Using Big Data Analytics for Money Laundering Detection – A Case Study

Shih-Che Lo †
Department of Industrial Management
National Taiwan University of Science and Technology, Taipei, Taiwan
Tel: (+886) 2-2737-6351, Email: sclo@mail.ntust.edu.tw

Tzung-Shian Li
Department of Industrial Management
National Taiwan University of Science and Technology, Taipei, Taiwan
Tel: (+886) 2-2733-3141 ext. 7479, Email: M10401203@mail.ntust.edu.tw

Abstract. Money laundering is the process that criminals conceal or disguise their crimes and redirect those proceeds into goods or services. Examples of illegal sources of income are betting operations, drug trafficking, illegal gambling and bribery. In this paper, we applied the big data analytics for a case company to detect possible money laundering activities. A partial database from an excel data file with 18,000 transactions along with brief summary report were analyzed beginning with data cleaning, traditional statistics analysis, and data mining process. Autocorrelation functions and partial autocorrelation functions were also conducted to analyze the relationships of attributes in the data set before performing the big data analytics methods. Finally, several time series forecasting methods, including regression methods, exponential smoothing methods, and predictive analytics were used to provide big data approaches and generate reports for decision makers as detection of money laundering activities. Computational results were implemented by using the Minitab™ and R software.

Keywords: big data analytics; time series forecasting; decision making; predictive analytics; data mining

1. INTRODUCTION

Money laundering is the process that illegal or dirty money is put through a cycle of transaction, so that it comes out the other end "look like" as legal or clean money. In other words, illegally acquired money is obscured through a series transfers and/or deals in order to make those funds to be appeared as legitimate cash. Nowadays, it poses a serious threat not only to financial institutions globally but also to the countries internationally. There are risks faced by financial institutions needed to be considered when complete any transactions, such as reputation risk, operational risk, concentration risk, and legal risk. Therefore, the governments and financial regulators require financial institutions to implement processes and procedures to detect money laundering as well as the financing of terrorism. Moreover, anti-money laundering (AML) is significant important to both national financial stability and international security (Gao and Weng, 2006).

Money launderers often use bank accounts of legitimate-looking businesses in some "famous" countries to circulate the dirty money through the financial system. However, suddenly injecting huge amounts of cash into those accounts usually attracts attentions significantly. The typical transacting behavior of that account will serve as a benchmark to measure the latest transactions by "flagged" customers. Money laundering usually involves three stages: placement, layering, and integration. Placement is the movement of cash from its source. On occasion the source can be easily disguised or misrepresented. This is followed by placing it into circulation through financial institutions, casinos, shops, and/or other businesses. Layering is to make it more difficult to detect and uncover a laundering activity. Integration is the movement of previously laundered money into the economy mainly through the banking system and thus such monies appear to be normal business earnings.

Liu and Zhang (2007) established agent-based AML system architecture to consideration of not only transaction information database in financial institutions, but also customer profiles information and external information. Lv et

al. (2008) proposed a RBF neural network model that can reach high correction rate in reducing false positive rate and enhancing detection rate remarkably, providing a new method to detect the suspicious transaction based on APC-III clustering algorithm and recursive least square algorithm for AML.

Gao (2009) designed a new cluster-based local outlier factor (CBLOF) algorithm to identify SMLTBPs that can effectively identify the synthetic data suspicious of money laundering transactions with a high processing speed and a satisfactory accuracy and used synthetic data experimentally to test its applicability and effectiveness. Khac and Kechadi (2010) presented a case study using a data mining-based approach for analyzing transactional data in an investment bank to detect money laundering patterns. Liu et al. (2011) presented a core decision tree algorithm that every financial data, they search from the node to the leaf gradually and in the leaf. Hong et al. (2015) proposed the optimal AML resource management with peer to peer anti-money laundering resource allocation model based on SMDP, considering both the maximal system reward and the system process cost of suspicious transaction reports.

