Fuzzy Systems and Data Mining X
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA241435

Correlation Analysis over Big Multidimensional Datasets: A Powerful Paradigm for Next-Generation Big Data Analytics Research−Definitions, Models, Implementations

Alfredo CUZZOCREA[a,b, 1]
[a] *iDEA Lab, University of Calabria, Rende, Italy*
[b] *Department of Computer Science, University of Paris City, Paris, France*
ORCiD ID: Alfredo Cuzzocrea https://orcid.org/0000-0002-7104-6415

Abstract. *Correlation analysis* has been a powerful paradigm to discover and analyze *hidden properties* and *patterns* of large-scale datasets for decades. At now, correlation analysis turns to be a perfect tool for supporting *big multidimensional data* analysis and mining, with a wide range of relevant properties, including the amenity of supporting meaningfully exploration and discovery of multidimensional ranges kept in such kind of datasets. These operators are thus the basis for several *multidimensional big data analytical tools* that can be designed and implemented on top of the foundations defined by correlation functions. In line this this scientific area, the talk will provide introduction and motivations, models and algorithms, and, finally, best-practices guidelines for effective and efficient implementations of correlation-analysis-based tools over big multidimensional datasets.

Keywords. Big data, Big data analytics, Correlation analysis, Advanced big data analytics

1. Introduction

Nowadays, the emergence of *big data* has revolutionized various sectors, ranging from scientific research to *Business Intelligence* (BI), by offering unprecedented opportunities for discovering hidden insights from vast and complex datasets. In particular, the ability to find meaningful patterns between variables has become a critical driver of progress in fields such as *healthcare*, *finance*, *environmental science*, and so forth (e.g., [1,2]). One of the most prominent analytical tools to achieve this is *correlation analysis*. Historically, correlation analysis has served as an essential method for understanding relationships between variables by quantifying the degree to which they move together or apart. Over time this has been found to be effective when it comes to various data-oriented issues, right from simple bilinear relationships to several variable complications. In this context,

[1] Corresponding Author. E-mail: alfredo.cuzzocrea@unical.it – This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France

as data volume, variety and velocity grow exponentially, the importance of correlation analysis amplifies, particularly when applied to *large-scale multidimensional datasets* (e.g., [3,4]).

The shift from traditional data analysis to big data analytics has increased the scope and volume of datasets while simultaneously posing fresh challenges in terms of complexity, dimensionality, and computing capability. Multidimensional data is common in a variety of disciplines, including sensor networks (e.g., [5]), geographical information systems (e.g., [6]), genomics (e.g., [7]), and financial markets (e.g., [8]), where datasets contain a large number of related variables that must be examined concurrently. While correlation analysis has long been used to uncover linear correlations between variables (e.g., [9]), applying it to multidimensional datasets necessitates dealing with a very complex data field in which patterns are difficult to detect. Despite its complexities, correlation analysis is particularly well-suited to big data analytics because it may be used as a strong exploratory technique to uncover previously unknown important patterns, anomalies, and connections (e.g., [10]).

Indeed, correlation analysis enables analysts to examine not just pairwise correlations but higher-order relationships that span multiple dimensions (e.g., [11]). These multidimensional correlations reveal intricate dependencies among variables, which are critical for applications like recommendation systems (e.g., [12,13]), risk management (e.g., [14]), and anomaly detection (e.g., [15,16]). Despite its advantages, the application of correlation analysis to big multidimensional datasets presents several challenges. One of the most prominent difficulties lies in managing the computational complexity inherent in analyzing large datasets with many dimensions. As the number of dimensions increases, so too does the complexity of the correlation functions required to analyze the data. This phenomenon, often referred to as the *curse of dimensionality* (e.g., [17,18]) results in computational bottlenecks, where the resources needed for analysis become prohibitively high. Thus, advanced techniques are required to filter out irrelevant data, mitigate the effects of noise, and ensure that correlation results are both accurate and reliable.

