Emanuele Baldacci (Eurostat)
Dario Buono (Eurostat)
Fabrice Gras (Eurostat)

The curse of dimensionality in official statistics?

Topic 3 – More rapid statistics and indicators on new phenomena

Keywords: Big Data Analysis, Classification, Dimensional reduction

Introduction

Statistical authorities need to produce accurate data faster and in a cost effective way, to become more responsive to users´ demands, while at the same time continuing to provide high quality output. One way to fulfil this is to make use of all new accessible data sources, as for example administrative data and big data. As a result, statistical offices will have to deal more and more with a "huge" number of time series, in particular for producing model based statistics. For example, timeliness or accuracy of several short-term macro-economic indicators can be improved by using "huge" data sets as literature review conducted in a project of Eurostat1 can show it. The challenges raised by the curse of dimensionality will need to be addressed more and more by statistical authorities. The article will then review some related methodological problems, but also illustrate how the curse of dimensionality could change practices in official statistics, in particular in the field of traceability, dissemination and communication.

Methods / Problem statement

At a first stage, this article will review some possible dangers that should be kept in mind when using a "huge" number of dimensions such as: data storage, data snooping, relevancy of the Euclidean distance in high dimension, limitation of machine learning in high-dimension computation time, sampling granularity, interpretability of the results, etc… At a second stage, some existing statistical methods derived from a Eurostat project on big data and nowcasting2 in order to deal with high-dimensional data sets will be briefly presented. Among those methods, it will be pointed out the selection of relevant variable through regularisation methods (LASSO) or Bayesian approaches, the extraction of latent factors through the use of factor models (dynamic or static) or the structuration of information through clustering methods and classification methods (k-mean algorithm, nearest neighbours). At a later stage, the article will show how these methods could introduce explicit and implicit nonlinearities that increase the curse of dimensionality, but also to some extent, that raised some challenges related to the activities of statistical authorities. For example, it will be shown how official statistics could be urged to change the approach for producing, storing and disseminating information.

Results / Proposed solution

In particular the presentation will focus on three specific points:

• The logistical issue related to the storage and access to data of a "huge" and various amount of time series, taken into account the ESS specificities.

• The methodological issues related to the increase use of model based statistics that select different variables along time and the need to consider more and more the estimation time span and the frequencies of the regressors as parameters.

• The communication and dissemination issues related to the presentation of statistics following complex probability distributions that cannot be summarised by a few set of parameters, which will imply to calculate or estimate the entire probability density function.

In order to answer to the issues presented in this article, some possible solutions are:

• Data virtualisation as a possible alternative for data storage.

• Use adapted distance and/or typology in order to deal with the curse of dimensionality.

• Develop methods for estimating intrinsic dimension of problem (see Taken's theorem).

• Develop a sound methodology in order estimate optimal time span estimation.

• Develop new dissemination tools based on the knowledge of the entire density function, in particular based on presentation of scenarios.

Conclusions

Using high dimensional datasets will most likely urge statistical authorities to follow a different approach, in particular to be conscious that the measurement of socio-economic variables will follow more and more non-linear processes that could not be described by probability distributions that could be easily described by few parameters. It will thus imply to adapt the way to observe the world through data taking into account at a greater extent uncertainty and complexity, which will in turn impact dissemination and communication activities of statistical authorities.