Industry 4.0 was proposed by the German Government, and represents the implementation of artificial intelligence, big data, and the Internet of Things (IoT) in the factories (Stock and Seliger, 2016). History about industrial revolution is that the first industrial revolution at the end of the 18th century saw the birth of manufacturing using machines. Next, the second industrial revolution came in the beginning of the 20th century that mass production lines were powered by electric energy. The third industrial revolution in 1970s changes from analogue and mechanical production into electronic and digital technology. Now, we are in the development and creation of the fourth industrial revolution (i.e., Industry 4.0) revolving around networks of manufacturing resources that are autonomous, capable of controlling themselves in response to different situations, self-configuring, knowledge-based, sensor-equipped, spatially dispersed, and incorporate the relevant planning and management systems (Iansiti and Lakhani, 2014; Lee et al., 2014; Adeyeri et al., 2015). Industry 4.0 relies heavily on the IoT that objects embedded with Information and Communicate Technology (ICT) detected by sensors and transferred by the sensor network (Varghese and Tandur, 2014). Cloud computing is also essential to support the billions of sensors, devices and the flow of information or data that they create (Wang and Dong, 2006). In Industry 4.0 era, powerful software is needed with the capacity to analyze all ICT information (i.e., Big Data) coming from manufacturing systems inside factories or in the whole supply chain in real-time. Through the IoT and the Big Data Analytics, companies can response quickly to business trends and plan better with accurate demand forecasting (Waller and Fawcett, 2013; Yen et al., 2014). The Cyber-Physical System (CPS) in Industry 4.0 is a system featuring a tight combination between the system's computational and physical elements (Jazdi, 2014). The CPS uses computations and communication deeply embedded in and interacting with physical processes to add new capabilities to physical system. Industry 4.0 will be more effective from order to delivery in the supply chain management by implementing in the environment of smart factory (Shrouf et al., 2014; Wang et al., 2016). Figure 1 shows a typical smart factory operation flowchart.

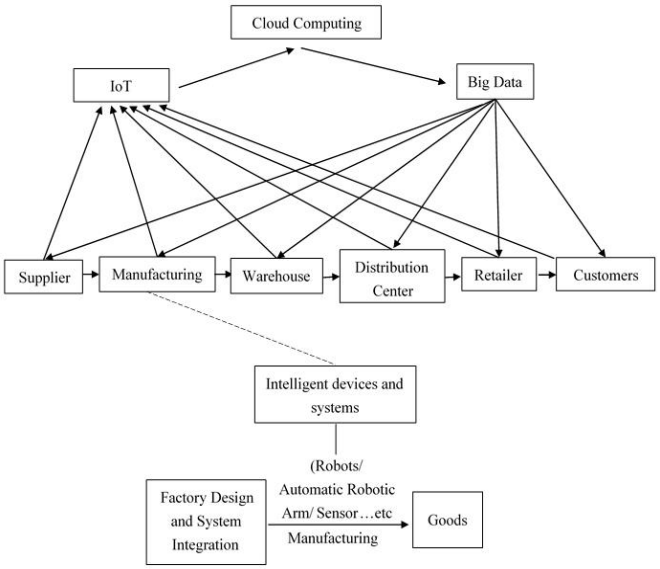

Figure 1: Smart factory operation flowchart.

The Big Data Analytics become popular in last few years, as it represents the hard work of researchers to achieve business intelligence by processing tremendously large amount of data (Parise et al., 2012). Although it is hard to give a precise definition of the Big Data since "big" as a word is fuzzy, the concept of the Big Data is classified as three main characteristics: velocity, volume and variety (3Vs). Velocity defined by the speed with which the data is fetched, processed and returned. Volumes encompassing varied percentage of meaningful data and handling mechanisms. Variety is concerned with the degree of variability of the data. The Big Data mostly has to do with unstructured data. Contrary to popular belief, however, structured data can also be classified as the Big Data and analyzed with Hadoop depending on other features of the data. Readers may refer to other literature for other definitions of the Big Data using 4Vs or 5Vs.

