ANALIZA DE INTELLIGENCE

EXPLORATORY INTELLIGENCE ANALISYS

Davide BARBIERI*

Abstract

Traditional data analysis and intelligence analysis are strongly focused on testing hypotheses, which can be either corroborated or rejected by means of collected observations. Still, we must acknowledge that many hypotheses are just assumptions or prejudices, which are unlikely to be supported by evidence. Researchers may therefore select the observations which comply with their ideas, in an attempt to prove they were right in making assumptions. To avoid this pitfall, analysts should not neglect the preliminary steps of exploratory data analysis, by means of which new – and possibly more robust – hypotheses are explored, diminishing the constraints imposed on the analysis by the investigators' creativity. Several descriptive and graphical techniques can be employed by intelligence analysts in order to succeed in this endeavor. This paper will give a summary account of some of them.

Keywords: *hypotheses finding, data exploration, descriptive techniques.*

Introduction

"Reality, contrary to fiction, does not need to be realistic".
L. Pirandello

In its simplest form, the intelligence cycle is a five-step process: planning, collection, processing, analysis and dissemination[1]. These steps overlap with those of traditional data analysis: state a problem, collect data to study it, process and analyze data, report results. Both of them are specific example of the knowledge production process as described by Popper (2002a, 2002b): problem 1 → hypothesis → test → problem 2... This is a purely deductive process: to address a problem, a tentative solution (hypothesis) is tested. Until it stands the challenge of evidence, the hypothesis is temporarily

* *Profesor, Universitatea din Ferrara, Italia.*
[1] *See for example https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books- and-monographs/analytic-culture-in-the-u-s-intelligence-community/chapter_4_systems_model.htm*

accepted. Otherwise it is rejected. Then, a new problem will arise, possibly more difficult, which will require a more creative solution, thus generating a virtually infinite sequence[2].

The discovery of antibiotics is a good example: infections caused by bacteria were the problem. Some tentative solutions were tested and soon rejected because there was no evidence of their effect. These outcomes redefined the problem as more difficult. Penicillin was then tested and – since it proved to be effective – it was accepted as a solution to infections. But new problems have arisen, like side effects and antibiotic-resistant bacteria.

According to Popper, no induction process is possible: hypotheses are always *a-priori* and scientists can only test them against evidence. They cannot be *verified* (proved to be true) strictly speaking, but just corroborated by means of the collected data, or rejected. Observations should be selected in order to challenge the current hypothesis, and data are not used to suggest new hypotheses.

This methodological framework has some limitations, including the scientist's creativity and a high risk for confirmation and selection bias. In fact, investigators have the tendency to collect data in order to confirm their hypotheses (Evans 1989, Nickerson 1998) rather than to challenge them. This bias could be avoided if the scientists only tried to counter (*falsify*, in Popper's terms) their hypotheses, which is not always the case[3].

Exploratory data analysis

John Tukey (1915 – 2000) was aware of the fact that traditional statistical analysis was mainly focused on hypotheses testing (*confirmatory data analysis*, CDA) rather than hypotheses finding. To overcome this pitfall, he suggested the adoption of *exploratory data analysis* (EDA, Tukey 1977). In his own words, EDA is a "detective work" and as such may bear several similarities with intelligence analysis. In this paper, I shall first describe the main features of EDA and then show some examples of how it can be applied to intelligence in order to elicit new and possibly more robust hypotheses.

Tukey asserted that EDA is essentially an attitude rather than a set of techniques. The aim of EDA is to find patterns and make tentative hypotheses, which will be later tested by means of CDA (Beherens 1997). In this sense, EDA can be considered a precursor of data mining, which is usually defined as "the process of discovering patterns in data" (Witten and Frank 2005, pp. 5, 9).

The American statistician suggested an extensive adoption of graphical methods, like bar charts, scatter plots and box plots. In particular, he

---

[2] *Popper refused to use the term cycle, since problems are always new.*

[3] *According to the Austrian philosopher, a theory can be considered scientific only if it can be falsified (Popper 2002a) – at least in principle – by means of some collected evidence.*

supported the use of five-number summaries (minimum, 1st quartile, median, 3rd quartile and maximum) for numerical variables. Such summaries make (almost) no assumptions[4] about the distribution of data and can be easily represented in graphical form by means of box plots.

As in data mining, the aim of which is to explore new and possibly counter-intuitive hypotheses, the value of a chart is greater if it allows you to see something you have never expected to see, thus overcoming the limits of the scientists' creativity and intuition. The amount of unnecessary and unwarranted assumptions in EDA is low, and thus it is more *data-driven* than *hypothesis-driven* as CDA. Such an exploratory approach will probably produce more robust hypotheses, which will stand the challenge of evidence, and therefore diminish the amount of necessary trials and subsequent errors.

