CEMLA
CENTER FOR LATIN AMERICAN MONETARY STUDIES

VIII MEETING OF THE FINANCIAL INFORMATION FORUM

OPENING REMARKS: **Dr. Manuel Ramos Francia**
Director General, *CEMLA*

CEMLA Mexico City
May 24, 2022

Emerging challenges for the use of Big Data in central banking

Good morning. It is a pleasure to welcome you to the VIII Meeting of the Financial Information Forum (FIF), organized by the Forum´s Executive Committee and hosted digitally by CEMLA. Spanning over three half-day sessions, this meeting features a distinguished group of international policymakers from Latin America, the U.S., and Europe, who will share and discuss their experience on emerging challenges for central banks' financial statistics.

This year's agenda will be focused on key topics for central banks' statistical units, including the incorporation of non-traditional data sources to banks' statistics, the construction of financial market indices in derivatives and non-bank financial intermediation sectors, and the implications of financial digitalization for central banks' statistics.

I wanted to start by thanking the Member Countries of the Forum´s Executive Committee, who have done an excellent job in arranging the agenda for this year's meeting. In particular, the support and collaboration from representatives from the central banks of Brazil, Chile, Mexico, and Peru has been crucial to define and prioritize the topics that will be discussed in the next days.

In addition, I would like to thank the speakers that will participate in the panel and keynote sessions. We are foremost honored to host Ron S. Jarmin, Deputy Director of the U.S. Census Bureau as our keynote speaker today. Ron has led numerous initiatives to incorporate data science in the Census Bureau. Moreover, he has recently edited a book on Big Data for Twenty-First Century Economic Statistics[1], published by the National Bureau of Economic Research. The book presents a thorough account of the opportunities and challenges of big data, topics that we expect to discuss with him later today.

Last but not least, let me thank our staff at CEMLA for assisting the Forum's Executive Committee in the organization of the meeting. Particularly, Matías Ossandon Busch from the Financial Stability Directorate and our IT team.

The advent of big data

The fundamental changes in our economies triggered by the Covid-19 pandemic have come along with the advent of an age of big data, a process that has been further reinforced by phenomena such as the shift towards e-commerce and digital finance

[1] See Website of the NBER

during the pandemic.

As policymakers, we have historically relied on official statistics often collected and published with a low frequency and important lags to monitor risks and assess the stance and future path of the economy. Moreover, we have gotten used to the comfortable situation in which statistics can be obtained by simply relying on regulatory obligations that allow public institutions to access a steady flow of data from traditional market players.

This environment has now dramatically changed. The rampant digitalization has led firms and individuals to interact with financial markets via their own smart phones, cloud computing, peer-to-peer mechanisms, and a plethora of globally interconnected digital platforms and financial products. The pandemic was in that sense only a catalyst of an already strong undergoing process in which traditional mechanisms to collect useful statistics to inform policy decisions were not coping with the speed and extent of changes in financial markets.

In this context, at least three complementary forces can explain the increasing importance of big data and other forms of non-traditional data processing for central banks.

First, markets' digitalization has left relevant economic activities and information channels outside the regulatory perimeter that allows public institutions to collect data. Second, the very nature of digitalization means that firms and individuals are producing themselves unprecedented amounts of data that remain stored online in the servers of firms with which they interact. Third, mounting evidence suggests that the degree of interconnection between digital and traditional bank based financial markets is creating relevant financial stability risks that require being assessed (IMF, 2022).

In this regard, exclusively relying on data reported to central banks entails the risks of leaving relevant portions of financial markets away from any monitoring capacity. Moreover, the pandemic has also highlighted how the speed with which economic shocks unfold cannot be properly monitored using traditional data sources, impairing central banks' capacity to react fast and with precision following their price and financial stability mandates.

How can central banks benefit from big data?

Despite the different definitions at hand, experts seem to agree that big data reflects the use of large non-traditional data that also requires special techniques – such as machine learning or natural language processing – to be processed (Doerr et al., 2021). This definition comprises structured administrative data such as credit registers or derivatives repositories, but also unstructured data, such as the one that can be retrieved via web scraping or by analyzing newspapers' text.

Size is, however, not the only or even the main feature of big data. The so-called 3Vs – volume, velocity, and variety – provide a better representation of key characteristics associated with this technology (Tissot, 2019).

These features provide also a hint on how central banks can benefit from incorporating big data techniques in their statistical frameworks. More than just adding larger datasets to their existing repositories, the invitation is for central banks to think about how big data approaches can improve the timing and sectoral coverage of the information with which policy decisions can be informed.

The good news is that central banks seem to recognize the challenges ahead. According to the BIS (Doerr, 2021), 80 percent of central banks as of 2020 reported using some form of big data, up from only 30 percent in 2015. While this sharp increase signals central banks' interest in incorporating big data, uses are still limited in many cases to research projects and pilot schemes. Let me stress two examples of how central banks can incorporate big data techniques aiming also at improving the way statistical units can assist informing policy decisions.

A first application relates to the need of improving central banks' capacity to monitor the state of the economy in a real-time fashion, an approach known as nowcasting. Nowcasting models can be fed with high-frequency non-traditional statistics, such as traffic or satellite information, trends in google searches, electronic payments records, or housing prices. For instance, indices based on google trends have proven to be useful to construct nowcasting models estimating labor market trends in different countries (D'Amuri and Marcucci, 2017).

