Benoît Cœuré: Policy analysis with big data

Speech by Benoît Cœuré, Member of the Executive Board of the ECB, at the conference on "Economic and Financial Regulation in the Era of Big Data", organised by the Banque de France, Paris, 24 November 2017

The recent financial crisis, and the euro area sovereign debt crisis that followed, were characterised by periods of increased heterogeneity, market fragmentation and sudden turns in economic activity. This often made it difficult for economic policymakers to understand and assess in real time the underlying forces driving economic behaviour. Both traditional statistical datasets and our models proved at times inadequate to support the decision-making process, reflecting long time lags, linear assumptions and the absence of more granular information.

These events increasingly boosted the efforts in policy circles to obtain timelier and richer data for policy analysis, in short big data.[1] This push towards more granular information was not a revolution, however. It can be argued that big data, under different guises, have been used as an input into policymaking since Adolphe Quételet's *Mémoire* in 1848. Since then, big data has been central to business cycle analysis, from the early work of Clément Juglar to the contributions of both Wesley Mitchell and the Cowles Commission, right up until today. According to central bank mythology, former Federal Reserve Chair Alan Greenspan would sit in his bathtub perusing sheets of statistics. And indeed, economic historians have analysed how the power of governments has been shaped by statistics – and vice versa.[2]

The most recent push towards more granular data was thus an evolution rather than a revolution, triggered in part by the emergence of new opportunities – themselves a reflection of rapid technological progress – and the experience gained over several years of crisis management.

This evolution is already bearing fruit. Policymakers today have access to a large number of micro datasets, often very different in nature and scope. Some are the result of new financial regulations. Others are by-products of increased use of technology. What they have in common, however, is that, if used appropriately and responsibly, they can help policymakers to extract more timely and diverse economic signals, and thus are a meaningful complement to existing official data.

In my remarks this morning, I will take stock of the progress made at the ECB, and in the central banking community more generally, on the use of more granular data in the conduct of monetary policy.[3] I will show that micro

data collected by central banks themselves, in particular data on transactions between banks, have already proven to be an additional valuable guide for policymakers in devising policy responses.

And I will show that data generated through the greater use of technology in our daily lives have an enormous potential to help policymakers overcome prevailing constraints on timely data availability, understand better the consequences of their policies and calibrate them accordingly, while also creating challenges.

Central banks as big data collectors

Let me start with the role of central banks as producers of big data.

Central banks do not have to be at the forefront of data collection, and we should not seek to displace private sector efforts. That said, there are areas in which central banks have started, or are about to start, collecting large amounts of data to help them monitor developments in financial markets and allow them to extract richer information about the transmission of monetary policy, which in turn helps us calibrate our policies.

As a showcase for these efforts, I would like to discuss two such initiatives, namely the money market statistical reporting (MMSR) data, which the ECB, in collaboration with the wider Eurosystem, has been collecting since July 2016, and the so-called AnaCredit project, short for "analytical credit datasets", which will produce its first results over the course of next year.[4]

MMSR data contain confidential daily information on the individual euro-denominated loans in the euro money market from the 52 largest euro area banks, which collectively account for approximately 80-85% of the total balance sheet of euro area banks. At present, this means collecting information on 10,000 daily transactions in the unsecured money market, with a daily volume of around €100 billion. We also collect data on around 30,000 daily transactions on secured loans, worth around €500 billion.

Earlier this week, the project took an important step forward with the publication of the first set of euro area money market statistics, covering each of the Eurosystem's reserve maintenance periods in 2017.[5] The published data cover the unsecured market. The aim is to publish data on the secured market in 2018 once we are satisfied that the data are of sufficiently high quality.

The sheer volume of data, combined with the high frequency with which it is collected, means that the standard verification process involving human beings is not feasible. Carrying out checks by algorithm, using machine learning techniques and artificial intelligence is one way of ensuring that data remain of high quality. The Eurosystem has a steep learning curve in front of itself.

These efforts will no doubt pay off. Indeed, MMSR data have proven highly useful to policymakers in their short period of existence. Let me give you an

example. The data help us assess the impact of the ECB's asset purchase programme (APP) on market functioning, as I explained in more detail in a speech I gave last week in Brussels.[6]

MMSR data show that since the ECB launched the APP there has been a marked shift away from trades backed by general collateral, which are traditionally used to manage cash. Instead, there has been a growing share of repos off the general collateral curve, termed specials. Take for example the German Bund market. Traditionally, only around 5% of bonds in the German repo market traded as special, but in the second half of last year, that share rose to 50%.