Tukey foresaw the advent of *data science*: the application of both EDA and CDA to large amounts of data (the so called *big data*). Data science comprises data mining, especially in its exploratory phase, and traditional statistics, especially in its confirmatory phase. The two disciplines tend to overlap and complement each other so much that nowadays it is common to refer to both as *statistical learning*[5]. The data science cycle can be synthetically represented as in Figure 1.

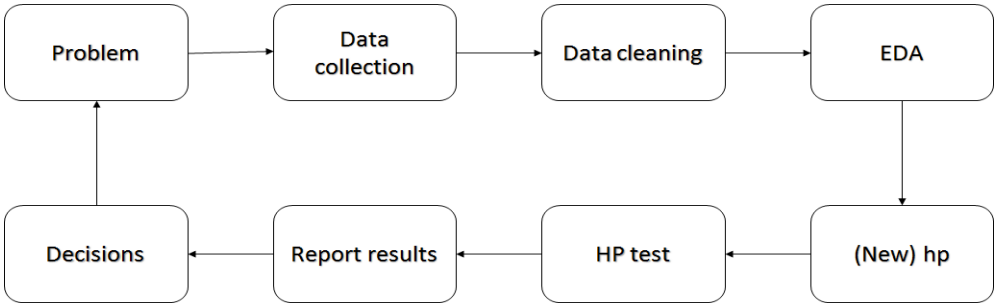

Figure 1. The data science cycle.

The steps from the definition of the problem to EDA are concerned with hypothesis finding, while the following steps are concerned with hypothesis testing, till a final decision is taken (the main aim of data science is

[4] *Some assumptions are needed in order to define outliers: usually, they are considered values lower than Q1-1,5*IQR or greater that Q3+1,5*IQR, where Q1 is the first quartile, Q3 the third quartile and IQR=Q3-Q1 is the interquartile range. Thus, there is no easy way out from a deductive framework, as Popper has always maintained.*

[5] *Data mining applies machine learning algorithms to large datasets. For an introduction see James et al. (2015).*

decision support). Comparing this cycle to that of intelligence, we can say that data cleaning and preparation are performed in the processing step, while EDA is performed at the beginning of the analysis step.

Since attention is selective (and we cannot collect all possible data), a clear understanding of the problem at hand is always needed. Therefore, some domain specific knowledge (i.e. hypotheses, assumptions) is necessary. Data scientists are not autonomous in their task, but they need the collaboration of other scientists, like biologists, medical doctors or intelligence and security experts, according to the nature of the problem to be investigated. Usually though, EDA improves the understanding of the problem and increases domain specific knowledge. Also, when large datasets are available, a lot of data cleaning and pre-processing is needed, in order to get rid of unwanted noise and to prepare data for subsequent analysis.

Hypothesis testing has often been described as a trial, where the null or starting hypothesis (innocence) is supported by the evidence brought to court by the lawyer, and challenged by the evidence brought by the prosecutor, who endorses the alternative or competing hypothesis (guilt). The goal of CDA is to support the final decision (i.e. the verdict). EDA instead can be likened to indictment.

EDA and its applications to intelligence analysis

In order to describe a possible application of EDA to intelligence analysis, the RAND database, openly available to all researchers, was downloaded from the corporation's website[6]. The database contains a list of all terrorist attacks worldwide since 1968. A subset including 10 years, from 2000 to 2009, of attacks in Italy was selected. A total of 141 attacks was included. Each record in the dataset, corresponding to one observation (i.e. a terrorist attack) was described by means of the following fields (attributes): date, city, perpetrator, weapon, injuries and fatalities. MS Excel was used to store, clean and analyze the data. Dates had to be formatted according to a unique standard.

Bar charts and Pareto charts

A first analysis displays the distribution, in order of diminishing frequency, of terrorist attacks per perpetrator. All groups were aggregated according to the following classification: anarchists, communists, pro-Palestinians, autonomists and others. In case the attack was not claimed or the terrorist group was not identified by investigators, the perpetrator was labeled *unknown*. A cumulative frequency curve (*ogive*) was added to the bar chart, as in Pareto charts. Results are shown in Figure 2.

