross-correlations between autocorrelated time series with short-range dependence*, J. Appl. Stat. 50 (2023), pp. 2934–2950.
[5] H. Tian, R.Z. Cohen, C. Zhang, and Y. Mei, *Active learning-based multistage sequential decision-making model with application on common bile duct stone evaluation*, J. Appl. Stat. 50 (2023), pp. 2951–2969.
[6] K.W. Tsang, F. Tsung, and Z. Xu, *Knockoff procedure for false discovery rate control in high-dimensional data streams*, J. Appl. Stat. 50 (2023), pp. 2970–2983.
[7] Y. Zhan, R. Zhang, Y. Zhou, V. Stoerger, J. Hiller, T. Awada, and Y. Ge, *Rapid online plant leaf area change detection with high-throughput plant image data*, J. Appl. Stat. 50 (2023), pp. 2984–2998.
[8] Y. Zhao, X. Huo, and Y. Mei, *Hot-spots detection in count data by Poisson assisted smooth sparse tensor decomposition*, J. Appl. Stat. 50 (2023), pp. 2999–3029.

Yajun Mei
Principal Guest Editor
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA, USA
ymei@isye.gatech.edu
http://orcid.org/0000-0002-1015-990X

Jay Bartroff
Co-Editor
Department of Statistics and Data Sciences,
University of Texas at Austin, Austin, TX, USA

Jie Chen
Co-Editor
Division of Biostatistics and Data Science,
Department of Population Health Science,
Augusta University, Augusta, GA, USA

Georgios Fellouris
Co-Editor
Department of Statistics
University of Illinois at Urbana-Champaign,
Champaign, IL, USA

Ruizhi Zhang
Co-Editor
Department of Statistics,
University of Georgia, Athens, GA, USA