BIG DATA ANALYSIS THROUGH THE LENS OF BUSINESS INTELLIGENCE – WORLD CONFLICT INCIDENTS CASE STUDY (1989-2016)

Adrian BARBU [*]

Tudor RAȚ [**]

Abstract

Data analysis as a process and the specialized tools exploited in this regard, represent a powerful "weapon" available to anyone who wants to explore datasets, even for personal goals. The top business companies around the world are highly linked and dependent on the outputs revealed by the experts who manipulate various types of data on two coordinates: development and improvement; limitation of risks and vulnerabilities. Data analysis is not the perquisite of business branch; it is also intensively used by the intelligence agencies, mostly in areas like SIGINT or OSINT. Nowadays, data analysts are required to handle an enormous amount of data, which in many cases represents the biggest challenge for them and inferentially for the analytics. The technological evolution entails a capacity of rapid reaction to continuous transformation of data flows and the ability to conserve an accurate manner of sense-making in order to be able to provide useful data-based intelligence.

Keywords: *data, analysis, intelligence, interactive, visuals, forecast*

What is data analysis?

A report developed by Forbes Insights and Ernst & Young, shows that almost two-thirds of companies[1], which have well-developed analytics strategies and are using advanced tools of data analysis raised their profits by 15% in 2016. Technology and telecommunications are the two industries that are using with prepostency this type of tools. Companies and organizations that based their business activity on data analysis strategies, reveal that the

[*] Expert from Romanian Intelligence Service
[**] Expert from Romanian Intelligence Service
[1] 1,518 companies across a range of industries

ability of those who extract, manipulate and interpret datasets improved the competitiveness of the organization (Forbes Insights, 2017).

The advent of technology and the enormous amount of data which flows through the IT&C infrastructure requires specific methods, techniques and tools to capitalize collected data. The new era that provides a huge flow of various types of data advanced innovator concepts of analysing smaller or bigger datasets and shows us that we can use data as a strategic asset. This kind of approach supports the management department in identifying trends, opportunities, vulnerabilities and risks.

The business area views data analysis as a systematic operation of processing and shaping raw data, to sight out evaluations and to carry out logical inferences in order to frame analytical conclusions.

Data analysis is based on statistics and its focus relies on the relationships between variables. It is a widely used method in many fields (economy, marketing, sociology, intelligence etc.) because it delivers long-range insights about a specific problem taking in account large datasets.

In terms of intelligence activity, data analysis is mostly viewed as a sub-component of intelligence analysis and it refers to applying a set of cognitive methods to assess data and to test hypotheses. The result can be used as self-contained analytic product or it can be embodied in a comprehensive analytical framework.

Working with data implies both advantages and disadvantages. One of the most important benefits consists in summarizing large quantities of information, visualizing them in a graphic format and extracting the informational crux. Besides the measurements and the statistical feature, it is very important to use the proper visual tools to facilitate the assessment of the output data, because for the human brain is easier to understand the whole quantity of information presented in a chart or diagram. When the conclusions are extracted on a visual basis, the process will evolve fluently (Chen et al., 2011, p. 85).

Data analysis could also be productive in: identifying connections between entities, phenomena and processes; identifying typologies of entities, phenomena and processes; assessing and predicting the evolution of entities, phenomena and processes which were analysed; validation/ invalidation of various hypotheses; supporting the decision act by delivering information in a timely manner; identifying vulnerabilities or dysfunctions (Eising, December 1, 2010).

Dealing with a colossal amount of data sweeps out various challenges for those who manipulate datasets. One of the biggest problems in managing the datasets in a proper way, in order to analyse them, is related to the

foregoing step of analysis - the processing. It is a time-consuming activity, and the analysts should have the capacity to carry out the trimming, refinement and structuring of datasets in a timely manner. Another inconvenience is related to difficulties of interpreting numerical data, especially because the analysts have to think in a comprehensive way and correlate the numerical data with different other complex facets of the subject analysed.

Big Data and its 3 Vs

Data is generated with a high rate of velocity in every moment. All the processes made *via* the IT systems produce mostly unstructured data and the analysts must possess a complete set of skills to be able to manage these disparate flows, in order to reduce or to cut out the noise. By "Big Data" we should understand a complex concept which encompasses the techniques to capture, process, analyse and visualize mammoth datasets in a reasonable timeframe which cannot be achieved through standard IT technologies (Mukherjee, Shaw, 2016, p. 2).

