From Big Data to Smart Data: Application to performance management

Amel Souifi, Zohra Cherfi-Boulanger, Zolghadri Marc, Maher Barkallah, Mohamed Haddar

From Big Data to Smart Data: Application to performance management

Amel Souifi [*] **Zohra Cherfi Boulanger** [**] **Marc Zolghadri** [*,****]
Maher Barkallah [***] **Mohamed Haddar** [***]

[*] *Quartz laboratory, SUPMECA, 3 Rue Fernand Hainaut, 93407, Saint Ouen, France (e-mail:amel.souifi@supmeca.fr).*
[**] *Roberval laboratory, Université de Technologie de Compiègne, 60203 Compiègne cedex, France (e-mail: zohra.cherfi-boulanger@utc.fr)*
[***] *LA2MP Ecole Nationale d'Ingénieurs de Sfax, Université de Sfax, LA2MP, Route Soukra Km 3.5, BP 1173, 3038 Sfax, Tunisia (e-mail: maher.barkallah@enis.tn)*
[****] *LAAS-CNRS, 7 Avenue du Colonel Roche, 31400 Toulouse, France (e-mail: marc.zolghadri@supmeca.fr).*

Abstract: In the context of digitalization, some companies are considering a transition to Industry 4.0 to ensure greater flexibility, productivity and responsiveness. The implementation of a relevant performance management system is then a real necessity to measure the degree of achievement of these objectives. In the era of Industry 4.0, the potential access to large amounts of data, i.e. Big Data, poses new challenges to the design and implementation of these systems. With the exponential growth of data generated from different sources, there is a need for extensive exploitation of data for performance management. Given the large volume of data, the speed at which it is generated and the variety of data sources, the manufacturing sector is facing with the challenge of creating value from large data sets. This paper introduces some potential benefits of Big Data for business and in particular its role in performance management systems. However, the key idea is that Big Data are not always neither available nor necessary. Authors focus on the concept of smart data, the result of the transformation of Big Data, and define a set of necessary and sufficient conditions the data should satisfy to be considered as Smart. The paper presents some methods of smart data extraction. Such smart data will be used to feed the performance management system in order to obtain more accurate, timely and representative key performance indicators.

Keywords: Big Data, Smart Data, Performance Management.

1. INTRODUCTION

Companies need to manage their processes continuously and improve their performance to achieve their objectives. Performance management systems, PMS in short, are an efficient tool to meet this purpose. Performance management translates the organization's critical success factors into a set of metrics to communicate critical objectives and support decision making (Bititci et al., 2015). It involves defining the organization's objectives, identifying the key performance indicators (KPIs) likely to measure the degree of achievement of the objective, evaluating performance level and implementing an action plan. Performance Measurement is much studied in the literature and several frameworks are proposed, including the SCOR model (APICS, 2017), the Balanced Scorecard (Kaplan et al., 2005), and the ECOGRAI framework (Bitton, 1990). The KPIs are used to quantify the company's activities and reflect its critical success factors, (Zhu et al., 2018). The ISO 22400 standard (ISO, 2014) has listed a set of criteria to ensure the usefulness of KPIs. These criteria include accuracy, timeliness, relevance, correctness and completeness. Since the indicators result from a combination of internal and/or external data, they should satisfy the same requirements.

In recent years, Big Data has received increasing attention from academic and industry. In the era of industry 4.0, noted hereafter I4.0, a huge amount of data is generated due to the interconnection of objects enabled by the internet of things and cyber-physical systems. Big Data is characterized in literature by 5Vs: Volume, Velocity, Variety, Veracity and Value (Elia et al., 2020). The volume stands for the big quantity of data generated, the variety for the heterogeneous sources and velocity for the high speed of generating the data. Veracity and value have been included to mention the concern with Big Data quality. While veracity refers to data uncertainty and doubt (Lozano et al., 2020), value is related to data use (Iafrate, 2015). Real-time data collection has great potential for the manufacturing industry if useful information can be extracted from it. At a first glance, Big Data seems to be valuable. But it is just data and it needs processing to be useful. In this context, we study the extraction of relevant pieces of relevant information (Smart data) from Big Data to support decision making.The processing of

data to extract relevant information is possible thanks to a wide range of statistical methods and artificial intelligence techniques. Our contribution here is limited to the methodology for extracting Smart data. We show that obtaining Smart data is dependant on the quality of data.

