Economic Commentaries

SVERIGES RIKSBANK

NO. 17, 2015

The Riksbank's future information supply in light of Big Data[1]

Jyry Hokkanen, Tor Jacobson at the monetary policy department and Deputy govenor Cecilia Skingsley and also Markus Tibblin at the monetary policy department[2]

Data and statistics are a cornerstone of the Riksbank's work. In recent years, the supply of data has increased dramatically and this trend is set to continue as an ever-greater amount of activities are stored automatically in different ways. This data revolution, which has given rise to concepts such as Big Data, challenges traditional thinking while placing new demands on processing and analysis. New analytical methods for Big Data are developing rapidly and there are now several applications that are of interest to central banks. To remain at the cutting edge, the Riksbank is working on an information strategy to ensure that relevant data and statistics are available for the decisions taken both today and in the future.

The concept of Big Data refers to the extremely large amounts of data that can now be retrieved and stored thanks to various technical breakthroughs. The concept of Big Data also includes streaming data and non-traditional data such as text. This data is often complex and in many cases unstructured, which means that it normally has to be processed and analysed using special methods.[3]

The management of large datasets is nothing new; it has been done by both researchers and statisticians for a long time. Nowadays, however, there are better technical solutions we can use to gather, store, structure and analyse increasingly large amounts of complex data, including all the new data generated by our presence on the Internet. This, in turn, creates the potential to manage, structure and extract knowledge from large, streaming or non-traditional datasets in a way that was previously impossible.

Big Data is a data revolution which will also affect how central banks use and analyse data. Aggregated time series data, published with a time-lag, has traditionally been used to follow economic development. Non-traditional data could further improve our understanding of economic development. If such data also is continuously observed, decision-makers can react more quickly to changes in development. However, in order for authorities like the Riksbank to able to benefit from these new, rapidly growing datasets, a well-considered strategy is required, based on the premise that data is a strategic resource. The data revolution challenges traditional thinking not just regarding data capture and analysis; it also requires new skills, new technology and a fit-for-purpose organisation.

Big Data – analysis methods and fields of application

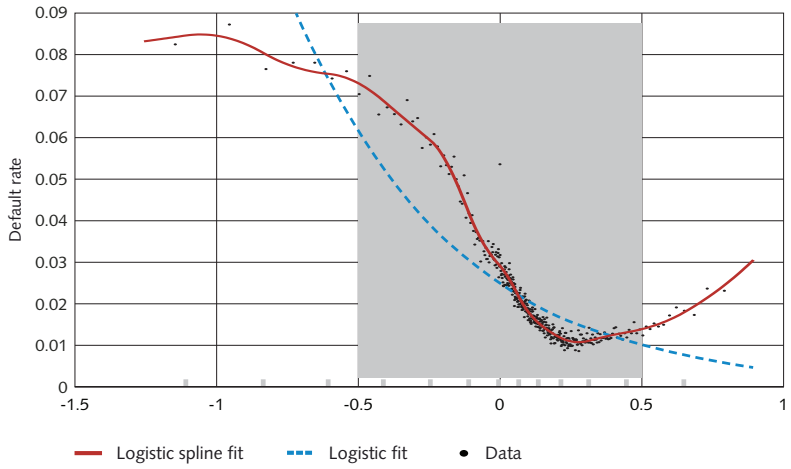
In a pedagogical and easily accessible article from 2014, Google's chief economist, Hal Varian, presents a set of analysis methods for managing big datasets, normally referred to as "machine learning", see Varian (2014). Machine learning methods have developed in the interface between statistics and computer science, and the term is an allusion to the fact that some of the methods are "self-specifying", that is, the algorithms build their own model design based on the appearance of the data, and thereby "learn" from it. Put simply, machine learning focuses, to a much greater extent than traditional econometric analysis, on predictions, or forecasts. There is therefore greater acceptance of a "black-box" approach, in which the analysis results are typically not based on economic theory.[4] When we have large datasets, the data can be divided up and used for different purposes. For example, the model

1. In September 2015, the Riksbank organised a workshop entitled "Big Data: Building data strategies for central banks in light of the data revolution". The event gathered together central banks, authorities, researchers and commercial users of Big Data. The presentations from the conference are available on the Riksbank's website at http://www.riksbank.se/sv/Press-och-publicerat/Nyheter/2015/Riksbanken-ordnar-workshop-om-big-data/.
2. We would like to thank Mikael Apel and Marianne Sterner for their comments.
3. Examples of what could be called Big Data are Twitter posts, information from the enormous trade repositories accumulated as a result of regulations such as MiFid or EMIR, and real-time (streaming) data from roadside cameras.
4. The expression "black box" refers in this context to the fact that the models' correlations or relationships cannot always be interpreted directly. They generate good forecasts, but it is not clear how.

can "train" itself or learn from a subset of observations and then be evaluated on a different subset. Traditional econometric analysis, adapted to and developed for small datasets and often aggregated data, is based on economic theory. In this respect, machine learning is less dependent on theory and more dependent on what the data looks like, i.e. it is data-driven.

