Symposium on Big Data in Finance, Retail and Commerce:
Statistical and Computational Challenges
Lisbon, 2-3 November 2017, Portugal

Exploratory (Big) Data Analysis

Albert Satorra (Speaker)
UPF, albert.satorra@upd.com

Catia Nicodemo
Oxford University, catia.nicodemo@gmail.com

Abstract
The collection of large amounts of databases, big data, (clinician, social security, facebook...), is becoming more common, as the research studies that use them. Usually this type of data are quite rich and with many potential benefits. However, the description and the analysis of big datasets could be hard to performance without the right techniques.

A primary goal of this paper is to present clearly and efficiently via statistical graphics, plots and information graphics to describe and explore big datasets. Effective data visualization helps to analyse and understand about data and evidence. It makes complex data more accessible, understandable and usable like to make comparisons or understanding causality. Charts are used to show patterns or relationships in the data for one or more variables facilitating the task to figure out the description and the possible correlation in the data.
In this paper we use the statistical software R that provides new tools to display in real-time changes and more illustrative graphics of the big databases, thus going beyond pie, bar and other charts. These illustrations veer away from the use of hundreds of rows, columns and attributes toward a more artistic visual representation of the data. In this paper we focus our attention in ***ggplot2*** is an R package for data visualization. It provides a powerful model of graphics that makes it easy to produce complex multi-layered graphics.
To design our study, we use two big databases: the Spanish Social Security data (Muestra Continua de Vidas Laborales) hereafter MCVL) in 2010[1] and the Hospital Episode Statistics data (HES) for UK [2].
The MCVL data comes from the register of the Social Security System (SS) in Spain for active people in the labor market, representing more than one

[1] More information here *http : //www.seg−social.es/prdi*00*/groups/public/documents/binario/*190
[2] See for more details here *http : //content.digital.nhs.uk/hes*

million people each year. The data set gives information of all of the historical relationships of any individual with the Social Security System (in terms of work and unemployment benefits). The data will include more than 15 million of observations per year.

The HES data provide information concerning all inpatients and outpatients admitted to NHS hospitals from 1989-90 onwards. It includes private patients treated in NHS hospitals, patients resident outside of England, and care delivered by Treatment Centres (including those in the independent sector) funded by the NHS. Each patient record contains detailed information, including: clinical information, patient characteristics, such as age and gender, and administrative and location information, such as method of admission, and the geography of treatment and residence. Since our focus is on GP influence upon admissions, our analysis concerns only the 'first admission' to the hospital, which the GP is most likely to influence, rather than admissions for continuing treatments. The database contains more than 80 million of observations per year.

Big data are data on a massive scale in terms of volume, intensity, and complexity that exceed the capacity of standard software tools. They present opportunities as well as challenges to statisticians. There are several statistical methods to analyse the big data like learning machine, Lasso, etc,. However very few evidence is on how to describe databases with huge number of observations.

The analysis in this study consist first in preparing the databases and after using *ggplot2* to explore the data and present several factors. In particular, we will study from the social security data the correlation between wage and age conditional on the fact to have a permanent or temporary contact, to be male or female, and the level of education. The scope is to analyse the difference across young and old people in a three-dimensional way. Traditional graphs, like plotting the two variable again each other, are useless. This analysis allow us to explore more deeply the data and present them in an easy way to understand from a general audience like policy makers, stakeholder, etc.

The health data are explored to look at the correlation between the referrals (a specialist visit in the hospital) and the treatment (a surgery) in the hospital at practice level. Among general practice in the UK exist a huge variation in terms of people referred to a hospital and people treated in the hospital. Before to analyse these data with statical complex models we want to present trough "ggplot2" the variation across practices and understand the possible co-factors that could driven certain results, like for example if the practice is in a poor or reach area.

Our study will bring a reference for people interested in exploring the data before to think which is the best model to predict the outcomes.

Keywords: Social Security Data, Health Data BIg Data