Munar et al. (2014) presented the Big Data architectural and design pattern offering horizontal scalability and no-SQL flexibility while at the same time meeting the stringent quality and resilience requirements of the banking software standards. They aimed at the adoption of these new technologies in the solution of massive data processing and analytics tasks in the financial institution. Holley et al. (2014) propose various aspects of the Big Data and the data models from its initially

defined schema such that data models can easily adapt to changes. Demchenko et al. (2014) discusses a nature of the Big Data that may originate from different scientific, industry and social activity domains and proposes improved Big Data definition that includes the following parts: Big Data properties data models and structures, data analytics, infrastructure and security. Gandomi and Haider (2015) introduced documents in which the basic concepts relating to big data. A key contribution of this research is to bring forth the oft-neglected dimensions of big data.

In this paper, we develop a procedure to detect possible money laundering activities with traditional statistics methods and Big Data Analytics. The paper is organized as follows: section 1 is the introduction and research background. Section 2 describes the methods that we try to implement on the dataset acquired from case company following section 3 for the story of case company briefly. Computational results are provided in section 4 and section 5 is the conclusions.

2. METHODOLOGY

In this section, we present the methods that we applied into the money laundering case study by the Big Data Analytics for both traditional statistics analysis and data mining techniques.

According to the dataset from the case bank, the core of using the R language and MinitabTM software to detection of money laundering activities pattern is shown in Figure 2.

2.2 Forecasting Methods

Fundamental forecasting methods were applied at the beginning of our analysis into dataset in order to fill-in blank or unidentified data fields. We presented some of the methods in this subsection.

2.2.1 Naïve forecasts

Naïve forecasts are the most cost-effective forecasting model, and provide a benchmark against more sophisticated models. This forecasting method is only suitable for time series data with short-term forecast capability. Using the naïve approach, forecasts are produced that are equal to the last observed value. If the time series is believed to have seasonality, seasonal naïve approach may be more appropriate where the forecasts are equal to the value from last season. The naïve method may also use a drift, which will take the last observation plus the average change from the first observation to the last observation. Three naïve forecasts formula are:

(1) Stable time series data:

$$F(t) = A(t-1), \qquad (1)$$

where t is time index, $F(\bullet)$ is the forecast value, and $A(\bullet)$ is the actual value. That is, the stable time series forecast is the same as the last actual observation.