Large datasets also provide greater freedom in the choice of functional form. The large amount of observations makes it possible to detect complex, non-linear relationships that describe the data's appearance better than the enforced linear relationships emanating from attempts to apply economic theory.[5] This is exemplified in the figure below for the correlation between Swedish limited companies' profits as a share of total assets and default (see Giordani, Jacobson, von Schedvin and Villani 2014). Using a so-called logistic probability model, a conventional model approach estimates a constantly declining default risk, the higher the profits are (the broken line). An option with a flexible, non-linear "spline" model captures the empirical relationship in the data better and clearly shows that "excess profits" are instead associated with an increased risk of default (the unbroken line that follows the observation values). The first, logistic, model underestimates the risks for low and high profit companies, but it overestimates the risks for companies that make normal profits. Unsurprisingly, we find that the non-linear model makes significantly better default risk forecasts than conventional linear models. This type of spline model based on data for entire populations of Swedish limited companies over a long time can, for example, be used to evaluate the credit risks associated with Swedish banks' corporate lending activities, as part of the Riksbank's financial stability monitoring remit.

Figure 1. Correlation between profits and default risk for Swedish limited companies
EBITDA/TA



The information captured from the Internet is often unstructured and consists of text or text fragments that need to be analysed.[6] The sources are many: for example, all types of news reporting, social media, market reports and market analyses.[7] A well-known example is the data generated by various online search engines. Bholat et al. (2015) show that there is a strong positive correlation between the British unemployment level and the number of Google queries for information on the terms for claiming unemployment benefit. This shows that the variations in search frequency on Google could therefore be used to capture current changes in important underlying variables such as total unemployment.

5. The decision to assign linear correlations in an empirical model often reflects a desire to take a linear, theoretical approach to data prediction, but may also be a necessary concession since the number of observations is insufficient to fully describe the empirical relationships.
6. The Internet also enables the capture of numerical data. A topical example of this is provided in a development project at the Riksbank which aims to collect prices of goods and services by "scraping" from the Internet in order to estimate and describe the development of inflation in real time.
7. The Centre for Central Banking at the Bank of England has presented a review of different text analysis methods which central banks can apply (see Bholat, Hansen, Santos and Schonhardt-Bailey, 2015).

Rönnqvist and Sarling (2015) show in a new study how advanced text analysis can be used to forecast distress events in individual banks.[8] In simple terms, the method utilises the fact that certain words from the same domain concept tend to occur in similar contexts. In other words, by carrying out a semantic analysis of words and of the content of entire articles, combinations of words and text that can predict distress events can be identified. The model can be used to generate a stress index for both individual bank institutions as well as entire business groups. Even though an increased stress index need not necessarily be interpreted negatively, it gives a signal of when it is reasonable to make an in-depth analysis of a specific bank. Since the model is based on textual data, it is possible to obtain descriptions of the events that drive the index via extracts from the source texts. The model can therefore enable us to both detect and describe events that are relevant to economic development.

A modern central bank needs an information strategy

In addition to the general growth in data, the increased regulation and monitoring of the financial sector in recent years has led to a sharp rise in the amount of data and statistics for analysis and monitoring.[9] This development is expected to continue at an ever-quicker pace and several central banks have reviewed their data and information strategies in order to benefit.

In the data strategies now being formulated, data is seen as a strategic asset for the whole central bank. Big Data (together with traditional data) is seen as something that can lead to a deeper understanding of underlying economic phenomena and provide early indications as to the state of the economy. For this vision to be realised, an overarching strategy is required for how the gathering, processing, storage and distribution of data is to be managed so that the entire organisation can benefit from the data analysis. The strategies are developed in collaboration between not only economists and analysts, the primary users of the data, but also lawyers and IT managers.

Implementation of the new strategies requires more resources, new skills and clearer coordination among the central bank's various policy areas. At the US Federal Reserve and the Bank of England, implementation has also involved organisational changes. The responsibility for overall data management in these organisations has been moved to a separate department – the Office of the Chief Data Officer. This department has recruited experts in the fields of data architecture, data compliance and data governance in order to develop and implement intra-bank systems and working routines for data. They are primarily responsible for developing and administrating the overall data model[10], establishing simple and secure processes for data gathering and processing, and ensuring that data is stored in a standardised and secure way but is at the same time easily accessible to different users. A similar project, although in a different form, has been implemented at the European Central Bank (ECB). An overall infrastructure and data model is being developed centrally within the ECB, and considerable focus is being put on structuring data gathering and storage and on making data easily accessible for users from several policy areas.

