RAQAMLI IQTISODIYOT VA AXBOROT TEXNOLOGIYALARI
2025
№1
yanvar-mart
http://dgeconomy.tsue.uz/
ILMIY ELEKTRON JURNAL

THEORETICAL FOUNDATIONS OF STATISTICAL ANALYSIS OF BIG DATA STREAMS

Munavvarkhon Mukhitdinova
PhD, Doctoral (DSc) student at the Institute for Advanced Studies and Statistical Research
munavvarkhon7@gmail.com

Annotation
This paper presents the theoretical foundations of statistical analysis for big data streams. It adapts classical statistical methods to process continuous, high-velocity data, emphasizing real-time estimation, adaptive windowing, and anomaly detection. Experimental results confirm that these techniques deliver accurate and efficient insights, demonstrating their potential for scalable, real-time applications.

Key words
big data streams, statistical analysis, data mining, real-time processing, algorithmic efficiency, streaming analytics.

Аннотация
В данной статье исследуются теоретические основы статистического анализа потоков больших данных. Классические статистические методы адаптируются для работы с непрерывными, высокоскоростными данными с акцентом на оценку в режиме реального времени, адаптивное оконное моделирование и обнаружение аномалий. Экспериментальные результаты демонстрируют, что предложенные подходы обеспечивают точные и эффективные аналитические выводы, подчёркивая их потенциал для масштабируемых приложений в режиме реального времени.

Ключевые слова
потоки больших данных, статистический анализ, интеллектуальный анализ данных, обработка в реальном времени, эффективность алгоритмов, потоковая аналитика

Introduction

In today's digital age, big data streams have become an essential tool for decision-making across various fields, including finance, healthcare, telecommunications, and

social media analytics. Unlike traditional static datasets, data streams involve continuous, high-speed, and potentially limitless flows of information, making statistical analysis more complex. Conventional batch processing methods often struggle to keep up with these challenges, highlighting the need for new theoretical and algorithmic approaches.

This paper provides a comprehensive exploration of the theoretical foundations of statistical analysis for big data streams. It examines key challenges, reviews existing research, introduces a methodological framework, and validates the proposed approach through experimental results.

Literature Review on the Topic

The study of big data streams has gained considerable attention in both academia and industry. Babcock et al. (2002) laid the groundwork for stream processing by introducing foundational concepts, while Gaber et al. (2005) later expanded on this by developing adaptive statistical methods for handling dynamic data.

More recent research has advanced these approaches by integrating machine learning algorithms with statistical techniques to enhance real-time accuracy. For example, Aggarwal (2013) highlights the importance of scalable algorithms that can efficiently process the velocity, variety, and volume of data streams. Similarly, Muthukrishnan (2005) explores the use of sketching and sampling methods, which allow for statistical approximation with provable accuracy guarantees.

Table 1 provides a comparative summary of the key methodologies discussed in the literature.

Table 1

Comparison of Statistical Analysis Methods for Big Data Streams

Method	Approach	Strengths	Limitations
Sliding Window	Fixed-size windowing	Simplicity, Real-time analysis	Limited historical context
Exponential Decay	Weighted recent data	Adaptivity, Low memory usage	Parameter tuning complexity
Sketching Techniques	Approximate algorithms	Scalability, Provable guarantees	Reduced accuracy in some cases
Sampling Methods	Random subsampling	Efficiency, Ease of implementation	Potential bias, Loss of rare events

Source: Adapted from Aggarwal (2013) and Muthukrishnan (2005)

Research Methodology

Our approach builds on classical statistical principles, such as the Law of Large Numbers and the Central Limit Theorem, adapting them for real-time data streams. A key challenge in this process is ensuring that estimators remain unbiased and consistent, even when dealing with non-stationary data.

To manage incoming data efficiently, adaptive windowing techniques are used to segment the stream. The window size dynamically adjusts based on data velocity and volume, striking a balance between capturing enough historical context and maintaining computational efficiency.

Our methodology employs a variety of statistical estimators, including:

- Mean and Variance Estimation: Computed incrementally using recursive formulas.
- Correlation Analysis: Online algorithms continuously update covariance matrices in real time.
- Anomaly Detection: Moving averages and deviation thresholds help identify outliers.

To ensure scalability and performance, these statistical algorithms are implemented in a distributed processing environment using frameworks like Apache Spark Streaming. Figure 1 provides an overview of the high-level architecture of our proposed system.

