

# Chemometrics: From Data Preprocessing to Fog Computing

Gerard G. Dumancas, Louisiana State University, Alexandria, USA

Ghalib Bello, Icahn School of Medicine at Mount Sinai, New York, USA

Jeff Hughes, RMIT University, Melbourne, Australia

Renita Murimi, Oklahoma Baptist University, Shawnee, USA

Lakshmi Viswanath, Oklahoma Baptist University, Shawnee, USA

Casey O. Orndorff, University of the Ozarks, Clarksville, USA

Glenda Fe G. Dumancas, Louisiana State University, Alexandria, USA

Jacy O'Dell, Oklahoma Baptist University, Claremore, USA

Prakash Ghimire, Louisiana State University, Alexandria, USA

Catherine Setijadi, Louisiana State University, Alexandria, USA

## ABSTRACT

The accumulation of data from various instrumental analytical instruments has paved a way for the application of chemometrics. Challenges, however, exist in processing, analyzing, visualizing, and storing these data. Chemometrics is a relatively young area of analytical chemistry that involves the use of statistics and computer applications in chemistry. This article will discuss various computational and storage tools of big data analytics within the context of analytical chemistry with examples, applications, and usage details in relation to fog computing. The future of fog computing in chemometrics will also be discussed. The article will dedicate particular emphasis to preprocessing techniques, statistical and machine learning methodology for data mining and analysis, tools for big data visualization, and state-of-the-art applications for data storage using fog computing.

## KEYWORDS

Big Data Analytics, Chemometrics, Fog Computing, Partial Least Squares, Pattern Recognition, Principal Component Analysis, Principal Component Regression

## INTRODUCTION

The rise of several hyphenated analytical techniques and their applications have led to the development of various chemometric methods in order to come up with meaningful information from the data generated by these instruments (Kumar, Bansal, Sarma & Rawal, 2014). The applications of chemometrics are extensive, ranging from multicomponent analysis in spectroscopy to the areas of bioinformatics, molecular genetics, and genetic epidemiology in recent years (Dumancas, 2012; Dumancas et. al., 2014; Dumancas et. al., 2015).

One of the areas of chemometrics is in Process Analytical Technology (PAT). PAT is an initiative designed to improve the efficiencies of both the manufacturing and regulatory processes by utilizing an integrated approach to quality analysis. One of the central cores of PAT is data analysis (Willis,

DOI: 10.4018/IJFC.2019010101

2004), which encompasses various chemometric tools. Thus, the advances that are now visible in PAT using chemometrics involve both the use of analytical instrumentation and mathematical methods for multivariate data analysis (Bogomolov, 2011; Dubrovkin, 2014; Kessler, 2013; Pomerantsev & Rodionova, 2012). The primary driving forces that led to the success of PAT would be the development of novel analytical methods and the continuous expansion of their applications (Dubrovkin, 2014).

As mentioned earlier, there has been rapid growth of data due to the rise of various analytical instruments. However, the main challenge comes from processing these data in a facile manner. In certain cases, multiple sensors are studying the same variables or compounds of interest. As such, the process of Data Fusion, a subclass of Chemometrics, is now considered an important topic (Esteban et. al., 2005; Ovalles & Rechsteiner, Jr., 2015). Multi-sensor Data Fusion is a tool used to combine the data from multiple sensors with the overall goal of providing a more reliable and accurate output (Castanedo, 2013; Rashinka & Krushnasamy, 2017). The Joint Directors of Laboratories (JDL) defines data fusion as a “multi-level and multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from various sources” (Steinberg et. al., 1999). The corresponding informational models emanating from data fusion should simulate extremely complex problems by fitting to the massive amount of empirical semi-structured and unstructured data (Isaeva et. al., 2012). Consequently, the algorithmic support and the interface of a computerized analytical system (often with limited computer resources) should be adjustable to systems with features of new types. Such challenge arising from analytical information management led to several new perspectives and solutions, such as the concept of Cloud Computing, all of which are part of the development of “Big Data Approach” (BDA) (Dubrovkin, 2014). Cloud Computing can simply be defined as the operation of computer power or storage on remote servers by means of a network. Using the Cloud, very high-level services with high computational power is now possible. Fog computing, on the other hand, constitutes the layer below Cloud computing in connected Things (Paret and Huon, 2017). In other words, Fog Computing is an extension of the Cloud Computing paradigm to the edge of the network, thus enabling a new breed of applications and services (Bonomi et al, 2012).