The rest of the paper is organized as follows. We present first characteristics of Big Data and their potential benefits for business and we focus on the concept of smart data. Then, we introduce a simulation model to generate simulation data to be used later and we define a set of conditions to get Smart data. Finally we present some techniques to process Big Data transforming it into Smart Data.

2. DATA-DRIVEN DECISION MAKING

2.1 Big data uses

With the increasing rate of data generation, companies can take advantage of it to improve their performance. Massive data makes it easier to describe different aspects of complex systems and evaluate their performance. It could enable business process monitoring and then improve productivity and minimizes inventory. For instance, in logistics, Radio Frequency identification (RFID) enables collecting huge amounts of real-time data to support logistics planning and scheduling (Zhong et al., 2015). Marketing is also taking advantage from collecting and analyzing consumers' data to fulfill their needs via targeted products and create value (Elia et al., 2020). The use of new technologies and social media has enabled the generation of large amounts of structured and unstructured data. The exploitation of this data would improve the decision-making process, better understand business and reduce risks (Mello and Martins, 2019). The processing of external data allows the company to understand the evolution of its environment and to know its opportunities and threats to better define their strategies. Organizations should stay connected to their environment especially with the rapid transformation of the industry revolutionizing processes and customer needs. The fall of the Finnish giant Nokia shows that it is not enough to control processes, but innovation is needed to meet the needs of customers and keep competitive. Big Data has also shown potential for crisis management. In response to the Covid-19 pandemic, Taiwan analyzed data available in its health insurance, immigration and customs databases to facilitate case identification (Wang et al., 2020).

As data are of great importance for different fields, it will be necessary to define which data are useful. In this sense, we introduce the concept of smart data.

2.2 Smart Data

A small clean dataset can provide more ideas than a large amount of noisy data. Small data is data *intentionally* collected (Faraway and Augustin, 2018) to meet company's requirements like sales and customer data, financial data and production statistics. It is a structured data which is managed typically by Data Base Management System (DBMS). With the emergence of connected devices, sensors, wireless sensor networks (WSN) and internet of things, huge amounts of data are generated and collected. Big Data may exist in usable or unusable form and do not represent an end in itself (Bellinger et al., 2004). In fact, "an ounce of information is worth a pound of data. An ounce of knowledge is worth a pound of information, an ounce of understanding is worth a pound of knowledge"(Ackoff, 1989).

In this context, the concept of Smart Data has been introduced. García-Gil et al. (2019) define Smart data as "high quality, clean and valuable data". They identify three attributes for data to be smart: accuracy, actionability and agility. Actionable data is used to drive scalable actions to maximize business objectives. Agility is about the availability of data in real time and its capacity to adapt to the changing environment. Smart Data is also seen as actionable knowledge (Lenk et al., 2015) resulted from an intelligent processing and transformation of unstructured massive data. Smart Data refers also to the way of bringing together and analyzing different data sources to support decision making and action (Iafrate, 2015).

3. STUDY OF CONDITIONS FOR SMART DATA

3.1 Case study

Data used in the rest of the paper is collected from a simulation model of a perfume filling line.The production line consists of nine workstations inline (figure 1) to produce perfume bottles with a given throughput of 1500p/hour. To simulate a Big Data set, 74 performance indicators related to productivity such as setup time, number of failures of each machine and conveyor, good units processes by each workstation, good units conveyed, number of unis accumulated on conveyors and output rates for all workstations and conveyors, were collected every minute over a month (24/7). The data set is represented by a matrix of size (43200 x 74).

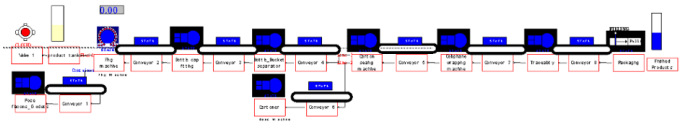

Fig. 1. Simulation model

3.2 Conditions for Smart Data

Smart Data can be extracted from small and Big Data to support decision making and improve business performance. We define a function R to measure the data relevance, defined by 3 parameters: data quality, quantity and cost. To get Smart Data, we need then to set conditions for each parameter to optimize R.