[6] *http://www.rand.org/nsrd/projects/terrorism-incidents/download.html*

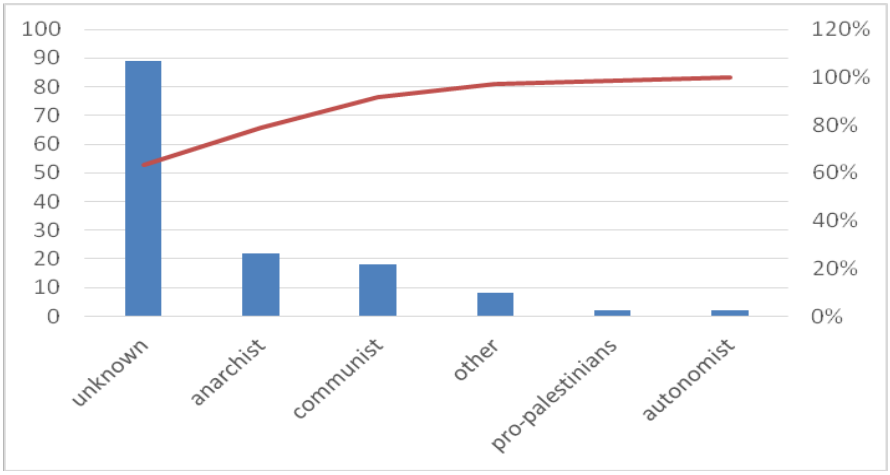

Figure 2. Terrorist attacks by perpetrator.

Absolute frequencies (counts) are displayed on the left y-axis, while cumulative frequencies are displayed on the right y-axis. Pareto charts are an important decision support tool, which allow policy makers to allocate wisely their limited resources (time, personnel, money, technology etc.). In this case, no matter how many resources were engaged to monitor political extremists, the majority of attacks (63%) was not claimed. Probably, more resources should be allocated to preliminary investigations.

Scatter plots and correlations

A second exploratory analysis tries to correlate[7] the population of 10 cities and the number of terrorist attacks. Data are shown in Table 1.

City	Population (k)	Attacks	Relative freq.
Milan	1,345	26	31%
Rome	2,864	18	22%
Genoa	586	7	8%
Turin	890	7	8%
Bologna	386	6	7%
Cagliari	154	5	6%
Florence	382	5	6%

[7] *For an introduction to correlation and regression, see Mann 2010, pp. 565-623.*

Pisa	89	3	4%
Treviso	83	3	4%
Viterbo	67	3	4%
Total		83	

Table 1. First 10 cities by number of terrorist attacks.

This analysis can be performed graphically by means of a scatter plot and a fitted line, as in Figure 3. Population is shown on the x-axis while the number of attacks on the y-axis.

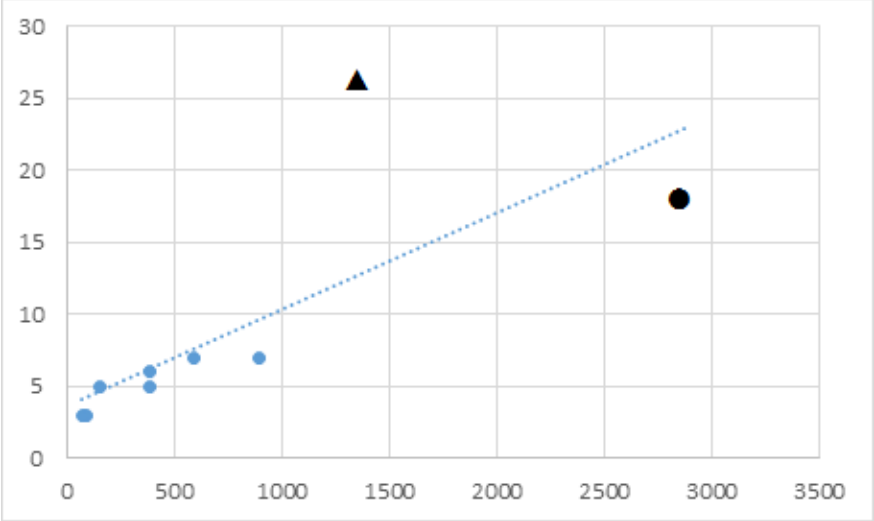

Figure 3. Correlation between population and number of attacks.

There seems to be a linear correlation between the population and the number of terrorist attacks. Still, the number of attacks in Milan (▲) is evidently higher than expected. Rome (●) is instead in line with other smaller cities. This fact was not immediately evident from the database itself.

Time series analysis

Time series analysis[8] is not a typical EDA technique, but it can be used effectively to explore the data. In the following example (Figure 4), attacks are aggregated by year (segmented line; it must be considered that a terrorist

[8] *For an introduction see Brockwell and Davis (2002).*

attack is a fairly rare – even if tragic – event). Moving average[9] (n=3) is shown by means of a smoothed line, and the trend as a dotted line. It is evident – regardless of the common perception – that after a peak in the early 2000s, attacks were diminishing, at least until 2008.