According to Douglas Laney[2], the concept of "Big Data" is defined by three major elements:

- volume – collected data originates from a variety of sources (business transactions, social media, machine-to-machine data);
- velocity – large amount of data streams with higher and higher speed and must be dealt in a timely manner;
- variety – input data is collected in all types of formats (text, video, photo, audio in unstructured, semi-structured or structured databases) (SAS Analytical Solutions, 2017).

Big Data analysis, as well as the analysis of smaller datasets can reveal descriptive, predictive and/ or prescriptive insights about entities, phenomena or processes. Each type of outlook can be summarized in a question which defines the main purpose of the insight:

1. descriptive – "what happened in the past?";
2. predictive – "what might happen next?";
3. prescriptive – "how do I deal with this?" (Su, 2017, pp. 5-6).

The mentioned analysis methods were developed and used long time before the concept of Big Data. Although, the characteristics of the "new era of data" have shown the usability of analytics in delivering fast and actionable insights for the decision-makers providing both an early-warning tool and a well-developed instrument for the stakeholders.

[2] VP and distinguished analyst, chief data officer at Gartner

Cognitive biases in data analysis

Regarding the role of cognition in data analysis John Wilder Tukey stated: "The basic general intent of data analysis is to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognisable to the data analyser. [...] At all stages of data analysis the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use and want it" (Wilder Tukey,1962, pp. 1-67).

All types of analyses rely on the ability of human mind to interpret and assign meaning to the collected and processed data, usually in accordance with a specific context or goal.

Foregoing the analytical process, a logical plan and a sense making schema should be developed, in order to avoid or to limit the cognitive biases. Commonly, there are four main types of biases that darken the analyst's judgement during the operation with large datasets.

The first one is the confirmation bias which refers to the need of proving a hypothesis. The analyst tends to lean on data that might certify the initial assumption. In this case, the full picture of the scenario might be left out, because the analysed data does not represent the relevant information (Smyth, July, 2017).

The second main type is the sampling bias, also known as selection bias. Occasionally, the available timeframe for the requested analysis is limited and one needs to extract a sample from the entire amount of data. The sampling procedure is usually achieved by applying a range of statistical techniques and a well-designed randomization. Errors may appear if the sample is not representative for the whole dataset that will be analysed (Crawford, Kate, April 1st, 2013).

The third one is related to the illusory correlations made between data variables. This bias occurs when a correlation between two variables is detected when no relationship actually exists. Also, this could evolve in connecting a cause to an effect even when there is no link between them (Grisanti, April 30, 2015).

Last but not least, the so-called *apophenia* is the tendency of humans to see patterns in randomness. This cognitive bias, similar to the previous one, may occur when people identify connections between meaningless data, usually due to the lack of expertise and experience (Grisanti, April 30, 2015).

The human resource cannot be removed from the process of analysis, hence the cognitive biases will linger and the analysts need to be aware of the errors that might shadow the output. In order to thwart blundering interferences in the judgment process, there are a couple of actions that could

limit the cognitive biases: collect data from as many points as possible, even if they seem irrelevant at first glance; create clear and structured datasets, free of subjectivity; maintain an impartial behaviour in manipulating datasets.

Likewise, analysing Big Data by using business intelligence (BI) specialized tools represents a proper way to avoid cognitive biases. This type of technological solution does not deliver only descriptive information based on the available datasets; it also uses strong mathematical algorithms which can provide accurate forecasts useful in the process of testing hypotheses.

Case study – conflict incidents 1989-2016

As was mentioned above, data analysis is a valuable asset in pointing out patterns and future developments of an event or phenomenon. Armed conflicts have always been one of the biggest threats for the national, regional and global security. Even more, nowadays the governments are facing multiple issues regarding conflict prevention and conflict management. Wars are now worn mostly between state and non-state actors and the strategies, tools, and tactics used by rebels, paramilitary groups, terrorist or any other non-state entities have changed in such way that classic counter-measures have become useless.