Data quality The amount of data collected has a potential for monitoring in industry but it brings also quality issues for manufacturers. They bring non-qualities related to the process of data collection, storage, processing and analysis. For example, when estimating the proportion of doctor visits for flu-like illnesses by aggregating Google search queries, Google Flu Trends (GFT) predicted twice the number of visits recorded by the Centers for Disease Control and Prevention (CDC), according to (Lazer et al.,

2014). Researchers have defined dimensions for data quality mainly accuracy, reliability, completeness, consistency and timeliness. For this purpose, data cleaning is essential before data exploitation. Metadata is also necessary to ensure data quality as it inform about the data sources, time of collection, whom collected it and for which purpose.

Data quantity A sufficient amount of data is required to ensure the representativeness of the business and evaluate its overall performance. But a large volume of data is not always useful. For example, in order to detect the position of the elements of a conveyor, we could place sensors covering the entire conveyor and collect data continuously. But this quantity, although its representativeness and completeness, is not necessary. If we consider the parameters monitored on a process as variables, we will need to identify (i) the right number of variables collected (ii) at a right and relevant collection frequency. The "amount of information" required depends on the decision-making level (Townsend et al., 2018). As top managers make strategic decisions, they have more interest in aggregated data from the production shop and external information to respond to stakeholders' expectations. Tactical and operational decision-making levels need more frequent data to ensure the achievement of objectives in mid/ short terms. Then, the frequency of data collection increases from strategic to operational decision-making level. The uncertainty about the process is another factor to be taken into consideration for determining the sufficient amount of data. It is the result of the complexity of the system under study and/or the lack of expertise of the decision-maker. (Galbraith, 1995) pointed that the amount of information required is a function of the task uncertainty to achieve the expected level of performance. As I4.0 is based on concepts such as cyber-physical systems, the Internet of Things and the Internet of Services, it combines the digital and physical worlds and thus increases the complexity of production systems. More data must therefore be processed to make the tasks more transparent and reduce uncertainty. According to (Daft and Lengel, 1986), organizations process data to reduce uncertainty and equivocality. When (Galbraith, 1995) defined uncertainty as "the difference between the amount of information required to perform the task and the amount of information already possessed by the organization", (Daft and Lengel, 1986) presented equivocality as ambiguity and lack of understanding resulting in conflicting interpretations about an organizational situation. They pointed that technology, interdepartmental relationships and environment are the sources of organizational uncertainty and equivocality. As one of the goals of I4.0 is agility, decisions must be made faster, which requires more frequent data collection for faster decision making. According to Nyquist–Shannon sampling theorem, a reconstruction of the original signal requires a sampling frequency greater than twice the frequency of the sampled signal.

As mentioned above a large amount of data does not guarantee their usefulness. In order to confirm this hypothesis, we rely on Shannon's entropy to assess the amount of information contained in a data set (Cover and Thomas, 2012). Let X be a discrete random variable defined on X and probability distribution mass function $p(x) = P(X = x)$, $x \in X$. The entropy $H(X)$ is defined by:

$$H(X) = -\sum_{x \in X} p(x) \log(p(x)) \tag{1}$$

The conditional entropy of a random variable Y given another variable X is defined by:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log(p(y|x)) \tag{2}$$

If we consider a set of variables $X_1, X_2, \ldots, X_n$ drawn according to $p(x_1, x_2, \ldots, x_n)$, then the entropy of these variables is the sum of the conditional entropies:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_i - 1, \ldots, X_1) \tag{3}$$

To illustrate the relationship between the amount of information contained in a data set and the period of data collection, we used data from the simulation model described above. Two calculations are then made:

- First scenario:The model is then forced to extract data at various periods of collection, from one minute to one month, i.e simulation data are collected only once at the end of simulation. The figure 2 shows that the amount of information contained in the data decreases as the period of data collection increases. This decrease is significant by passing, for example, from a period of one minute to three hours (the entropy passes from 9.8 to 5.8nats (natural unit of information). But by passing from one minute to three minutes the entropy decreases by 7% and therefore it is possible to sacrifice some information in order to reduce the quantity of data and reduce storage and processing costs.