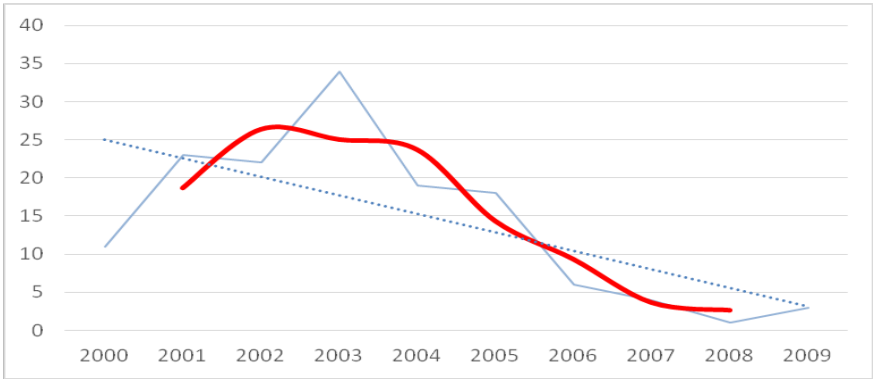

Figure 4. Time series of terrorist attacks from year 2000 to 2009.

Time series analysis can eventually find some periodicity in the sequence of events. Periodicity can be leveraged for prediction, risk assessment, and to optimize resource allocation (Barbieri 2016).

Spurious correlations

Especially when exploring large amounts of data, the probability to find some *spurious* (i.e. non-causal) correlations is very high[10]. As it is often stated by statisticians, *correlation is not causation*. The aim of science is to *explain* (i.e. to find cause-and-effect relationships) the observed phenomenon, be it terrorism, finance, or epidemiology. Nonetheless, this is not necessarily the main purpose of intelligence analysis.

While allowing scientists (criminologists or psychiatrists, economists, medical doctors etc.) to find an explanation for the extrapolated patterns, intelligence agencies can immediately leverage the newly acquired information to further their investigative activities. Similarly, marketing experts use the correlations found in collected data to improve their business, by means, for example, of cross- and up-selling.

On the one hand, not all correlations are useful to understand and explain a phenomenon, on the other not all correlations which appear

[9] *The mean of n data around the central value.*
[10] *For some examples see: http://tylervigen.com/spurious-correlations*

counter-intuitive or contrary to established scientific knowledge should be rejected. Real progress usually goes against the traditional explanations and the history of science is scattered of such cases.

Conclusions

The techniques described in this paper can be used by intelligence analysts to explore and investigate their data prior to establishing the competing hypotheses to be tested. Most of the conjectures which may be done after the adoption of some graphic and descriptive techniques could not be envisaged or assessed by simply looking at the hundreds (or thousands, in many cases) lines of a dataset, since the underlying patterns and information are not immediately visible to the human eye.

Relying solely on the creativity and intuition of the analysts can be a serious limiting factor, especially after the most obvious hypotheses have already been taken into proper consideration. Time spent looking at the data, exploring possible patterns, is always useful. The bottom line, which supports EDA as an investigative tool, is that reality can be less *intuitive* than expected, and therefore CDA alone may fail to give a comprehensive picture of the problem at hand and of its possible solutions.

References

1. Barbieri D, Air terrorist attacks: A time series analysis, ARMLET, Bucharest (Romania), 21-23 September 2016.
2. Beherens JT, Principles and procedures of exploratory data analysis, *Psychological Methods*, 2(2): 131-160,1997.
3. Brockwell PJ and Davis RA, *Introduction to time series and forecasting*, Springer, New York (USA), 2nd edition, 2002.
4. Evans J St B T, *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum, 1989.
5. James G, Witten D, Hastie T, Tibshirani R, *An Introduction to Statistical Learning*, Springer, New York (USA), 6th edition, 2015.
6. Mann PS, *Introductory Statistics*, John Wiley & Sons, Hoboken (NJ), USA, 2010.
7. Nickerson R S, Confirmation Bias: A Ubiquitous Phenomenon in Many Guises, *Review of General Psychology*, 2(2): 175–220, June 1998.
8. Popper K, *Conjectures and Refutations*, Routledge, New York (USA) 2002a.
9. Popper K, *The Logic of Scientific Discovery*, Routledge, London (UK) and New York (USA), 2nd ed, 2002b.
10. Tukey J, *Exploratory Data Analysis*, Pearson, 1st ed, 1977.
11. Witten IH and Frank E, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann Publishers (Elsevier), San Francisco, CA (USA), 2005.