These data are also useful to construct high frequency indices, such as the New York Fed's Weekly Economic Index. This index relies on high-frequency non-traditional data sources such as weekly unemployment insurance claims and railroad traffic records to proxy for the stance of economic activity[2]. Credit registers and derivatives repositories, such as the one developed by the Central Bank of Chile and that will be discussed in tomorrow's panel session, are also examples of novel approaches to data.

Another useful application relates to the use of machine learning and natural language processing approaches to construct economic indices. This approach is based on the notion different emphases in public statements, news, and other forms of written communication can proxy for economic sentiments and expectations. Studies have shown, for instance, that language processing can be used to identify forward guidance elements in monetary policy statements, which can be used to forecast market performance (Hansen and McMahon, 2016). These techniques can also be used to construct policy and economic uncertainty indices (Baker et al., 2016).

Challenges for use of big data and machine learning

While central banks reassess their response to emerging data needs, it is also important to consider challenges and potential risks related to a thoughtless impulse towards non-traditional data sources, which are certainly not a silver bullet. Let me stress a few of these challenges that may add to the discussions that will follow in the next sessions.

[2] See NY Fed Website

One aspect to consider is the possibility of introducing biases and measurement errors when relying on non-traditional data. Gender and socioeconomic gaps are likely to be introduced if we expect to track real-time economic dynamics by looking at online sources. These biases can be triggered both by the unequal access to digital markets and by the differential willingness of individuals and firms to share their data with private companies. This is certainly a relevant problem in Latin America given the still limited digital literacy in our societies.

Biases and measurement errors can also stem from problems with data quality, data cleaning procedures, the treatment of missing information, and the overall representativeness of the collected data. Here, one should keep in mind that unstructured datasets are in most cases a byproduct of economic activities whose main purpose is not to keep sound data repositories, calling for caution regarding their use.

Another related source of bias can emerge from problems with algorithmic fairness. Machine learning algorithms used to interpret unstructured datasets are, like any model, subjected to measurement errors. For example, an algorithm can be trained to identify signals of future financial crisis. However, if most past crises were triggered by stress in the housing market the model may underestimate the importance of, say, derivative markets in a new crisis. This backward-locking nature of algorithms' training can end up with economic indices that cannot properly inform decision makers.

The use of machine learning techniques is certainly a topic that deserves its own considerations. Despite its benefits in terms of reaching more accurate predictions or identifying market sentiments via text analysis, these methods do not aim at identifying causal relationships. When allowing the data to tell which model is the most accurate, other relevant variables could be left out of the analysis. Moreover, text-analysis applications are also exposed to an analyst's subjective decisions on how to define, say, a 'positive' economic outlook within a text. Finally, and as with any methodological innovation, we still lack a proper set of best practices to judge models' soundness.

Finally, a few other operational aspects should be considered. For example, a recent survey conducted by the Irving Fisher Committee shows that central banks consider the lack of an appropriate IT infrastructure and highly-trained staff as their main concern regarding the adoption of big data (Doerr, 2021). This concern is more acute in EMEs. Data protection and legal issues are also key, as the data storage in private firms can impose limits to large-scale data collection. Finally, introducing big data and machine learning methods can impose a non-trivial burden on central banks carbon footprints, an aspect that will likely gain in importance in the coming years.

Final remarks

These latter concerns should not discourage discussing about the many useful applications of non-traditional data sources and big data for central banks. In contrast, they highlight the importance of international cooperation to foster the adoption of new data approaches while moderating the associated risks. Central banks' cooperation in the form of technical assistance, training activities, and even joint IT infrastructures can, for instance, help to overcome the challenges ahead.

At CEMLA we are contributing to this effort with our Course on Machine Learning and Central Banking and other Technical Assistance Programs.

I would like to welcome you again and emphasize that the work of the Financial Information Forum is key to foster the dialogue between central banks on common challenges in their statistical units. At CEMLA we are grateful to host this year's meeting and we are looking forward to promote further collaboration and research initiatives among central banks that can improve our common understanding of relevant data challenges, especially in these uncertain times.

I wish you a fruitful discussion during the meeting. Thank you for your attention.

References

- Baker, S., N. Bloom, and S. Davis (2016). Measuring economic policy uncertainty. Quarterly Journal of Economics, vol 131, no 4, pp 1593–636.
- D'Amuri, F. and J. Marcucci (2017). The predictive power of Google searches in forecasting unemployment. International Journal of Forecasting, vol 33, no 4, pp 801–816.
- Doerr, S., L. Gambacorta, and J.M. Serena Garralda (2021). Big data and machine learning in central banking. BIS Working Papers 930, Bank for International Settlements.
- Hansen, S. and M. McMahon (2016). Shocking language: understanding the macroeconomic effects of central bank communication. Journal of International Economics, vol 99, no 1, pp 114–33.
- International Monetary Fund (2022). Global Financial Stability Report: Implications of the War in Ukraine. Washington, DC, April.
- Tissot, B. (2019). Financial big data and policy work: opportunities and challenges, Eurostat, Statistical Working Papers, KS-TC-19-001-EN-N.