In this manuscript, the major aspects of Big Data utilization and processing in Analytical Chemistry (Chemometrics), specifically some commonly used algorithmic and instrumental techniques and aspects of computerized analytical systems, will be discussed. An interesting discussion will also be the role of fog computing in chemometrics.

## **BACKGROUND**

Chemometrics is a rapidly evolving field, which has multitude of applications in both descriptive and predictive problems in experimental life sciences, especially in Chemistry. It is considered to be a highly interdisciplinary area which includes Multivariate Statistics, Computer Science, and Applied Mathematics utilizing methods employed in core data analytics. The primordial goal of chemometrics is to address problems in various fields including that of Biochemistry, Medicine, Chemistry, Chemical Engineering, and Biology among others (Khanmohammadi, 2014).

As mentioned earlier, the field of Chemometrics has a wide array of applications especially in Biology and Medicine. For example, Support Vector Machines (SVMs) and Partial Least Squares Discriminant Analysis (PLS-DA) are widely used techniques for classification purposes involving microorganisms, medical diagnosis using spectroscopy, and metabolomics using Coupled Chromatography and Nuclear Magnetic Resonance Spectrometry (Brereton, 2007). Studies for the determination of quality of medicinal plants and standardization of herbal drugs utilize liquid and gas chromatography with various chemometric techniques such as Principle Component Analysis (PCA) and Linear Discriminate Analysis (LDA), among others, to analyze complex chemical compositions of these medicines. The aforementioned techniques continue to be expanded as new methods for quality estimation and are regularly explored (Bansal, Chhabra, Rawal, & Sharma, 2014).

40 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/chemometrics/219359](http://www.igi-global.com/article/chemometrics/219359)

## Related Content

---

### Advanced Brain Tumor Detection System

Monica S. Kumar, Swathi K. Bhat and Vaishali R. Thakare (2020). *International Journal of Fog Computing* (pp. 31-45).

[www.irma-international.org/article/advanced-brain-tumor-detection-system/266475](http://www.irma-international.org/article/advanced-brain-tumor-detection-system/266475)

### Designing and Analysis of Antenna Using Back Propagation Network

Rajeev Kumar, Ritu Vijay and Surjit Singh (2019). *Handbook of Research on the IoT, Cloud Computing, and Wireless Network Optimization* (pp. 453-490).

[www.irma-international.org/chapter/designing-and-analysis-of-antenna-using-back-propagation-network/225730](http://www.irma-international.org/chapter/designing-and-analysis-of-antenna-using-back-propagation-network/225730)

### Texture-Based Evolutionary Method for Cancer Classification in Histopathology

Kiran Fatima and Hammad Majeed (2016). *Managing and Processing Big Data in Cloud Computing* (pp. 55-69).

[www.irma-international.org/chapter/texture-based-evolutionary-method-for-cancer-classification-in-histopathology/143339](http://www.irma-international.org/chapter/texture-based-evolutionary-method-for-cancer-classification-in-histopathology/143339)

### IoT-Fog-Blockchain Framework: Opportunities and Challenges

Tanweer Alam (2020). *International Journal of Fog Computing* (pp. 1-20).

[www.irma-international.org/article/iot-fog-blockchain-framework/266473](http://www.irma-international.org/article/iot-fog-blockchain-framework/266473)

### Fog Computing Qos Review and Open Challenges

R. Babu, K. Jayashree and R. Abirami (2018). *International Journal of Fog Computing* (pp. 109-118).

[www.irma-international.org/article/fog-computing-qos-review-and-open-challenges/210568](http://www.irma-international.org/article/fog-computing-qos-review-and-open-challenges/210568)