In light of these drawbacks, current research has centered on creating new models, algorithms, and tools for doing efficient and effective correlation analysis on large multidimensional data sets. Several computational methodologies, including dimensionality reduction techniques, parallel processing, and distributed computing, have been developed to overcome the scalability challenge. For instance, techniques like as *Principal Component Analysis* (PCA) (e.g., [19]) and *Single Value Decomposition* (SVD) (e.g., [20]) can minimize the number of dimensions in a dataset while maintaining the most essential variables, simplifying the study without sacrificing vital information. These developments, together with the development of powerful algorithms capable of handling high-dimensional spaces, have prepared the way for next-generation big data analytics that fully use correlation analysis.

Following these considerations, this paper aims to advance the current state of research in this domain by providing a comprehensive examination of correlation analysis applied to big multidimensional datasets. In doing so, it aims to contribute to both the theoretical and practical dimensions of big data analytics. Specifically, this paper will explore the fundamental definitions and models that reinforce correlation analysis in a multidimensional context, by offering insights into the mathematical and statistical foundations of the technique.

2. Related Work

In this Section, we provide a comprehensive overview of key research proposals that are closely aligned with our work on correlation analysis over big multidimensional datasets. We examine a range of studies that have addressed similar challenges in managing and analyzing high-dimensional data, with a particular focus on methods developed to uncover complex correlations across multiple variables.

[21] proposes eBoF, a novel visualization approach based on the *Bag-of-Features* (BoF) model, aimed at analyzing time-varying ensemble data. The eBoF method extracts simple, monotonic intervals from target variables in ensemble scalar data, which serve as local feature patches. Previous works in *ensemble data visualization* have mainly focused on static models or traditional clustering techniques, which often lose *geo-spatial information* and provide less detailed insights into temporal dynamics. Existing methods like topic modeling or clustering algorithms tend to overlook the temporal correlations and geographic features critical for comprehensive data analysis. eBoF addresses these limitations by incorporating *temporal correlation* into *feature clustering*, preserving spatial information that is often lost in conventional methods. The authors build on these limitations by introducing an approach that clusters *feature patches-based* on their temporal similarity, which offers a probability distribution across different clusters. This enables more insightful clustering results, which are validated using domain knowledge. Through case studies and performance evaluations, the authors demonstrate the effectiveness of eBoF in *ensemble simulation data analysis*, by comparing it against traditional methods in order to show its superior ability to retain *spatial-temporal insights*.

In [22], the authors introduce a novel approach leveraging *Canonical Correlation Analysis* (CCA) to explore the intricate relationships between electricity consumption, gas consumption, and climate factors within a *smart grid environment*. Unlike traditional approaches, which predominantly focused on *coarse-grained time* series analyses, which often rely on methods like *Autoregressive Distributed Lag* (ARDL) models to investigate long-term relationships over monthly or yearly intervals, this research addresses the emerging need for *fine-grained*, *high-resolution* data analysis. The increasing deployment of smart grids and *Internet of Things* (IoT) devices has enabled the collection of multidimensional, hourly consumption data, capturing complex interactions between internal factors, such as consumer behavior, and external variables, including climate, location, and building types. Traditional CCA methods, while effective for small-scale datasets, struggle with the computational demands and complexity introduced by these large, dynamic, and multidimensional datasets. This paper significantly extends the application of CCA by proposing an optimized model, called *Canonical Correlation Analysis with an Optimal Result Selection Mechanism* (CCASM), which is specifically designed to handle the volume and granularity of big data. CCASM improves upon classical CCA by offering enhanced computational efficiency and the ability to manage large-scale multivariable datasets, providing a more scalable and accurate means of analyzing fine-grained relationships in energy consumption patterns. By segmenting consumers based on geographic and climatic factors, this model not only allows for a more detailed understanding of consumption behaviors but also enables cross-sectional analysis across various demographic and environmental conditions.

[23] addresses the computational challenges of *Generalized Canonical Correlation Analysis* (GCCA) in handling *large-scale multi-view* data. While classical GCCA has been extensively studied for integrating information across multiple feature spaces, many

existing algorithms face significant memory and computational demands, limiting their scalability. To overcome these issues, the authors propose *scalable* and *distributed* GCCA algorithms that reduce complexity by scaling linearly with the number of data points, while also efficiently handling sparse data. Previous works have introduced formulations like *sum-of-correlations* (SUMCOR) and *maximal variation* (MAX-VAR) for GCCA, but these often relied on *deflation-based methods*, which can lead to error propagation and computational inefficiencies. This paper improves upon those approaches by avoiding bottlenecks such as whitening processes, which traditionally increase computational load. By leveraging distributed computing, the proposed method enables more efficient analysis of large-scale datasets, which significantly enhances the scalability of GCCA for big data applications.

