Irving Fisher Committee on Central Bank Statistics

BIS

Granular data: new horizons and challenges for central banks

Joshua Brault, Maryam Haghighi, Bruno Tissot[1]

Executive summary

Central banks, as both producers and users of statistical information, have been increasingly interested in **making the most of the wealth of data available in modern societies**. But they are also fully aware of the need to address the important challenges associated with dealing with granular data sets. These data sets typically comprise micro-level records (such as at the individual account, instrument or transaction level) as well as disaggregated data (that are below the level of aggregated statistics and have a higher likelihood of identifying reporting units). The challenges they pose include not only the large size of these data sets and their often limited quality, but also their complexity due to the high level of details they entail and the sheer variety of their formats.

Despite these challenges, **granular data are appealing for a variety of reasons**. The most obvious is that they can offer a level of precision and comprehension into economic phenomena that aggregation may typically mask or hide. Further, granular data provide sufficient flexibility and adaptability to tackle a wide scope of issues. The insights gained can enable central banks to have a better understanding of emerging behaviours and lead to more targeted and effective policy prescriptions. Additionally, since many micro data sources are a by-product of the increasingly digitalised world, they provide information that can be timelier and more resilient to disruption, such as during financial crises. Moreover, granular data can support the production of novel economic indicators (for example on economic confidence or uncertainty) as well as experimenting with alternative ones. Finally, they have proved to usefully complement traditional macroeconomic statistics, for instance by helping to assess real-time accuracy and address gaps or missing data.

While granular data offer a level of detail and insight not previously possible with aggregate sources, **central banks must confront two important issues**. First, these data inherently lack the thorough production processes and quality assessments of typical macroeconomic statistics. Second, to provide value they must be situated within a coherent framework and contextualised so that they can ultimately inform policy decisions.

[1] Senior Data Scientist, Data Science Innovations Division, Bank of Canada (JBrault@bank-banque-canada.ca); Director of Enterprise Data Science and Insights, Bank of Canada (MHaghighi@bank-banque-canada.ca); and Head of Statistics and Research Support, BIS & Head of the Secretariat of the Irving Fisher Committee on Central Bank Statistics (IFC) (Bruno.Tissot@bis.org).

The views expressed here are those of the authors and do not necessarily reflect those of the BIS, the Bank of Canada, the IFC or any of the institutions represented at the event.

We thank Douglas Araujo, Gloria Peña and Olivier Sirello for supporting comments and helpful suggestions.

Fortunately, central banks' experience shows that **data science and its large set of new tools and techniques can be very effective at tackling these issues**. Four aspects deserve careful consideration:

- **Granular data in their raw form are often unstructured, unverified and dispersed**.[2] Consequently, lengthy pipelines are typically required to receive, transform and validate the information collected before it can be useful. Due to the vastness of the data, these operations can hardly be labour-driven and require a high degree of automation. Fortunately, recent advances in artificial intelligence (AI) and machine learning (ML) offer great opportunities to harness the insights from granular data, by performing a wide variety of tasks such as anomaly detection, real-time monitoring, pattern recognition and predictive analytics.
- The defining feature of granular data is the possibility of identifying individual reporting units. While **there are serious considerations related to privacy**, central banks in isolation might not be able to process the vast quantity of granular data and benefit fully from the insights they offer. Finding the means to collaborate with other authorities, the industry and/or academia without jeopardising privacy or confidentiality requires new and innovative ways to work with and share granular data.
- A third important feature relates to the governance and standardisation of granular data. Unlike aggregate statistics, for which common frameworks are usually in place to ensure consistency and comparability, **harmonising granular data across countries is more challenging** for a variety of reasons. These include varying levels of technological infrastructure, idiosyncratic regulatory and legal frameworks, as well as different degrees of data availability and access. Given the interconnected nature of today's globalised world, developing universally adopted frameworks and norms related to granular data appears essential.
- A majority of key policy models and analyses at central banks have been built over the years based on macroeconomic statistics. Important challenges need to be addressed in order to update or provide **new models that leverage granular data**, and to reconcile granular data with their aggregate counterparts.