2.1 Money Laundering Detection Flowchart

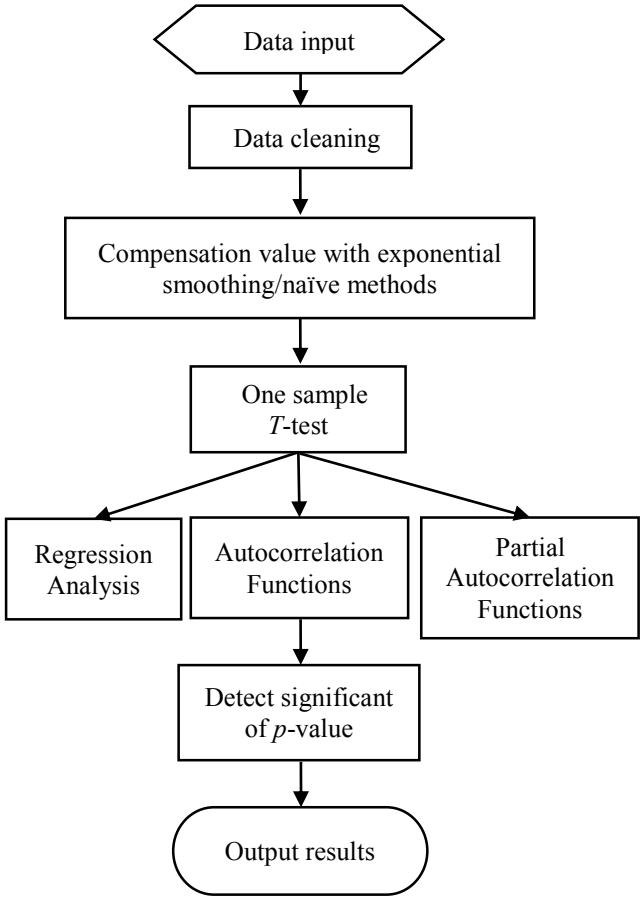

Figure 2: Money laundering detection flowchart.

(2) Seasonal variations:

$$F(t) = A(t-n), \qquad (2)$$

where a cycle lasts n periods. That is, seasonal variations forecast is the same as the last actual observation when we were in the same point in the cycle.

(3) Data with trends:

$$F(t) = A(t-1) + (A(t-1) - A(t-2)). \qquad (3)$$

There is constant trend, the change from $t-2$ to $t-1$ will be exactly as the change from $t-1$ to t.

2.2.2 Simple Exponential Smoothing

Exponential smoothing methods give larger weights to more recent observations, and the weights decrease exponentially as the observations become more distant. These methods are most effective when the parameters describing the time series are changing slowly over time. The Simple Exponential Smoothing method is used for forecasting a time series when there is no trend or seasonal pattern.

The exponential smoothing formula is:

$$F(t) = \alpha A(t) + (1 - \alpha)F(t - 1), \qquad (4)$$

where α is the smoothing constant, $\alpha \in (0, 1)$.

2.2.3 Autocorrelation Function

The autocorrelation function proposed by Box and Jenkins (1970) can be used for the following two purposes:

1. To detect non-randomness in data;
2. To identify an appropriate time series model if the data are not random.

This function is a plot of the autocorrelation as a function of lag. The autocorrelation is simply the ordinary Pearson product moment correlation of a time series with itself at a specified lag. The autocorrelation at lag 0 is the correlation of the series with its unlagged self, or 1. The autocorrelation at lag 1 is the correlation of the series with itself lagged one step; the autocorrelation at lag 2 is the correlation of the series with itself lagged 2 steps; and so forth.

Given measurements, Y_1, Y_2, ..., Y_N at time X_1, X_2, ..., X_N, the lag k autocorrelation function is defined as

$$\rho_k = \frac{\text{covariance at lag } k}{\text{variance}} = \frac{\sum (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum (Y_t - \bar{Y})^2}. \qquad (5)$$

Although the time variable, X, is not used in the formula for autocorrelation, the assumption is that the observations are equi-spaced. Autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times X_i and X_{i+k}.

When the autocorrelation is used to detect non-randomness, it is usually only the first (lag 1) autocorrelation that is of interest. When the autocorrelation is used to identify an appropriate time series model, the autocorrelations are usually plotted for many lags.

2.2.4 Partial Autocorrelation Function

This function is a plot of the partial autocorrelations versus lag. The partial autocorrelation at a given lag is the autocorrelation that is not accounted for by autocorrelations at shorter lags.

The partial autocorrelation coefficient of order k is evaluated by regressing y_t against y_{t-1}, …, y_{t-k}:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_k y_{t-k}. \qquad (6)$$

Then calculate the correlation of the residuals of this regression with y_{t-k}. The partial autocorrelation as the autocorrelation which remains at lag s after the effects of shorter lags (1, 2, ..., k–1) have been removed by regression.

2.2.5 Regression Analysis

Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables. A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.

In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad (7)$$

where:

Y = dependent variable;
X = independent variable;
β_0 = Y-intercept;
β_1 = slope of the line;
ε = error variable.

The magnitude and direction of that relation are given by the slope parameter (β_1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (β_0). An error term (ε) captures the amount of variation not predicted by the slope and intercept terms. The regression coefficient (R^2) shows how well the values fit the data.