In [24], they delve into the application of correlation analysis within medical science, addressing the critical challenge of managing imprecise or uncertain data. It introduces an *improved Z-test* for correlation, explicitly designed to account for the inherent uncertainty often present in real-life datasets. While previous studies have predominantly relied on classical Z-tests, which assume precise and well-defined data, practical applications in fields like medicine frequently involve data with indeterminate values or interval-based measurements due to factors such as measurement error or variability in biological processes. To overcome these limitations, this paper extends conventional statistical approaches by incorporating *neutrosophic statistics*, a framework developed to handle *uncertainty*, *imprecision*, and *indeterminacy* in data. The improved Z-test is applied to assess the correlation between medical variables, such as *heartbeat* and *temperature*, where uncertainty in measurements is common. By comparing the performance of the proposed method against traditional statistical tests, the authors demonstrate its superiority in scenarios where data precision is uncertain, which highlights its practical significance in medical decision-making and diagnostics. This work not only improves the robustness of correlation analysis in medical contexts but also extends the applicability of statistical methods in fields where data reliability is a challenge.

When working with big, complicated datasets, correlation analysis is a frequently utilized method for exploratory data analysis. Three popular correlation techniques are compared with *Kendall's Rank Correlation, Spearman's Rank Correlation, and Pearson's Product-Moment Correlation*. In [25], they examine the association between the operational condition of industrial pumps and more than 207,000 variables in a *high-dimensional vibration dataset* using Spearman's rank correlation coefficient, a non-parametric measure. Because Spearman's correlation can capture monotonic correlations between variables and is resilient when dealing with *non-normally* distributed data, it was selected as the preferred method. Because of this, it works especially well in situations involving *non-linear* or *distribution-free* data. They employ Spearman's method to pinpoint important factors that have a strong correlation with the states of the pump. This allows for a more focused approach to further analysis, particularly in the development of unsupervised machine learning models. [25] makes a further contribution by using the *R Programming Language* and effective computational techniques to manage the enormous dataset and calculate correlation values. Their findings are important for machine diagnostics and predictive maintenance since the vibration analysis they conducted identifies mechanical problems early on, which may assist avoid expensive machine failures.

In [26], the authors investigate how China's targeted poverty alleviation through education may be addressed through the use of machine learning models and

spatiotemporal correlation analysis. The *Average Education Years* (AEY) and *Gross Domestic Product per Capital* (GDP/C) are the main subjects of their research, which finds a strong positive association between the two. The authors using a combination of remote sensing data and geospatial analysis, offering insights into how enhancing education may fight poverty, examine the distribution of poverty among provinces. In order to identify important factors impacting education levels, such as population distribution and economic development, the study uses *Principal Component Analysis* (PCA). A *Linear and Residual Integration Model* (LRIM) is then constructed. This model uses a *Back Propagation* (BP) neural network to handle nonlinear residuals and linear time series models (ARIMA) to forecast trends in educational progress. Furthermore, using nighttime light data, they categorize impoverished counties using *Random Forest* algorithms, attaining great accuracy in the spatiotemporal analysis of poverty distribution from 1995 to 2010. Their method combines *cutting-edge* big data processing and machine learning techniques to deliver practical insights for educational policy meant to combat poverty. By fusing classical statistical models with spatial analysis, this work advances the discipline and provides a fresh viewpoint on focused education-based initiatives to reduce poverty.