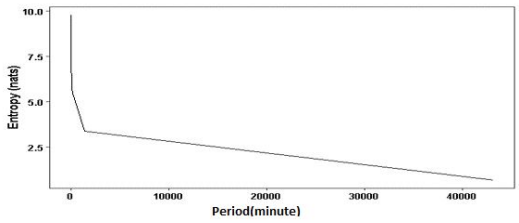

Fig. 2. The entropy values for different periods of data collection

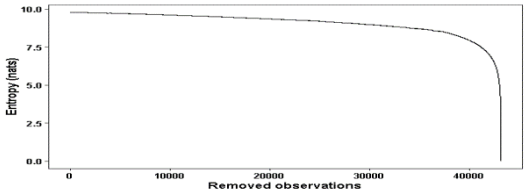

Fig. 3. Variation of entropy related to removed observations

- Second scenario: To simulate the loss of information resulting from the elimination of observations, data points were removed randomly, step by step, from the data set. The entropy variation starts to be significant only after the removal of 30 000 observations (figure 3). This shows that useful information is contained in a reduced number of observations.

- Conclusion: These observations show that higher period of data collection does not necessarily imply more information. However, a periodic removal of observations has a greater influence on the amount of information lost from data than a random removal of values.

On the other hand, high dimensionality data present problems related to their manipulation mainly their visualization, analysis and modelling. Therefore, a reduction of dimensionality is necessary while keeping the maximum amount of information. Then the challenge is to find a compromise between a sufficient and manageable amount of data. A large amount of data aims at improving the representativeness of the business and reducing the uncertainty. So representativeness increases with the increase of data quantity. But this increase in quantity is not without consequences. Indeed, it leads to additional costs and difficulty to manage the data. If we represent the data representativeness as an increasing function of data quantity while data manageability as a decreasing function that tends towards 0 when the data quantity is very large, we need to find a quantity of data Q* that optimizes both functions(figure 4).

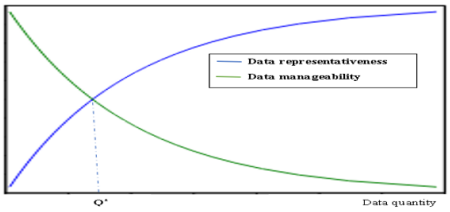

Fig. 4. Data representativeness and manageability as functions of data quantity

Affordable data cost The acquisition of large data analysis solutions involves high additional costs that may be unjustifiable if the data are of low quality and value. By comparing costs of Big Data Analytics appliances, (Jacob et al., 2018) reported that three year-costs for use of IBM PureData for Analytics and a Hadoop Cluster (Cloudera) are respectively 39 millions and 50 millions.Significant investments required for Big data Analytics is not sufficient to ensure performance improvement (Mikalef et al., 2019). Therefore, it is necessary to clearly define the company's strategy and objectives in relation to Big Data and assess its return on investment.

4. PROCESS TO GET SMART DATA

In order to obtain smart data, it is essential that the data used meet the necessary and sufficient conditions mentioned above, i.e. quality data, in sufficient quantity and at an affordable cost. Then process of extracting smart data starts by data pre-processing, analysis,interpretation and visualization. Data pre-processing is an essential phase to get actionable knowledge from Big Data. It includes many techniques mainly data cleaning, integration, discretization and dimensionality reduction techniques like principal component analysis. We propose here to use some statistical methods of data pre-processing in order to reduce the dimensionality of the data and to define a sufficient number of observations in a way that does not overload the data set. The methods to be used and their sequence depend only on the nature of the data and their use. If we were in the context of defining the amount of data to be collected, we could start by defining a frequency of data collection and then reduce the dimensionality. In our case, we will start by reducing the dimensionality so that we could use the time series segmentation methods implemented on R software. As variables are numerical, we use principal component analysis to reduce data dimensionality and then clustering of variables to discover insights and preliminary knowledge. Then, we are interested in time series segmentation to define a a frequency for data collection.