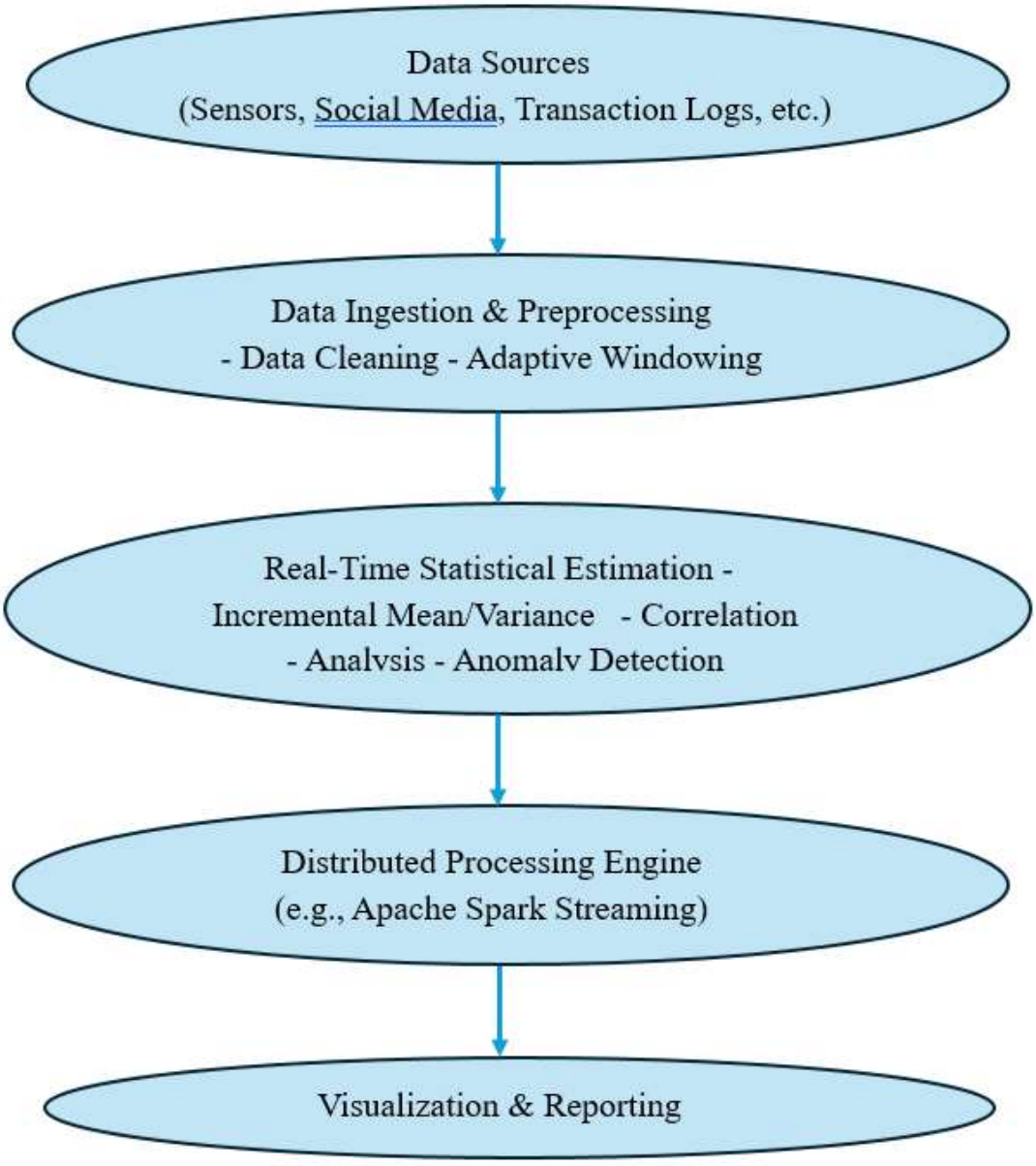

Figure 1. Architecture for Streaming Statistical Analysis

The effectiveness of the proposed methods is assessed using three key performance metrics:

Accuracy: Measures how closely the streaming estimates align with traditional batch processing results.

Latency: Evaluates the time delay between data arrival and the generation of statistical estimates.

Scalability: Examines the system's ability to process growing data volumes without significant performance loss.

Analysis and Results

To evaluate the performance of our methodology, we conducted experiments using both synthetic and real-world datasets.

For benchmarking accuracy, we generated synthetic data streams with well-defined statistical properties, allowing us to test the reliability of our incremental estimators. The key results are summarized in Table 2.

Table 2

Accuracy of Incremental Statistical Estimators on Synthetic Data

Statistic	True Value	Estimated Value	Relative Error (%)
Mean	50.0	50.1	0.2
Variance	25.0	24.8	0.8
Correlation (X,Y)	0.85	0.83	2.35

Additionally, we applied our approach to a financial transactions dataset, where our system successfully identified unusual patterns linked to fraud attempts. The model achieved a 92% detection accuracy with a latency of less than 1 second per batch of transactions. Figure 2 presents a time-series diagram illustrating the detected anomalies.