Using data mining techniques and clustering algorithms, the authors investigate the association between middle school student subjective well-being and physical activity in this paper. In order to gauge the amount of physical activity and subjective well-being of 1,848 students from five cities in Sichuan Province, China, the authors performed a study and gathered information via questionnaires. They discovered that student overall levels of physical activity were inadequate, with girls exhibiting noticeably lower levels of exercise. A strong relationship between life happiness, emotional balance, and exercise frequency was found by the data analysis. Pupils who exercised on a regular basis reported feeling better about themselves subjectively than those who did not. The scientists optimized data processing and increased accuracy in assessing the correlations between variables by using a clustering technique based on swarm intelligence, which is comparable to ant colony models. Overall, [27] highlights the value of physical education in enhancing overall well-being and offers insightful recommendations for enhancing middle school kids physical and mental health. The authors come to the conclusion that encouraging young people to participate in sports, especially girls, is crucial to advancing their mental and physical development.

Authors of [28] use the *Hilbert-Huang Transform* (HHT) to analyze spatial spatiotemporal large data, addressing the difficulties in doing so. They understand that *non-stationary* signals, which are a feature of geographic data, cannot be processed by conventional techniques like *Fourier* or *wavelet* transformations. Authors suggest utilizing the HHT to calculate the instantaneous frequency of geographic data in order to get around these restrictions. This method offers a more precise and flexible representation of non-stationary and non-linear signals. They supplement this by assessing correlations between non-stationary variables across several scales using an absolute *entropy-based* correlation technique that is based on *Kullback-Leibler* (KL) divergence. This method captures both local and global correlations in the data, which enables a more thorough understanding of the connections between geographic elements. The authors used five geographic elements from Beijing and Tianjin across several decade population, civil vehicles, business volume of post and telecommunications, local phone users, and undergraduate students to validate their strategy. They show that, especially when dealing with *non-stationary* data, their method outperforms better-established correlation techniques like *spectral angle mapping* and *Pearson correlation*.

In summary, this research clearly demonstrates that by utilizing these enhanced correlation insights, the suggested method not only provides more precise correlation analysis for spatial spatiotemporal large data but also paves the way for future research, including predictive modeling.

3. Correlation Analysis over Big Multidimensional Datasets

In this Section, we introduce the main approach of this paper, which consists of presenting a comprehensive framework designed to facilitate efficient correlation analysis over big multidimensional datasets.

In the context of the rapidly evolving landscape of big data analytics, the need for frameworks that can handle complex, hierarchical data structures while preserving privacy is paramount. This is where the Drill-CODA framework [29] brings significant advancements to the field of multidimensional big data analytics.

Drill-CODA is a composite framework designed to support drill-across multidimensional big data analytics over big co-occurrence aggregate hierarchical data, while simultaneously preserving privacy. The framework addresses critical challenges faced by big data analytics, including the complexity and scale of real-life hierarchical datasets. By combining data processing metaphors and multidimensional analysis principles, Drill-CODA enhances the expressive power and accuracy of decision-making processes. In [29], Drill-CODA is presented as a framework that allows for the execution of drill-across queries in a way that both respects the hierarchical structure of the data and ensures privacy through aggregation and co-occurrence analysis. This innovative approach allows the extraction of valuable insights from big hierarchical data while maintaining privacy constraints through anonymization and multidimensional aggregation.

Figure 1 shows our proposed reference architecture for supporting efficient and effective correlation analysis over big multidimensional datasets.

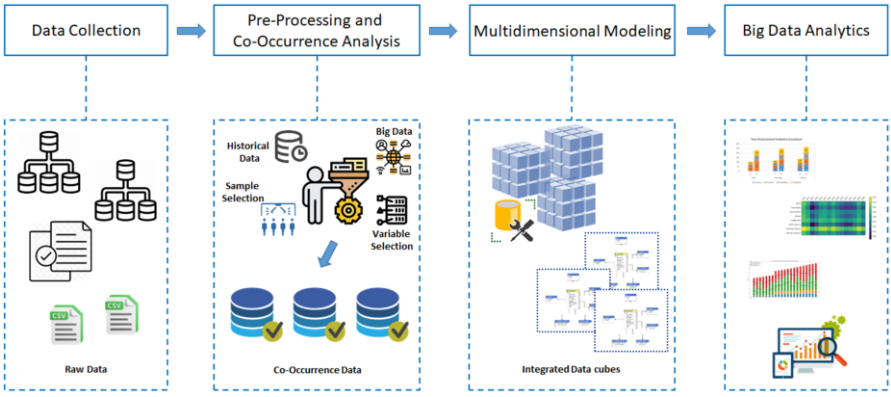

Figure 1. Reference Architecture for Supporting Correlation Analysis over Big Multidimensional Datasets.