Using *Tableau*, a business analytics service and interactive data visualization tool[3], a scientifically validated dataset[4] comprising the armed conflicts between 1989 and 2016 (excluding Syria) and the associated variables were explored in order to develop relevant analytic conclusions that might help in subsequent investigations on this subject (UCDP Conflict Encyclopedia).

The focus of the analysis was on two main components:

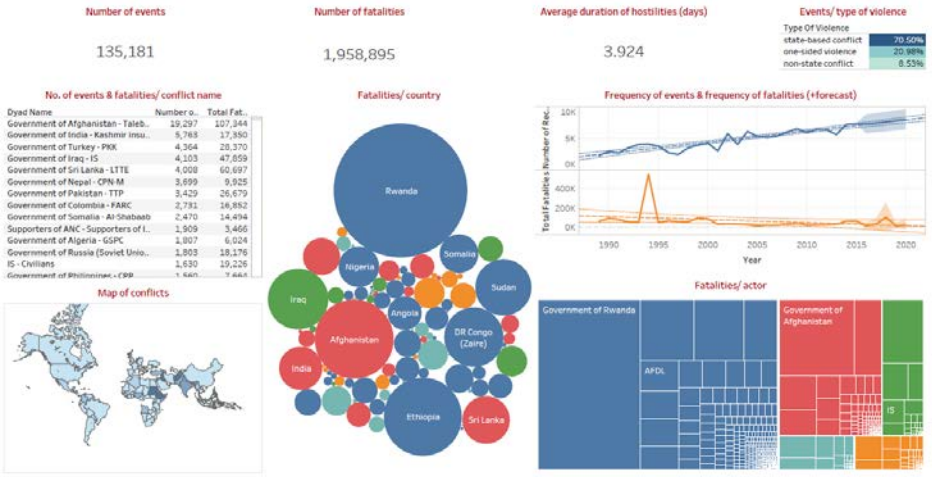
- number of events and fatalities in terms of country, region and actors;
- frequency of events and fatalities and possible evolutions.

[3] Tableau puts analytics in the hands of the user. By enabling individual creativity and exploration from the ground floor, businesses now the ability to adapt and outperform competition through intuitive data visualization and analysis. Tableau can connect to virtually any data source, be it corporate data warehouse, Microsoft Excel or web-based data. It gives users immediate insights by transforming their data into beautiful, interactive visualizations in a matter of seconds. What took expensive teams days or months to develop now is achieved through the use of a user-friendly drag-and-drop interface.

[4] The UCDP Conflict Encyclopedia (UCDP database) is an online, free of charge, database. It is updated and revised several times per year and contains detailed descriptive information on armed conflicts, peace agreements, and several other aspects of organised violence. Coverage is global with information from 1946 and onwards.

One of the issues that we need to be aware of from the beginning is that we deal with a heterogeneous dataset. The standard deviation of the variables shows that the values are spread out and characterized by irregularity. This means that values associated with various conflicts, like fatalities and incidents are far-off from the average. The best example is that of the Rwandan genocide.

As we could notice, in a period of 27 years, there were 135,181 events[5] that generated 1,958,895 fatalities among the belligerents and also civilian population. 70,50% of the incidents were state-based conflicts, 20,98% one-sided violence and 8,53% non-state conflicts.



The conflict with the highest number of incidents was the Afghan Civil War (19,297), followed by the insurgency in the Kashmir region (5,763), Turkish Government - PKK conflict (4,364) and the conflict between the Iraqi Government and Daesh (4,103). However, the most violent conflict, in terms of fatalities, was the Rwandan genocide (511,531), followed by Afghan civil war (107,344), Eritrean-Ethiopian war (97,435), and Sri Lankan civil war (60,697).

On one hand, we can remark that the highest rate of fatalities has occurred in the African region, particularly in Rwanda, Ethiopia, Democratic Republic of Congo, Sudan, Nigeria and Somalia. On the other hand, the highest

[5] An event is defined by UCDP as an individual incident of lethal violence occurring at a given time and place.

rate of events is related to the Middle East and Far East conflicts, especially in Afghanistan, India, Nepal, Pakistan, Sri Lanka and the Philippines.

Last year (Jan-Dec 2016), most of the 53,954 victims (approximate 70%) were caused by the conflicts between terrorist groups/ Islamic fundamentalists and different governmental forces, with prepostency in Middle East and North Africa.