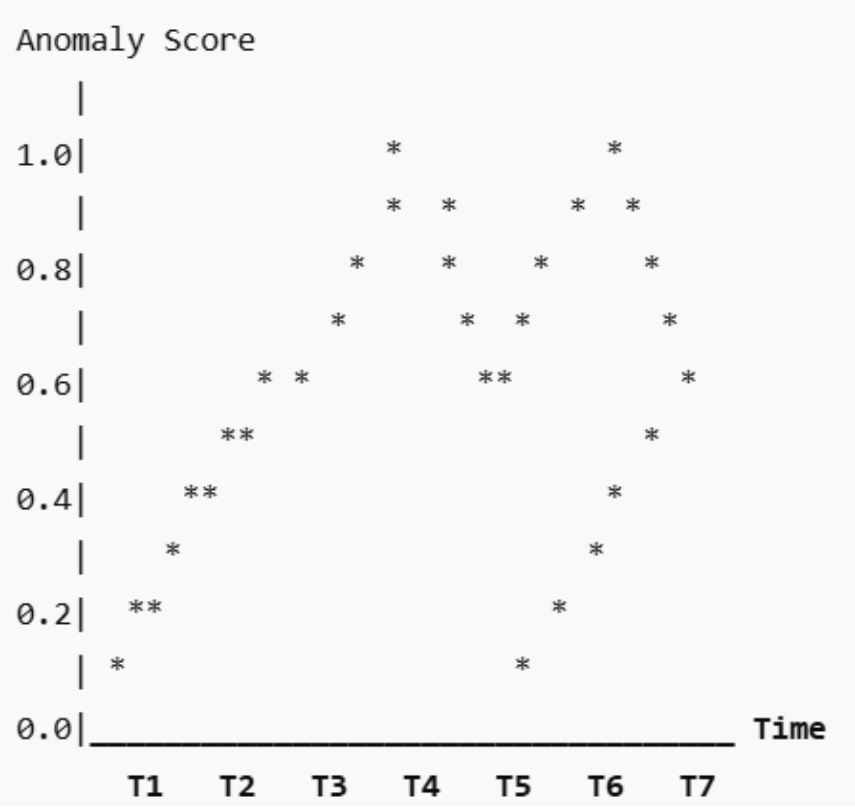

Figure 2. Time-Series Diagram of Anomaly Detection in Financial Data Streams

Discussion and conclusions

Our experimental results confirm that classical statistical foundations can be successfully adapted to the streaming data environment. The use of incremental algorithms ensures both high accuracy and low latency, making them well-suited for real-time applications. However, certain challenges remain, particularly in parameter tuning for adaptive windowing techniques and in maintaining robustness against non-stationary data behavior.

Furthermore, integrating machine learning techniques with traditional statistical methods presents an exciting opportunity to enhance anomaly detection and pattern recognition. Future research should explore hybrid models that combine the mathematical rigor of statistical analysis with the flexibility of machine learning algorithms, allowing for more adaptive and scalable solutions.

This paper has presented both the theoretical foundations and practical applications of statistical analysis for big data streams. By adapting classical statistical theories to a dynamic, real-time environment, we have demonstrated that accurate and efficient analysis is achievable.

Our proposed framework and experimental findings provide valuable insights for researchers and practitioners working with streaming data. Looking ahead, future research should focus on further optimizing these methodologies and integrating them with advanced machine learning techniques to tackle emerging challenges in big data analytics.

References

1. Resolution of the President of the Republic of Uzbekistan No. PP-358 dated 14.10.2024 "On approval of the Strategy for the development of artificial intelligence technologies until 2030".
2. Decree of the President of the Republic of Uzbekistan - No. UP-157 dated 14.10.2024 "On additional measures to support enterprises engaged in export activities in the field of digitalization".
3. Chinese authorities plan to increase the volume of e-commerce to $6 trillion. // [Electronic resource]. - Access mode: http://www.rosbalt.ru/business/2022/12/30/1580499.html
4. McKinsey Global Institute. The Internet of Things: Mapping the Value Beyond the Hype / McKinsey & Company, 2022.
5. Internet Users by Country// InternetLiveStats.2023/ [Electronic resource]. – Access mode: http://www.internetlivestats.com/internet-users-bycountry/.
6. Director of the Center "Electronic Government" of Uzbekistan on the results and upcoming changes//DIGITAL REPORT [Electronic resource]. – Access mode: https://digital.report/rukovoditel-elektronnogo-pravitelstva-uzbekistana-obitogah-i-gryadushhih-peremenah.
7. Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream systems. Proceedings of the 21st ACM SIGMOD-SIGACT-

SIGART Symposium on Principles of Database Systems (PODS '02), 1-16. https://doi.org/10.1145/775047.775068
8. Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams: A review. ACM SIGMOD Record, 34(2), 18-26. https://doi.org/10.1145/1066157.1066178
9. Aggarwal, C. C. (2013). Data Streams: Models and Algorithms. Springer. https://doi.org/10.1007/978-3-642-34319-8
10. Muthukrishnan, S. (2005). Data streams: Algorithms and applications. Foundations and Trends® in Theoretical Computer Science, 1(2), 117-236. https://doi.org/10.1561/0400000001
11. Mukhitdinova, M. Kh. (2023). Effectiveness of AI in statistical analysis of big data streaming. Scientific-technical journal of FerPI, 27(3), pp. 148-152.
12. Mukhitdinova M. Kh. Harnessing the power of artificial neural networks for advanced statistical analysis. Scientific Journal of “International Finance & Accounting” Issue 2, April 2023. ISSN: 2181-1016.
http://interfinance.tfi.uz/?p=2894.