As shown in Figure 1, our architecture encompasses several functional components that work collaboratively to enable correlation analysis over big multidimensional datasets, which we describe as follows:

- *Raw Data Collection Layer*: it is the foundation of the architecture, responsible for gathering all relevant data from various sources. These data come from multiple types of sources, including transactional databases, IoT sensors, social media platforms, and so forth.
- *Pre-Processing and Co-Occurrence Analysis Layer*: Once the raw data is collected, the next phase consists of a two-stage process. The first one is *data pre-processing*, which consists of cleaning data by removing duplicates, handling missing values, correcting data inconsistencies, as well as converting the data into a uniform format. The next phase is to apply *co-occurrence analysis* over the pre-processed datasets, which aims to identify patterns and correlations between different variables (e.g., how frequently two or more attributes occur together). Moreover, co-occurrence analysis can be used for *privacy-preserving aggregation*, as *sensitive* data can be anonymized during this stage.
- *Multidimensional Modeling Layer*: in this step, the co-occurrence data are organized into a *multidimensional data model*. This involves structuring data into OLAP data cubes, where dimensions such as *time* and *location* can be analyzed. This multidimensional modeling enables complex analyses like slicing, dicing, and pivoting of data.
- *Big Data Analytics and Visualization Layer*: this is the final layer of the architecture, where large-scale correlation analyses are performed over pre-processed and multidimensional data. This layer is designed to exploit the full potential of big data, by enabling highly efficient computation of complex queries and delivering valuable insights from vast and high-dimensional data collections. It supports a variety of analytical methods, including correlation analysis, trend discovery, and pattern recognition, through seamless integration with OLAP cubes.

However, it should be noted that in addition to the architecture implementation, [29] provides an extensive experimental assessment and evaluation of the Drill-CODA framework. The results demonstrated the framework's effectiveness in handling large-scale data processing while preserving privacy. The experimental setup involved real-life datasets, where Drill-CODA's performance in terms of processing time, scalability, and privacy preservation was rigorously tested. This further validates the practical application of the architecture described.

4. Conclusions and Future Work

Starting from the foundations of big data analytics, and the extended multidimensional big data analytics paradigm, this paper has presented models, issues and algorithms of correlation analysis tools for supporting big data analytics, along with the overview of a state-of-the-art proposal.

Future work is mainly oriented to *performance aspects* (e.g., [30-35]), which play a leading role when big multidimensional datasets must be accessed and processed.

Acknowledgments

This research is supported by the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing within the NextGenerationEU program (Project Code: PNRR CN00000013).

References

[1] Jones DE, Ghandehari H, Facelli JC. A Review of the Applications of Data Mining and Machine Learning for the Prediction of Biomedical Properties of Nanoparticles. *Computer Methods and Programs in Biomedicine* 2016; 132: 93–103, doi: https://doi.org/10.1016/j.cmpb.2016.04.025

[2] Focardi SM, Fabozzi FJ. *The Mathematics of Financial Modeling and Investment Management*, John Wiley & Sons, 2004.

[3] Russom P. Big Data Analytics. *TDWI Best Practices Report, Fourth Quarter* 2011; 19(4): 1–34.

[4] Tsai C, Lai C, Chao H, Vasilakos AV. Big Data Analytics: A Survey. *Journal of Big Data* 2015; 2: 21, doi: https://doi.org/10.1186/s40537-015-0030-3

[5] Rao P, Kwon J, Lee S, Subramaniam LV. Advanced Big Data Management and Analytics for Ubiquitous Sensors. *International Journal of Distributed Sensor Networks* 2015; 11: 174894, doi: https://doi.org/10.1155/2015/174894

[6] Stefanidis A, Crooks A, Radzikowski J. Harvesting Ambient Geospatial Information from Social Media Feeds. *GeoJournal* 2013; 78: 319–338, doi: https://doi.org/10.1007/s10708-011-9438-2