In terms of non-state conflicts, the hostilities were the most active in South Africa, between the supporters of Inkatha Freedom Party and the supporters of African National Congress[6]. There is a big variety of actors related to this type of conflict, from partisans of different political parties or political movements[7], organized crime groups[8] and various ethnic or religious communities, to terrorist and radical Islamic organizations[9]. It is also noticeable the disposal of the type of actors by continents and regions: supporters of opposing political organizations in Africa, organized crime groups and urban guerrillas in Latin America and terrorist groups in Middle East and North Africa.

Conclusions

Visualising and analysing in an interactive manner this type of dataset, with more than 130,000 entries, revealed useful insights regarding interstate and intrastate conflicts. The complex options of filtering, cross-filtering, easy capabilities of exploring and mathematical algorithms available in the BI tool allowed pattern identification and trend detection. The outputs could serve as a basis for future assessments on the evolution of specific violent incidents or the development of the whole phenomenon.

On the African continent, armed conflicts are more violent than in any other part of the world, even if there were fewer events, compared to Middle East and North Africa where the number of incidents was the highest. This is an essential mark of the characteristics of conflicts depending on the source region and it could be also useful in order to understand the particularities of the societies involved.

As non-state actors, the South American criminal organizations, paramilitary groups and the Islamic jihadists produced over 9 out of 10

---

[6] 1,909 incidents; the second non-state conflict with the highest rate of events was the conflict between Juarez Cartel and Sinaloa Cartel

[7] e.g. United Democratic Front (India), National Socialist Council of Nagaland, Muhajir Quami Movement (Pakistan)

[8] e.g. Gulf Cartel, Sinaloa Cartel, Los Zetas, Tijuana Cartel

[9] e.g. Hezbollah, Al-Qaeda in Arabic Peninsula, Daesh

victims. The forecasted trend for this type of violence is heavily ascending. If we refer to ongoing state-based conflicts, the Afghan civil war is still the most "prolific" in terms of fatalities and number of incidents.

Regarding the frequency of incidents in the last years, after a narrowing trend from 2011 to 2013, we can observe a continuous escalation. Furthermore, the forecast tool predicts that the level will increase lineal at least until 2020.

With regard to the frequency of fatalities, there was a slow increase from 2013 to 2014, albeit the trend-line is descending since 1995. The number of victims will grow till 2018, followed by a decrease in 2019 and a slow increase in 2020.

References:

1. Chen, Min, Trefethen, Anne, Banares-Alcantara, Rene , Jirotka, Marina, Coecke, Bob, (2011), *From Data Analysis and Visualization to Casuality*, Washington: IEEE Computer Society.

2. Crawford, Kate, (April 1st, 2013), *The Hidden Biases in Big Data*, accessed 27 August 2017, https://hbr.org/2013/04/the-hidden-biases-in-big-data.

3. Eising, Martin, (December 1, 2010), *Data Analysis Overview*, accessed 15 August 2017, http://www.dashboardinsight.com/.

4. Forbes Insights, (2017), *Data & Advanced Analytics: High Stakes, High Rewards*, accessed 5 September 2017, https://insights.forbes.com/advanced-analytics-high-stakes-high-rewards/?aliId=88748382.

5. Grisanti, Julie, (2017), *Trust the Data: How to Counteract Human Cognitive Biases*, accessed 27 August 2017, http://www.aunalytics.com/trust-the-data/.

6. Mukherjee, Samiddha, Shaw, Ravi, (2016), *Big Data – Concepts, Applications, Challenges and Future Scope*, International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, February.

7. SAS Analytical Solutions, (2017), *Big Data – What it is and why it matters*, accessed 6 September 2017, https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

8. Smyth, Daniel, (July, 2017), *Four cognitive biases that affect big data analysis*, accessed 27 August 2017, http://bigdata-madesimple.com/four-cognitive-biases-that-affect-big-data-analysis/.

9. Su, Xiaomeng, (2017), *Introduction to Big Data*, Norwegian University of Science and Technology, Trondheim.

10. UCDP Conflict Encyclopedia (UCDP database): www.ucdp.uu.se, Uppsala University.

11. Wilder Tukey, John, (1962), *The future of data analysis*, The Annals of Mathematical Statistics, Vol. 33, pp. 1-67.