Looking forward, making the most of the opportunities provided by granular data calls for **developing appropriate mitigation measures** to safeguard their security, address their quality problems and ensure the usefulness of the information provided in a transparent way, especially for supporting policy decisions.

[2] Fortunately, the quality of granular data sets used by central banks (eg those derived from supervisory reporting systems, accounting data) is typically much better compared with the "average". Moreover, important projects have been undertaken since the Great Financial Crisis to make further progress, for instance in Europe with the development of the Banks' Integrated Reporting Dictionary (BIRD) and of the Analytical Credit Database (AnaCredit); see Israël and Tissot (2021). Yet dealing with micro data sets nevertheless requires important validation work, a telling example being the data collected by trade repositories (IFC (2018)). Moreover, one important challenge is to integrate/reconcile information coming from different reporting exercises, especially for statistical, prudential and resolution purposes. In Europe, and as argued by Casa, authorities have already launched strategic initiatives to rationalise, standardise and integrate existing reporting frameworks – cf the ongoing Eurosystem Integrated Reporting Framework (IReF) initiative (ECB (2022)).

1. Introduction

The defining feature of granular data is their level of detail, broadly encompassing both disaggregated data and micro data. On the one side, disaggregated data are below the level of aggregates and are more likely to identify economic agents. On the other side, micro data refer to individual reporting units (persons or entities) which can be identified (IAG (2017)).[3] Examples of granular data include social media interactions, firm-level financial statements, credit and bank transaction records, news articles, sector-specific output and productivity metrics, online job postings, and geospatial climate data.

Historically, the quantity and scope of granular data have been limited, primarily due to the labour-intensive and costly nature of their collection, processing and storage. Another reason was the primary focus on macroeconomic fluctuations and the related subdued interest devoted to micro-level developments. **Central banks have thus traditionally relied on aggregate data sources** to analyse economic conditions and make data-informed policy decisions. These sources, which are typically accurate, reliable and methodologically consistent, include statistics on national accounts, the labour market and prices. Yet aggregate data have several drawbacks, not least limited frequency, timeliness and level of details. Further, as highlighted during the Covid-19 pandemic, major disruptions to statistical production processes may hamper the supply of traditional indicators, leaving central banks in the uncomfortable spot of having to conduct policy without their usual information inputs. Expanding the data offering, particularly with alternative and more granular sources, can be essential to mitigate such risks and foster statistical resiliency (Jahangir-Abdoelrahman and Tissot (2023)).

The situation is now rapidly changing. Macroeconomic analyses have increasingly recognised the merit of incorporating micro-level insights, especially since the Great Financial Crisis of 2007–09 (GFC).[4] Moreover, **an unprecedented quantity of real-time and near real-time granular data are now available** with the ongoing data revolution, reflecting both the increasingly digitalised world and the proliferation of the Internet of Things (IoT) associated with the development of connected devices. This wealth of data presents central banks with a plethora of opportunities to foster innovation with respect to their work, as argued by Pablo García Silva (Irving Fisher Committee on Central Bank Statistics (IFC)). However, using granular data is also accompanied by new risks that central banks must now confront.

In this context, the IFC and the Bank of Canada co-organised "Granular data: new horizons and challenges for central banks" as a satellite seminar to the 64th World Statistics Congress (WSC) of the International Statistical Institute (ISI) in 2023.[5] This event was an opportunity to take a comprehensive view of the uses of granular data by central banks, highlighting their analytical benefit, the required tools and

[3] It is worth noting that while granular data and big data are related notions, they are conceptually different. Granular data are defined by the level of detail, whereas big data are characterised by particular features such as their volume, velocity and variety (IFC (2017)). Hence, granular data can form part of big data, but not all big data are granular.