[7] O'Driscoll A, Daugelaite J, Sleator RD. 'Big Data', Hadoop and Cloud Computing in Genomics. *Journal of Biomedical Informatics* 2013; 46(5): 774–781, doi: https://doi.org/10.1016/j.jbi.2013.07.001

[8] Ahmed M, Choudhury N, Uddin S. Anomaly Detection on Big Data in Financial Markets. In: *9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 2017 July 31 - August 03; Sydney, Australia. pp. 998–1001, doi: https://doi.org/10.1145/3110025.3119402

[9] Spearman C. The Proof and Measurement of Association between Two Things. *International Journal of Epidemiology* 2010; 39(5): 1137–1150, doi: https://doi.org/10.1093/ije/dyq191

[10] Tran H, Hu J. Privacy-Preserving Big Data Analytics: A Comprehensive Survey. *Journal of Parallel and Distributed Computing* 2019; 134: 207–218, doi: https://doi.org/10.1016/j.jpdc.2019.08.007

[11] Zhou X, Liang W, Wang KI, Yang LT. Deep Correlation Mining Based on Hierarchical Hybrid Networks for Heterogeneous Big Data Recommendations. *IEEE Transactions on Computational Social Systems* 2021; 8(1): 171–178, doi: https://doi.org/10.1109/TCSS.2020.2987846

[12] Wu Z, Yin W, Cao J, Xu J, Cuzzocrea A. Community Detection in Multi-Relational Social Networks. In: *14th International Conference on Web Information Systems Engineering*; 2013 October 13-15; Nanjing, China. pp. 43–56, doi: https://doi.org/10.1007/978-3-642-41154-0_4

[13] Leung CK, Kajal A, Won Y, Choi JMC. Big Data Analytics for Personalized Recommendation Systems. In: *IEEE DASC/PiCom/DataCom/CyberSciTech*. 2019 August 5-8; Fukuoka, Japan; pp. 1060–1065, doi: https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00190

[14] Yang R, Yu L, Zhao Y, Yu H, Xu G, Wu Y, Liu Z. Big Data Analytics for Financial Market Volatility Forecast based on Support Vector Machine. *International Journal of Information Management* 2020; 50: 452–462, doi: https://doi.org/10.1016/j.ijinfomgt.2019.05.027

[15] Chandola V, Banerjee A, Kumar V. Anomaly Detection: A Survey. *ACM Computing Surveys* 2009; 41(3): 1–58, doi: https://doi.org/10.1145/1541880.1541882

[16] Cuzzocrea A. CAMS: OLAPing Multidimensional Data Streams Efficiently. In: *11th International Conference on Data Warehousing and Knowledge Discovery*; 2009 August 31 - September 2; Linz, Austria. pp. 48–62, doi: https://doi.org/10.1007/978-3-642-03730-6_5

[17] Cuzzocrea A, Darmont J, Mahboubi H. Fragmenting Very Large XML Data Warehouses via K-Means Clustering Algorithm. *International Journal of Business Intelligence and Data Mining* 2009; 4(3/4): 301–328, doi: https://doi.org/10.1504/IJBIDM.2009.029076

[18] Vervliet N, Debals O, Sorber L, De Lathauwer L. Breaking the Curse of Dimensionality Using Decompositions of Incomplete Tensors: Tensor-Based Scientific Computing in Big Data Analysis. *IEEE Signal Processing Magazine* 2014; 31(5): 71–79, doi: https://doi.org/10.1109/MSP.2014.2329429

[19] Wold S, Esbensen K, Geladi P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* 1987, 2(1-3): 37–52.

[20] Golub GH, Reinsch C. Singular Value Decomposition and Least Squares Solutions. In: *Handbook for Automatic Computation: Volume II: Linear Algebra*, Springer, pp. 134–151, 1971.