Being aware of the distribution of trades and the premium placed on special bonds enables us to assess the impact of the APP on market functioning. To ease potential frictions, we decided last December to allow cash to be used as collateral in our securities lending programme, and to permit APP purchases below the deposit facility rate. As a result of these decisions, the "specialness" premium on long-term German bonds has declined, and the share of bonds that trade special is now around 30%.

A further notable area of use for MMSR data has emerged from growing concerns about the reliability and robustness of current risk-free benchmark rates for the euro area.[7] Banks have become increasingly reluctant to participate in benchmark panels, owing to concerns about potential litigation, compliance risks and costs. The resulting uncertainty in the integrity of reference rates represents a potentially serious source of vulnerability and systemic risk.

Against this backdrop, the ECB recently announced that it intends to provide a new overnight unsecured interest rate, using MMSR transaction data. The new rate is intended to complement existing benchmark rates produced by the private sector and serve as a backstop reference rate, with the aim to start publishing by 2020.[8] This move is motivated by our desire to mitigate the potential adverse impact on the monetary transmission mechanism and on financial stability from the lack of reliable benchmarks. The market facilitating role played by the ECB in this field is consistent with the tasks conferred upon it by its Statute.

The process will involve broad consultation with market participants, end-users and other public authorities. A first public consultation will be launched before the end of 2017.

The use of MMSR data implies that the new overnight interest rate will differ from the current EONIA benchmark in a number of ways.[9] Accordingly, any decision pertaining to this new benchmark rate does not bear direct consequences for the choice of our operational target.

Let me now say a brief word about the second initiative I mentioned earlier, namely the AnaCredit project, which will push the frontier of big data use at the ECB further out.[10]

AnaCredit will deliver loan-by-loan information, mostly on a monthly basis,

on credit to companies and other legal entities extended by euro area banks. Early estimates point to around 70 million exposures reported every month, representing loans granted by about 4,500 credit institutions to more than 15 million counterparties.[11] In view of these large expected data volumes, a state-of-the-art IT infrastructure is currently under development, which will also take into account the need to ensure the adequate protection of confidentiality.

AnaCredit data collection has been designed to produce a complete picture of the credit exposure of the reporting population. The information collected comprises almost 100 different "attributes" covering various aspects of the credit exposure, such as the outstanding amount, maturity, interest rate, collateral or guarantee, information on the counterparty, and many other things.

Access to highly granular loan-level information will be a major step forward in helping policymakers to analyse and monitor credit developments and to assess the impact of their decisions on bank lending. As you know, bank lending plays a key role in the euro area, where the share of loans in the total external financing of small and medium-sized enterprises (SMEs) is considerably higher than in the Anglo-Saxon world, where market financing plays a more important role.

Granular AnaCredit data will help us look beyond aggregates and extract the underlying developments. Granular loan-by-loan data will allow us to know the characteristics of specific groups of counterparties involved in each transaction, without of course revealing their identities. This means that we will be able to assess the driving forces behind aggregate developments and distinguish genuine and healthy growth from potential exuberance. This is crucial for policymakers.

And, finally, although no data will be collected specifically for supervisory purposes in the initial stages of the project, the information will also be very useful in many areas of banking supervision. I am thinking in particular of the information on the link between lenders and the unique identification of counterparties across the entire lender population.

The use of big data in policy analysis

Let me now turn to the broader issue of the use by central banks of technology-driven data.

An early and prominent example of online data being used for policy analysis is the "Billion Prices Project", which was launched by the Massachusetts Institute of Technology (MIT) in 2008.[12] Today, the project publishes *daily* online price indices for more than 20 countries, based on a technology called "web scraping", where price information is collected automatically by machine from hundreds of retailers that also have physical outlets. By 2015, about 15 million prices were being collected every day from 900 retailers.

The advantages for policymakers are plain to see. Online inflation data, if of high quality, are much timelier than current price statistics and may

cover a much larger number of products. Eurostat currently collects, via the national statistical institutes, roughly three million prices for the Harmonised Index of Consumer Prices every month in the 28 EU Member States and it publishes the index 17 days after the end of each month.

Online price data can therefore be used to improve short-term forecasts and to check the robustness and reliability of current price indices. Mixed-frequency models, for example, could be used to enhance existing forecast models that are based on monthly data.

Sveriges Riksbank has already tested the use of online price data. The results indicate that the data add value when it comes to forecasting short-term developments in consumer prices for fruit and vegetables.[13] Preliminary explorative analysis at the ECB confirms the predictive power of online price data, provided they are sufficiently granular.