4.1 Principal component analysis

Principal Component Analysis (PCA) is powerful to explore the structure of multidimensional data. The PCA aims to reduce the dimensionality of a data set containing a large number of variables while retaining the maximum of the variation present in the data set (Jolliffe and Cadima, 2016). Let X be an (n*m) matrix of observations. The n rows refer to individuals and the m columns to variables.The PCA allows to synthesize the m variables in p artificial variable with $p \ll m$. The new artificial variables are named principal components. The main limitation is the difficulty to use the extracted variables only to understand the data (Dash et al., 1997).

- Scenario: Since high dimensional data are very difficult to manage and value, and therefore far from being relevant (Saporta, 2006), we initially applied PCA to our matrix of simulation data of size (43200 x 74). The result shows that 78% of the data can be synthesized into two main components. The choice of the number of dimensions to be retained is an important issue in the application of PCA (Saporta, 2006). An empirical method consists in studying the eigenvalue decay curve (Scree plot) (figure 5) in order to detect a bend indicating a change in structure. By studying the correlation between the initial variables and the synthetic components, we can interpret the first synthetic variable as the production slowdown and the second component as the production rate.
- Conclusion: By applying PCA data is summarized into a reduced number of synthetic variables. This aggregation of data, while informative, is still insufficient to be able to exploit the data. Indeed, this synthesis of variables is abstract and difficult to interpret.The inadequacy of PCA results leads us to apply the clustering of variables to facilitate the interpretation of PCA results and the reduction of dimensionality.

4.2 Clustering of variables

Clustering is another basic means of exploratory data analysis (Pedrycz, 2005). It consists of forming groups from the data based on a measure of similarity. The data in the same group have high similarities to each other but low similarities to the data in the other groups (Zhang et al., 2017). Clustering can be applied to variables in order to identify homogeneous groups of variables that are highly correlated and therefore bringing the same information.

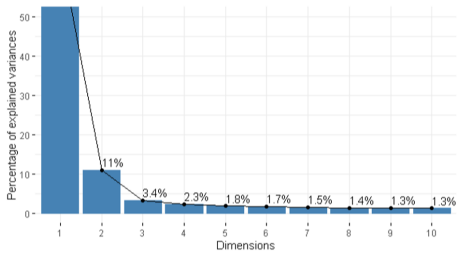

Fig. 5. Scree plot:Evolution of the variance explained by each dimension for the ten first dimensions

Clustering around latent variables (CLV) approaches lump together correlated variables and synthesize them into a variable capturing the maximum information. CLV techniques overcome the limitations of the analysis in main components essentially the difficulty of interpreting the results (Wang et al., 2017).

- Scenario: We applied clustering around latent variables using the commands of R's ClustVarLV package. The clustering results in figure 6 show that the clustering criterion drops significantly when moving from 2 clusters into one cluster.
- Conclusion: The application of variable grouping confirmed the results of the PCA. The first group includes variables related to the number of breakdowns and downtime. The second cluster relates to output rates. Clustering facilitates the interpretation of the main components.

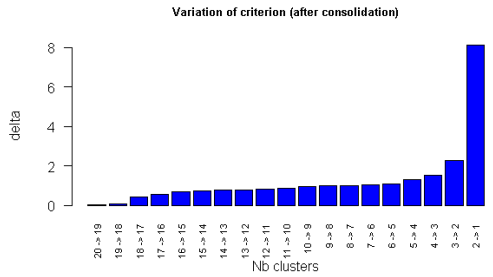

Fig. 6. Variation of the clustering criterion when passing from K to $(K-1)$ clusters of variables

4.3 Data as time series

Production systems are multi-components and multi-states systems. Then, to evaluate a company's performance, several variables and are monitored over time. The data representing the evolution of these variables define multivariate time series (MTS). Time series can be used to reflect the interaction between these components and capture their dynamics (Lenk et al., 2015). MTS mining allow patterns and rules discovery in high dimensional data. MTS mining tasks include preprocessing, clustering, segmentation, classification, prediction and visualization (Lovrić et al., 2014). Since performance indicators must describe the state of the system under study, they must be representative and accurate. As data can be collected in real time, a definition of the periodicity of indicator calculation is necessary. The periodicity can be formulated as a problem of segmentation of MTS for several reasons.