[4] See for example Gabaix (2011) for the interest of using firm-level data to analyse aggregate shocks.

[5] Proceedings of ISI WSCs are available at Proceedings & Abstracts | ISI (isi-web.org).

approaches needed to unlock their full value, and the challenges associated with using granular data. The various contributions, referred to in this overview and included in this *IFC Bulletin*, discussed the general benefits provided by granular data (section 2), the opportunities offered by technical innovation (section 3) as well as the key risks faced when using granular data and the associated mitigation measures in central banks (section 4).

2. General benefits – novel insights from granular data

Granular data offer several benefits and novel insights to central banks, whose experience underscores four key areas of interest:

i. Granular data provide deeper and more flexible insights across a broad spectrum of topics.

ii. The wealth of granular data originating from the increasingly digitalised world facilitates faster collection and utilisation of statistics, with a timelier and more resilient production process.

iii. The richness of granular data along with improvements in technological capacity can yield entirely new types of economic indicators, while also allowing statisticians to refine and improve methodologies for the existing ones.

iv. Granular data can also act as a valuable complement to available statistical aggregates, helping to address issues such as real-time measurement error and the imputation of missing data.

Deeper and more flexible insights

The wealth of granular data, both in scope and detail, can support a more complete understanding of economic phenomena, with deeper and more flexible insights into emerging patterns, trends and relationships.

The first way which granular data sets can aid central banks is **the precision and nuance of the information** they offer. With traditional data, aggregation leads to a loss of information and limits the scope of questions that can be adequately answered. A major reason for this pertains to heterogeneity, especially across different sectors, geographic regions or demographic groups. For instance, a recent Bank of Canada study shows that the degree of economic uncertainty may vary significantly across sectors and can be driven by idiosyncratic developments in specific areas. Moreover, combining different individual positions into "normal" macro aggregates raises the risks of overlooking tail risks, as evidenced during the GFC. In contrast, granular data enable a much more comprehensive view of economic developments and the impacts of policy decisions, not only on macroeconomic aggregates but also on the distributions of those aggregates (Israël and Tissot (2021)). This sentiment was echoed by the former President of the European Central Bank a few years ago (Draghi (2016)):

"*Disaggregated data are indeed necessary to identify and analyse the heterogeneity that characterises the real world. For central banks this is particularly*

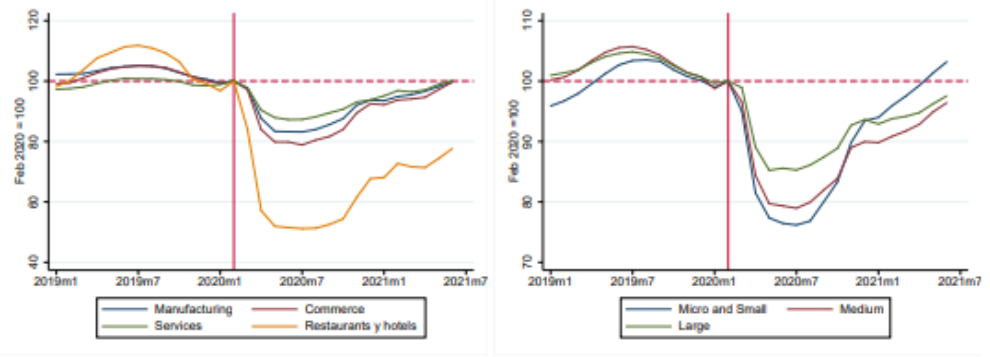
important: to implement policy in the most effective way, we need to know how our policy actions affect all sectors of the economy."

This type of nuance was also **especially helpful during the Covid-19 pandemic** in 2020–22 when the situation was shifting rapidly, making it difficult to assess the current state and future trajectory of the economy.[6] These challenges were further complicated by the fact that both the spread of the virus and associated containment measures had disproportionate impacts on different regions, sectors and demographic groups.