This suggests that online prices may complement existing price data. Daily online data can also add significant value to policymaking in periods of high uncertainty. As research has shown, would the full Billion Prices Project data have been available in 2008, the turning point in US inflation following the demise of Lehman Brothers could have been identified months before it showed up in the official US consumer price Index.[14]

Online prices are, however, only one source of potential interest for policymakers. Large-scale barcode scanner data, for example, are an alternative avenue that researchers at the ECB have explored in the past to investigate factors that determine prices and the degree of price dispersion, much in the tradition of our long-standing efforts under the umbrella of the Eurosystem Inflation Persistence Network.[15] The data, which consist of 3.5 million observations on the price and quantity of individual products sold, have confirmed that competition at producer and retail level is a key factor affecting micro price-setting.

Insights gained from the use of online and scanner data have also encouraged the ECB to make micro-price research a strategic research priority between 2018 and 2020. The first step is to pick up where the Inflation Persistence Network left off, by resuming the collection of micro data underlying the Consumer Price Index and the Producer Price Index and complementing them, where possible, with scanner and online price data.[16] Collaboration within the Eurosystem will once again be vital for the success of this ambitious project.

The ECB, with the help of the national central banks, also draws on big data to help improve business cycle analysis. Google search data, for example, have been suggested as a potentially valuable source of data for policymakers. Back in 2012, Hal Varian and co-author showed that Google searches can help predict economic activity.[17]

Building on this idea, ESCB staff have explored the possibility of nowcasting unemployment using volumes reported by Google Trends for a large number of search queries broadly related to unemployment. They find that many of the search terms indeed correlate with unemployment and may reduce forecasting

errors by up to 80% by comparison with naive benchmark models.[18] Job search and related Google search quotes have also been found to be strong predictors of variations in subjective wellbeing.[19] Similar initiatives are underway for more leading indicators.

Related to this, electronic payment data from credit cards and cash withdrawals from ATMs have been shown to help forecast private consumption as well as GDP growth, provided that they are made available in a timely manner.[20] These data are also less prone to revision than traditional national accounts data. Our internal findings so far suggest that, in some euro area countries, the correlation between payment data and private consumption is encouragingly strong, making their systematic use in forecasting an option for policymakers.

Technological progress in exploiting textual information can provide similar benefits. Research shows that information extracted from business newspapers can be used to nowcast quarterly GDP growth and outperforms traditional forecasting methods at turning points.[21]

Similar methodologies can also be used to extract information on how central bank communication is perceived by the public. Specifically, text mining techniques are employed by the ECB to determine whether its communication has been interpreted as dovish or hawkish by the media and to assess its impact in unconventional times.[22] Given the increasing role of communication in determining the monetary policy stance, such analysis may usefully complement the feedback extracted from financial markets.

Challenges in using big data for policy analysis

Overall, these examples highlight the fact that big data can help policymakers overcome some of the shortcomings of traditional macroeconomic time series. In short, these are: improved timeliness, richer detail on interactions and the ability to test and develop old and new theories using data not previously available. For sure, big data have also provided plenty of opportunities to deepen our understanding of behavioural economics, and how psychology can drive macroeconomic developments.

At the same time, some of the data raise new challenges, many – but not all – of them are related to the statistical production process itself.

For example, the sheer scale and variety of big data result in the difficulty of processing unstructured data. The database interfaces used by central banks were designed in the past mainly to handle time series data. Strides have certainly been made to cope with cross-sectional data, such as those used in supervision. But putting in place effective data processing methods and interfaces that permit the retrieval and visualisation of big data represents a considerable IT challenge.

It is also a challenge in term of human resources. When recruiting staff, central banks have not traditionally sought experience with techniques developed to cope with big data. Building up expertise in this new field takes time, and it also involves strategic planning on the recruitment side.

But central banks are not the only employers wanting these skills, and competing against the private sector for the limited pool of highly sought-after experts is likely to prove challenging.

Apart from these operational considerations, there are also analytical, legal and ethical concerns related to the use of such data.

For example, perennial pitfalls involved in analysing traditional data also apply to big data. One misconception about big data is the common view that we no longer need to concern ourselves with sample bias, as large volumes of data take precedence over standard sampling theory, and that such data provide a census.[23]

Any statistician will know that this is of course not true. Take a dataset that includes all tweets ever made. Despite being a census of tweets, people who tweet are likely to differ in age, preferences and behaviour from those that do not, so several segments of the population are underrepresented, or do not appear in the sample.

Making assumptions about household behaviour without adequately re-weighting the sample is likely to result in biased and inaccurate estimates. Traditional datasets, designed to be representative of the entire population, might therefore sometimes be less subject to sample biases, although they are significantly smaller in size.