3. CASE STUDY

In order to follow the IPP law, we disguised our case study bank as the AM Bank, which is a $40 billion asset regional bank holding company headquartered in New Orleans,

Louisiana. The AM Bank was founded in 1951. At its inception, the AM Bank mainly provided commercial banking services and served the greater Baton Rouge area. Within a decade, the AM Bank extended its operations and started serving small business in the state of Louisiana. Currently, the AM Bank provides consumer banking services, mortgage banking services, equipment leasing, wealth and investment management services, trust services, and brokerage services, as well as other financial products and services. The AM Bank is confident in the Bank's proprietary software that meticulously monitors transactions coming from Cash Intensive Businesses (CIBs), such as casinos and restaurants, as well as from Money Service Businesses (MSBs), such as currency exchanges and money transmitters. It is worried that certain money laundering activities may go unnoticed if the AM Bank's suspicious activity monitoring and detection platform does not undergo material enhancements in the area of predictive analytics.

The AM Bank has been particularly keen on tracking Bank's customers spending habits over time and building profiles for all customers. The majority of the AM Bank customers have predictable transacting habits and it is the sole responsibility of the AM Bank to uncover those transacting patterns by diligently and accurately mining the data that is being automatically collected and fed into the AM Bank's repositories round the clock.

In this case study, a partial database on monthly aggregate incoming wire transfer amounts into the bank account of one of its customers, a U.S. chain of stores selling swimwear was acquired for our research. Based on business knowledge related to swimwear retail, the management of the AM Bank finds it extremely unlikely that a swimwear chain of that size will have an aggregate monthly incoming wire transfer amount exceeding $800,000 while other monthly transfer amount average about $409,740. Hence, through initial analysis, we believe that there should be unusual transactions exist.

4. EXPERIMENT RESULTS

Data cleansing is important at the beginning for any dataset to ensure that the data within a dataset is correct. During this process, records are checked for accuracy and consistency, and they are either corrected or deleted as necessary. A one sample *t*-test for the unusual transactions month is performed following from data cleaning process. Figure 3 shows the Minitab printout from the case data with the *p*-value = 0 at 95% confidence level. That is, there exists unusual transactions in this month.

One-Sample T

Test of μ = 828.171 vs $\neq$ 828.171

N	Mean	StDev	SE Mean	95% CI	T	P
83	409.74	37.05	4.07	(401.65, 417.83)	-102.89	0.000

Figure 3: Minitab printout of one sample *t*-test for the unusual transactions month.

Then, we used exponential smoothing and naïve forecasts to compensation values missing in the dataset with R software.

The dataset is classified into three categories:

1. Original Data (label as "Original");
2. Inward Remittance (label as "In");
3. Outward Remittance (label as "Out").

Next, we use the following three methods to detection of money laundering activities:

1. Regression Analysis;
2. Autocorrelation Function;
3. Partial Autocorrelation Function.

4.1 Regression Analysis

We used regression analysis to identify the relationship between a dependent variable and independent variables whether there were significant of *p*-value. We defined dependent variable is transaction amount and independent variable is time. We divided into four categories of individual data from classified data:

1. No category (label as "All");
2. Range of transaction amount (label as "Amount Range");
3. Week;
4. Day.

Table 1: Transactions analysis by weeks and by day of week from regression analysis.