With the huge amount of manufacturing data containing large numbers of records and many variables, it is not possible to exploit all of them directly by a decision-maker. To overcome this issue, sufficient amount of information is transferred to the decision points at appropriate times. Then it is useful to find a precise representation of the data that allows the reduction of their dimensionality. For this purpose, time series segmentation can be applied. Time series segmentation aim at stable periods of time location, change points identification or time series compression (Abonyi et al., 2005). Hence, segmentation enables the detection of change points in the data structure and thus reflects changes in the system's behavior. In the hierarchy of performance indicators, the lowest level KPIs are derived from process measures and higher-level KPIs are calculated from lower-level KPIs (ISO, 2014). Therefore, the definition of data collection frequency and processing points is an essential to determine the periodicity of generating and monitoring KPIs. There are always delays in making decisions based on available information and in taking action (Forrester, 1997). In this case a representation of data at each time interval would be sufficient to assess performance. A multivariate time series $T = \{x_k | 1 \leq k \leq N\}$ is defined by a sequence of N observations, ordered in time, on a set of n variables. An observation x_k is given by $x_k = [x_{1k}, x_{2k}, \ldots, x_{nk}]^T$. Given the large number of variables, we propose to apply principal component analysis to the data and perform segmentation on the first two components obtained.

- Scenario: We applied PCA on 24-hours simulation data (a sub-matrix (1440*74) of the initial matrix). then, we use the segmentation function based on dynamic programming algorithm to perform the segmentation of the data consisting of projections of individuals on principal components. Results are given by figure 7 that regroups two sub-figures each representing the variation of the first and second synthetic variables as a function of time. This figure shows that we can divide the time series into 5 homogeneous segments (given by the five different colours). The starting positions of each segment are:1, 263, 512, 878 and 1222, which means a change in system behaviour every 5 hours on average.

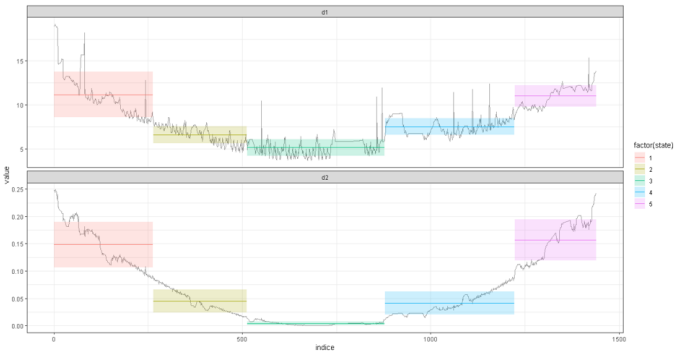

Fig. 7. Data segmentation based on principal components

- Conclusion: The results of data segmentation showed that we can synthesize observations to a reduced number (five observations), which makes it easier for a decision-maker to access and analyze data.

5. CONCLUSION

Transforming Big Data into Smart Data minimizes costs for data storage, transfer, processing and analysis. Smart Data can be used to develop digital twins of different elements of the production shop floor. This enables the data to be integrated through a set of models and facilitates performance measurement and management. The challenge is then to be able to exploit the smart data to dynamically feed the models of the digital twin.

REFERENCES

Abonyi, J., Feil, B., Nemeth, S., and Arva, P. (2005). Modified gath–geva clustering for fuzzy segmentation of multivariate time-series. *Fuzzy Sets and Systems*, 149(1), 39–56.

Ackoff, R.L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1), 3–9.

APICS (2017). *Supply Chain Operations Reference Model SCOR Version 12.0*.

Bellinger, G., Castro, D., and Mills, A. (2004). Data, information, knowledge, and wisdom.

Bititci, U.S., Garengo, P., Ates, A., and Nudurupati, S.S. (2015). Value of maturity models in performance measurement. *International journal of production research*, 53(10), 3062–3085.

Bitton, M. (1990). *ECOGRAI: Méthode de conception et d'implantation de systèmes de mesure de performances pour organisations industrielles*. Ph.D. thesis, Bordeaux 1.

Cover, T.M. and Thomas, J.A. (2012). *Elements of information theory*. John Wiley & Sons.

Daft, R.L. and Lengel, R.H. (1986). Organizational information requirements, media richness and structural design. *Management science*, 32(5), 554–571.