Similarly, big data also involve analytical challenges related to, for example, econometric identification. This is also referred to as the "curse of dimensionality", (sometimes called the "large p, small n paradigm") where there are many parameters but few observations.[24] Dynamic factor models, or Bayesian shrinkage, are methods which can help address the difficulties arising from the high dimensionality of data and, in fact, the term "big data" first appeared in econometrics in 2000 in relation to these models.[25] But some of these methods are still being developed, in particular for cases involving many observations – a situation which we will increasingly face over time as our datasets grow even larger.

Furthermore, the ability of big data to link individuals across various datasets raises important questions about privacy and confidentiality. A special Eurobarometer survey in 2015 showed that two-thirds of respondents were concerned about not having complete control over the information they provided online. A quarter of respondents considered it a risk that personal information could be shared with third parties, such as government agencies.

With this in mind, the European Parliament in April last year approved the EU's General Data Protection Regulation (GDPR), which will come into force in May 2018. This Regulation will provide strong personal data protection, while giving individuals greater control over their own data.[26] But statistical agencies and central banks are already taking serious steps to ensure that privacy standards are fully upheld when they make use of new big data, such as credit card transactions or Google search data. Anonymity in transaction data helps build trust and allows policymakers to leverage new technology to help take better policy decisions for the common good.

Differences in national data protection regulations, in turn, together with differences in data standards, also pose a challenge, in the form of our *global* ability to collect, aggregate, disseminate and share data. In today's interconnected world, a global approach to data collection is a prerequisite for developing a coherent view of the global economy and financial system, and for identifying vulnerabilities. But many legal hurdles to cross-border data sharing remain to be overcome, and implementation gaps to be filled, if the intended benefits of the data harmonisation reforms in the financial field are to be fully achieved.[27]

Regulation also plays a big role in decisions about accessibility. Consider the revised Payment Services Directive (PSD2), which will enter into force in a few weeks. One outcome of the new Directive is that banks will soon lose the monopoly they have on their customers' transaction data.

The implication is that new fintech firms will be able to use these data to design new tailor-made products and to allow real-time access to financial services anywhere and at any time. This can be expected to foster productivity. At the same time, regulators need to be mindful of the potential financial stability implications, in particular if incumbents fail to rise to the challenge, and fintech firms thus start to crowd banks out of a large range of financial services. Granting access to big data therefore also has the potential to shift economic structures.[28]

It even has the potential to change the course of policy. So far, policy has largely relied on data produced by governments, often in the form of abstract concepts, such as GDP.[29] But the more widespread ability to create individual statistics threatens the hegemony of policymakers in defining the parameters of policy. One risk is that individuals will use other, readily available sources of data that could undermine confidence in official statistics and, hence, in the central bank's commitment to price stability.

Just as there are concerns about "fake news" dominating social media, there is a risk of "fake", or at least poor quality, statistics driving out better quality ones in public discourse. And just as false reports may quickly go viral, reliance on internet search data in the assessment of the economic situation may become self-fulfilling. Actions by economic agents could become less anchored to actual activity and more prone to manias and panics, with obvious implications for economic and financial stability.

The challenges here are twofold for policymakers. The first is to meet the demand for more individually tailored information. Some statistical offices, for example, already offer the possibility to calculate a personalised inflation index using official price quotes, but with expenditure weights provided by the user. The second challenge is to recognise that public perceptions of what we should target may differ from the abstract definitions we employ. Over time, engagement with the public may result in changes to definitions, and even in the conduct of policy.

This brings me to my last challenge, which has a philosophical dimension to it. Over time major central banks have tended to become more transparent, in the belief that communication would aid the effectiveness of monetary

policy.[30] This has led to growing interaction between central banks and financial markets, simply because communication is not a one-way street.

You can see what I'm getting at. New technological advances may create a new "monkey in the mirror" mimetic loop[31], this time not through financial markets, but through big data. That is, we may one day be tempted to draft our monetary policy statements and speeches in the light of how they will be comprehended and interpreted by artificial intelligence algorithms. So in the future big data may work two ways, with central banks acting on, and reacting to, it, with consequences which remain to be understood.

Conclusion

Let me conclude.

Central banks have made considerable progress in recent years in integrating big new datasets into their policy analysis and decision-making. Granular data collected by central banks themselves have, in particular, become an indispensable source of information for policymakers.

Yet, it is probably fair to say that we are still exploring how to use much of the big data that new technology opens up to us. Although evidence is growing that online data may provide tangible benefits for short-term forecasting, more research is needed to ensure that the data are of sufficient quality and reliability to systematically inform policymaking.

The potential of such data to enrich central bank analysis in the future is considerable, however, as are the challenges that come along with it. For this reason, I encourage all of you to continue your efforts and to work in the same collaborative spirit as you have done so far.

Thank you.