Big Data gives rise to new challenges for central banks

Rapid developments and the consequent changes in the way data is perceived have posed a number of new challenges for central banks. Despite the advantages of being able to easily access data from other policy areas, it has been difficult for different departments at central banks to adjust their data management immediately to an overarching framework. The Federal Reserve and Bank of England succeeded in counteracting departments' reluctance by introducing an intra-bank data council.

8. The data for their prediction models are of two types: firstly, 6.6 million news articles in Reuters's open archive from the period 2007-2014, and secondly, 243 identified "stress events" that occurred during the period 2007-end of June 2012 and that concerned one of the 101 systemically important major European banks included in the study.
9. For example, new derivative market regulations have resulted in large amounts of detailed data being collected. These sources are currently estimated to contain more than 15 billion data points.
10. A model that describes how various data is interrelated.

The data council takes decisions on overall issues linked to the gathering, processing, storage and distribution of data. Proposals for changes to data management are presented and revised in this council, and intra-bank decisions about data can be taken with every part of the central bank represented. The data council is an important factor in the change process, and has ensured that central bank departments have been comfortable adjusting to the overarching data principles established.

Analysing the new sources of data has also been a challenge in practical terms. Firstly, it has been difficult to recruit skilled people within the field of "data science"[11], and secondly, the IT capacity of the central bank has not always lived up to the requirements laid down for processing large amounts of data. The methods of analysis also often generate complex models, which makes interpreting results difficult and the mediation of policy recommendations even more so.

The future of the Riksbank's information supply

The Riksbank's responsibility for monetary policy and financial stability requires access to data from many different sources for analyses and forecasts. A cornerstone of data acquisition has traditionally been the statistics regulated by the Riksbank itself in regulations for financial institutions, as well as other statistics and data (either public or purchased). Like other central banks, the Riksbank also needs to review how data is managed effectively internally, and how data – in many cases highly detailed – is dealt with from new sources, which could also contain unstructured information and which requires another form of processing and analysis than has previously been the norm. In order to ensure information is handled in an appropriate manner, a vision for information supply is therefore required, along with an accompanying strategy which guides how data is acquired and processed.

The Riksbank's vision for its information supply can be expressed simply by *The right data in every situation.* The strategy therefore requires relevant data, if not already present at the Riksbank, to be acquired in a familiar way, at a known cost and from the right source. In addition to this, the Riksbank is also required to know how data is made available to the right person at the right time, and to ensure that decisions and responsibilities are clarified in terms of retrieving, storing and accessing data. The need to retrieve and analyse principally non-structured data places particularly high requirements on analysis tools and data infrastructure, and also demands active international monitoring. The ambition to have access to relevant data is resource-intensive and must always be set against the need for analysis, costs, obligation to provide information, and more.

Conclusions – the data revolution makes new working methods possible

The quantity of available data is enormous and is continuing to increase at an ever-greater pace. Along with new methods of analysis, new technologies have made it possible to analyse Big Data. Several central banks think they can exploit this development to create better foundations for their policy decisions and have therefore changed their strategies for managing data and statistics. New staff with skills in managing and analysing this data have been recruited to these central banks. This is partly to take care of the gathering, processing and storage of data, and partly to benefit from the content in new types of data, such as textual information or enormous quantities of microdata, by using new IT systems and methods of analysis.

Big Data also poses considerable challenges for central banks, both technically and methodologically. In addition, new considerations need to be made in terms of strategy, organisation, skills and budgets.

11. Data science is an interdisciplinary field within which different methods and systems are used to gather, process and analyse Big Data. The methods used are based on a combination of ideas from computer science, programming and numeric and statistical analysis. Big Data projects often involve individuals who have mutually complementary technical skills within these fields.

Relevant data and statistics are a cornerstone of the Riksbank's work. To remain at the cutting edge, the Riksbank therefore needs a strategy to ensure that relevant data is available for the decisions taken both today and in the future. The motto *the right data in every situation* means the need for analysis will govern which data is gathered, while acquiring, managing and analysing data will always need to be set against the costs and potential added value an analysis would provide.

References

Bholat, David, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey, 2015. Text mining for central banks. Centre for Central Banking Studies Handbook No. 33.

Giordani, Paolo, Tor Jacobson, Erik von Schedvin, and Mattias Villani, 2014. Taking the Twists into Account: Predicting Firm Bankruptcy Risk with Splines of Financial Ratios. *Journal of Financial and Quantitative Analysis*, 49, 1071-99.

Rönnqvist, Samuel and Peter Sarlin, 2015. Detect & Describe: Deep learning of bank stress in the news. *Proceedings of the 2015 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, forthcoming.

Sveriges riksbank, presentation material from the workshop, "Big Data: Building data strategies in light of the data revolution", www.riksbank.se/en/Press-och-publicerat/nyheter/2015/Riksbanken-ordnar-workshop-om-big-data/.

Varian, Hal R., 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28, 3-28.