[21] Ding Z, Han J, Qian R, Shen L, Chen S, Yu L, Zhu Y, Liu R. eBoF: Interactive Temporal Correlation Analysis for Ensemble Data Based on Bag-of-Features. *IEEE Transactions on Big Data* 2023; 9(6): 1726–1737, doi: https://doi.org/10.1109/TBDATA.2023.3324482

[22] Jiang Z, Yuan Q, Lin R, Yang F. Canonical Correlation Analysis and Visualization for Big Data in Smart Grid. *IEEE Journal of Emerging and Selected Topics in Circuits and Systems* 2023; 13(3): 702–711, doi: https://doi.org/10.1109/JETCAS.2023.3290418

[23] Fu X, Huang K, Papalexakis EE, Song HA, Talukdar PP, Sidiropoulos ND, Faloutsos C, Mitchell TM. Efficient and Distributed Generalized Canonical Correlation Analysis for Big Multiview Data. *IEEE Transactions on Knowledge and Data Engineering* 2019; 31(12): 2304–2318, doi: https://doi.org/10.1109/TKDE.2018.2875908

[24] Aslam M. Analysis of Imprecise Measurement Data Utilizing Z-Test for Correlation. *Journal of Big Data* 2024; 11(1): 4, doi: https://doi.org/10.1186/s40537-023-00873-7

[25] Xiao C, Ye J, Esteves RM, Rong C. Using Spearman's Correlation Coefficients for Exploratory Data Analysis on Big Dataset. *Concurrency and Computation: Practice and Experience* 2016; 28(14): 3866–3878, doi: https://doi.org/10.1002/cpe.3745

[26] Han Y, Liu L, Sui Q, Zhou J. Big Data Spatio-Temporal Correlation Analysis and LRIM Model Based Targeted Poverty Alleviation through Education. *ISPRS International Journal of Geo-Information* 2021; 10(12): 837, doi: https://doi.org/10.3390/ijgi10120837

[27] Song Q, Rong B. Correlation Analysis of Middle School Students' Happiness and Sports in the Context of Big Data. *International Journal of Web-Based Learning and Teaching Technologies* 2024; 19(1): 1–14, doi: https://doi.org/10.4018/ijwltt.337605

[28] Song W, Wang L, Xiang Y, Zomaya AY. Geographic Spatiotemporal Big Data Correlation Analysis via the Hilbert-Huang Transformation. *Journal of Computer and System Sciences* 2017; 89: 130–141, doi: https://doi.org/10.1016/j.jcss.2017.05.010

[29] Cuzzocrea A, Soufargi S. Privacy-Preserving Big Hierarchical Data Analytics via Co-Occurrence Analysis. In: *13th International Conference on Data Science, Technology and Applications*; 2024 July 9-11; Dijon, France. pp. 93–103, doi: https://doi.org/10.5220/0012767800003756

[30] Cuzzocrea A, Saccà D, Serafino P. Semantics-Aware Advanced OLAP Visualization of Multidimensional Data Cubes. *International Journal of Data Warehousing and Mining* 2007; 3(4): 1–30, doi: https://doi.org/10.4018/jdwm.2007100101

[31] Cuzzocrea A, Moussa R, Xu G. OLAP*: Effectively and Efficiently Supporting Parallel OLAP over Big Data. In: *3rd International Conference on Model and Data Engineering*; 2013 September 25-27; Amantea, Italy. pp. 38–49, doi: https://doi.org/10.1007/978-3-642-41366-7_4

[32] Cuzzocrea A. Improving Range-SUM Query Evaluation on Data Cubes via Polynomial Approximation. *Data & Knowledge Engineering* 2006; 56(2): 85–121, doi: https://doi.org/10.1016/j.datak.2005.03.011

[33] Yu B, Cuzzocrea A, Jeong DH, Maydebura S. On Managing Very Large Sensor-Network Data Using Bigtable. In: *12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*; 2012 May 13-16; Ottawa, Canada. pp. 918–922, doi: https://doi.org/10.1109/CCGrid.2012.150

[34] Jiang R, Lu R, Choo KR. Achieving High Performance and Privacy-Preserving Query over Encrypted Multidimensional Big Metering Data. *Future Generation Computer Systems* 2018; 78: 392–401, doi: https://doi.org/10.1016/j.future.2016.05.005

[35] Ramdane Y, Boussaid O, Boukraâ D, Kabachi N, Bentayeb F. Building a Novel Physical Design of a Distributed Big Data Warehouse over a Hadoop Cluster to Enhance OLAP Cube Query Performance. *Parallel Computing* 2022; 111: 102918, doi: https://doi.org/10.1016/j.parco.2022.102918