	All		Amount Range		Week		Day	
Original	Unclassified	0.070	Low	0.137	W1	0.321	MO	0.939
							TU	0.633
	IgnoreWE	0.226	Medial	0.236	W2	0.906	WE	0.043
							TH	0.690
	IgnoreFR	0.216	High	0.317	W3	0.886	FR	0.048
							SA	0.403
	IgnoreWE& FR	0.575			W4	0.902	SU	0.142
In	Unclassified	0.033	Low	0.540	W1	0.926	MO	0.771
							TU	0.437
	IgnoreWE	0.107	Medial	0.411	W2	0.909	WE	0.053
							TH	0.568
	IgnoreFR	0.136	High	0.803	W3	0.545	FR	0.025
							SA	0.386
	IgnoreWE& FR	0.371			W4	0.562	SU	0.056
Out	Unclassified	0.514	Low	0.592	W1	0.105	MO	0.674
							TU	0.328
	IgnoreWE	0.189	Medial	0.103	W2	0.909	WE	0.309
							TH	0.928
	IgnoreFR	0.432	High	0.755	W3	0.603	FR	0.763
							SA	0.595
	IgnoreWE& FR	0.130			W4	0.035	SU	0.400

From Table 1 we see that most of the *p*-value are not significant, but especially it approached significant of *p*-value in Wednesday and Friday of inward remittance at the 95% confidence level. Then, we deleted (ignored intentionally) Wednesday, Friday or both of them to observed significant degree of *p*-value from categorized data. But the results show insignificant of *p*-value. From this step, we acquired limited information about the transactions in this month.

4.2 ACF and PACF

We observed plots of autocorrelation functions and partial autocorrelation functions to understand what type of categorized data, and then judged the characteristic of data.

We also divided into four categories of individual data from classified data:

1. No category (label as "All");
2. Range of transaction amount (label as "Amount Range");
3. Week;
4. Day.

Table 2: Transactions analysis by weeks and by day of week from ACF and PACF.

	All		Amount Range		Week		Day	
Original	Unclassified	None	Low	Trend	W1	None	MO	None
							TU	None
	IgnoreWE	None	Medial	Trend	W2	None	WE	None
							TH	None
	IgnoreFR	None	High	Trend	W3	None	FR	None
							SA	None
	IgnoreWE& FR	None			W4	None	SU	None
In	Unclassified	None	Low	None	W1	None	MO	None
							TU	None
	IgnoreWE	None	Medial	None	W2	None	WE	None
							TH	None
	IgnoreFR	None	High	None	W3	None	FR	None
							SA	None
	IgnoreWE& FR	None			W4	None	SU	None
Out	Unclassified	Random	Low	Trend	W1	Random	MO	Random
							TU	Random
	IgnoreWE	Random	Medial	Trend	W2	Random	WE	Random
							TH	None
	IgnoreFR	Random	High	Random	W3	Random	FR	Random
							SA	None
	IgnoreWE& FR	Random			W4	Random	SU	Random

In Table 2, we noted that "None" meant no pattern, "Trend" meant tendency pattern of data, "Random" meant disorder pattern of data. The majority were no pattern characteristic of data with the exception of there were appearing tendency pattern which were understandable in "amount range." Notable exception is random pattern in outward remittance. This represented unusual transactions in this period of time (i.e., this unusual month).

5. CONCLUSIONS

In this paper, we applied the Big Data Analytics methods to detect possible money laundering activities at a particular time. After applying the Big Data Analytics, we discovered high degree significant of *p*-value in Wednesday and Friday of inward remittance at the 95% confidence level with regression analysis and there is random pattern in outward remittance with autocorrelation functions and partial autocorrelation functions. Hence, we believed that most probably money laundering times in this month happened in Wednesday and Friday of inward remittance and then transfer to outward remittance disorderly. Further techniques are required in order to discover more information about this dataset. A functional AML detection system may be able to design to detect unusual activities in real-time.

REFERENCES

Adeyeri, M.K., Mpofu, K., and Olukorede, A.T. (2015) Integration of agent technology into manufacturing enterprise: A review and platform for industry 4.0. *The 2015 International Conference on Industrial Engineering and Operations Management*, 1-10.

Box, G. and Jenkins, G. (1970) *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Demchenko, Y., De Laat, C., and Membrey, P. (2014) Defining architecture components of the big data ecosystem. *The 2014 International Conference on Collaboration Technologies and Systems*, 104-112.