Dash, M., Liu, H., and Yao, J. (1997). Dimensionality reduction of unsupervised data. In *Proceedings ninth ieee international conference on tools with artificial intelligence*, 532–539. IEEE.

Elia, G., Polimeno, G., Solazzo, G., and Passiante, G. (2020). A multi-dimension framework for value creation through big data. *Industrial Marketing Management*.

Faraway, J.J. and Augustin, N.H. (2018). When small data beats big data. *Statistics & Probability Letters*, 136, 142–145.

Forrester, J.W. (1997). Industrial dynamics. *Journal of the Operational Research Society*, 48(10), 1037–1041.

Galbraith, J.R. (1995). *Designing organizations: An executive briefing on strategy, structure, and process*. Jossey-Bass.

García-Gil, D., Luengo, J., García, S., and Herrera, F. (2019). Enabling smart data: noise filtering in big data classification. *Information Sciences*, 479, 135–152.

Iafrate, F. (2015). *From big data to smart data*, volume 1. John Wiley & Sons.

ISO (2014). Iso 22400-1:2014 automation systems and integration — key performance indicators (kpis) for manufacturing operations management — part 1: Overview, concepts and terminology. International Standard Organization (ISO).

Jacob, R.N., Shirwadkar, P.G., and Singh, M.A. (2018). Analysis of ibm puredata system with hadoop implementations for structured analytics. *International Research Journal of Engineering and Technology (IRJET)*, 05, 1258–1260.

Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

Kaplan, R.S., Norton, D.P., et al. (2005). The balanced scorecard: measures that drive performance. *Harvard business review*, 83(7), 172.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176), 1203–1205.

Lenk, A., Bonorden, L., Hellmanns, A., Roedder, N., and Jaehnichen, S. (2015). Towards a taxonomy of standards in smart data. In *2015 IEEE International Conference on Big Data (Big Data)*, 1749–1754. IEEE.

Lovrić, M., Milanović, M., and Stamenković, M. (2014). Algoritmic methods for segmentation of time series: An overview. *Journal of Contemporary Economic and Business Issues*, 1(1), 31–53.

Lozano, M.G., Brynielsson, J., Franke, U., Rosell, M., Tjörnhammar, E., Varga, S., and Vlassov, V. (2020). Veracity assessment of online data. *Decision Support Systems*, 129, 113132.

Mello, R. and Martins, R.A. (2019). Can big data analytics enhance performance measurement systems? *IEEE Engineering Management Review*, 47(1), 52–57.

Mikalef, P., Boura, M., Lekakos, G., and Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. *Journal of Business Research*, 98, 261–276.

Pedrycz, W. (2005). *Knowledge-based clustering: from data to information granules*. John Wiley & Sons.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.

Townsend, M., Le Quoc, T., Kapoor, G., Hu, H., Zhou, W., and Piramuthu, S. (2018). Real-time business data acquisition: How frequent is frequent enough? *Information & Management*, 55(4), 422–429.

Wang, C.J., Ng, C.Y., and Brook, R.H. (2020). Response to covid-19 in taiwan: big data analytics, new technology, and proactive testing. *Jama*, 323(14), 1341–1342.

Wang, E., Alp, N., Shi, J., Wang, C., Zhang, X., and Chen, H. (2017). Multi-criteria building energy performance benchmarking through variable clustering based compromise topsis with objective entropy weighting. *Energy*, 125, 197–210.

Zhang, Q., Zhu, C., Yang, L.T., Chen, Z., Zhao, L., and Li, P. (2017). An incremental cfs algorithm for clustering large data in industrial internet of things. *IEEE Transactions on Industrial Informatics*, 13(3), 1193–1201.

Zhong, R.Y., Huang, G.Q., Lan, S., Dai, Q., Chen, X., and Zhang, T. (2015). A big data approach for logistics trajectory discovery from rfid-enabled production data. *International Journal of Production Economics*, 165, 260–272.

Zhu, L., Johnsson, C., Varisco, M., and Schiraldi, M.M. (2018). Key performance indicators for manufacturing operations management–gap analysis between process industrial needs and iso 22400 standard. *Procedia Manufacturing*, 25, 82–88.