Gandomi, A. and Haider, M. (2015) Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, **35**(2), 137-144.

Gao, Z. (2009) Application of cluster-based local outlier factor algorithm in anti-money laundering. *The 2009 International Conference on Management and Service Science*, 1-4.

Gao, Z.-A. and Weng L.-F. (2006) Transfer price-based money laundering in international trade. *The 2006 International Conference on Management Science and Engineering*, 1128-1132.

Holley, K., Sivakumar, G., and Kannan, K. (2014) Enrichment patterns for big data. *The 2014 IEEE International Congress on Big Data*, 796-799.

Hong, X., Liang, H., Cai, L.X., Gao, Z., and Sun, L. (2015) Peer to peer anti-money laundering resource allocation based on semi-markov decision process. *The 2015 IEEE Global Communications Conference*, 1-6.

Iansiti, M. and Lakhani, K.R. (2014) Digital ubiquity: How connections, sensors, and data are revolutionizing business. *Harvard Business Review*, **92**(11), 91-99.

Jazdi, N. (2014) Cyber physical systems in the context of Industry 4.0. *The 2014 IEEE International Conference on Automation, Quality and Testing, Robotics*, 1-4.

Khac, N.A.L. and Kechadi, M-T. (2010) Application of data mining for anti-money laundering detection: A case study. *The 2010 IEEE International Conference on Data Mining Workshops*, 577-584.

Lee, J., Kao, H.-A., and Yang, S. (2014) Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia CIRP*, **16**, 3-8.

Liu, R., Qian, X.-L., Mao, S., and Zhu, S.-Z. (2011) Research on anti-money laundering based on core decision tree algorithm. *The 2011 Chinese Control and Decision Conference*, 4322-4325.

Liu X. and Zhang, P. (2007) An agent based anti-money laundering system architecture for financial supervision. *The 2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 5472-5475.

Lv, L.-T., Ji, N., and Zhang, J.-L. (2008) A RBF neural network model for anti-money laundering. *The 2008 International Conference on Wavelet Analysis and Pattern Recognition*, **1**, 209-215.

Munar, A., Chiner, E., and Sales, I. (2014) A big data financial information management architecture for global banking. *The 2014 International Conference on Future Internet of Things and Cloud*, 385-388.

Parise, S., Iyer, B., and Vesset, D. (2012) Four strategies to capture and create value from big data. *Ivey Business Journal*, **76**(4), 1-5.

Shrouf, F., Ordieres, J., and Miragliotta, G. (2014) Smart factories in industry 4.0: a review of the concept and of energy management approached in production based on the Internet of things paradigm. *The 2014 IEEE International Conference on Industrial Engineering and Engineering Management*, 697-701.

Stock, T. and Seliger, G. (2016) Opportunities of sustainable manufacturing in Industry 4.0. *Procedia CIRP*, **40**, 536-541.

Varghese, A. and Tandur, D. (2014) Wireless requirements and challenges in Industry 4.0. *The 2014 International Conference on Contemporary Computing and Informatics*, 634-638.

Waller, M.A. and Fawcett, S.E. (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, **34**(2), 77-84.

Wang, S., Wan, J., Zhang, D., Li, D., and Zhang, C. (2016) Towards smart factory for Industry 4.0: A self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, 101, 158-168.

Wang, X. and Dong, G. (2009) Research on money laundering detection based on improved minimum spanning tree clustering and its application. *The 2009 Second International Symposium on Knowledge Acquisition and Modeling*, 62-64.

Yen, C.-T., Liu, Y.-C., Lin, C.-C., Kao, C.-C., Wang, W.-B., and Hsu, Y.-R. (2014) Advanced manufacturing solution to industry 4.0 trend through sensing network and Cloud Computing technologies. *The 2014 IEEE International Conference on Automation Science and Engineering*, 1150-1152.