One telling example of how granular data could help to inform policy in this context was the **use of firm-level administrative registers** in Chile (Albagli et al (2023)). This rich data set provides precise insights on firm-level production, firm-to-firm transactions, firm-to-bank credit transactions, employer-employee interactions, and access to credit and employment support programs. This information proved key to understanding how firms' margins of adjustments (in terms of changes in output, capital, labour, intermediate inputs or productivity) responded to the pandemic as well as to better gauging firm-level heterogeneity. For instance, while employment fell in all industries during the pandemic, the decline was particularly strong for firms in food and hospitality; additionally, employment losses occurred primarily in micro, small and medium-sized enterprises (Graph 1).

Employment in Chile during the Covid-19 pandemic: evolution across sectors and types of firm

Graph 1



(a) Sector (b) Size

Note: Seasonally adjusted employment, excluding workers enrolled in the employment protection program, normalised at 100 in February 2020.

Source: Albagli et al (2023).

Such detailed insights can in turn allow policymakers to **undertake and calibrate more targeted policy responses**. Indeed, the Central Bank of Chile, for example, took a number of exceptional measures during the pandemic to specifically address the

[6] For instance, Chetty et al (2024) were able to build a publicly available database that tracks US economic activity at a granular level in real time using anonymised data from private companies.

challenges faced by credit borrowers.[7] Of note was the opening of the Conditional Financing Facility for Increased Loans (FCIC), a credit line for commercial banks with the aim of increasing lending to consumers and smaller businesses. Turning to the case of Peru, the use of entity-level data has proved helpful to analyse the monetary policy transmission channel, depending on the level of leverage of financial institutions.

Another example is the widespread use of **granular data sets covering credit and bank transactions**. These can provide high-frequency glimpses into household behaviour and spending patterns, allowing policymakers to gauge economic developments and formulate appropriate policy responses. For example, one study conducted in Italy used transaction data from the retail settlement system to produce real-time estimates and forecasts of final consumption and its components. It shows that the addition of transaction data significantly improved both nowcasting and forecasting performance, confirming other research on gross domestic product (GDP) forecasts (Aprigliano et al (2019)) and on the use of debit card transactions in Norway (Aastveit et al (2020)). Similarly, Buda et al (2022) show that transaction-level data from Spanish retail accounts can be combined according to national accounting principles to produce highly accurate real-time measures of aggregate consumption.

The **second way in which granular data can provide greater insights is the adaptability that they offer**. Such flexibility is particularly important for conducting evidence-based policies. For instance, since the impact of monetary policy decisions can have long and variable lags, it is paramount that central banks are cognisant of not just the current state of the economy, but how the economy will evolve over the subsequent months and years.

This type of adaptability was particularly useful to policymakers in Canada during the Covid-19 pandemic, when questions arose about the actual level of inflation facing households. The consumer price index (CPI) is computed using a fixed basket of goods and services, with expenditure weights typically updated every two years; however, consumption patterns shifted dramatically during the pandemic. This raised concerns that reported figures could be quite different than the inflation households were experiencing. To address this issue, the Bank of Canada and Statistics Canada partnered to construct an adjusted CPI measure based on a real-time basket of goods and services, drawing on a variety of information sources, including consumer credit card purchases, survey data and transaction data from grocery retailers. Interestingly, this alternative measure showed that, after adjusting for changes in consumption patterns, the inflation faced by consumers was only slightly higher than what was reported in the official numbers (Huynh et al (2020)).

Yet, making the most of the flexibility provided by granular data calls for **having a coherent statistical framework to ensure the correct integration of micro-level observations into macroeconomic aggregates**. It is essential to support top-down approaches that allow "zooming in" on observed macro developments. Conversely, it helps to "zoom out" and compute idiosyncratic aggregates addressing specific policy needs, concerns or research questions (IFC (2021a)). One example of this two-way micro-macro interaction is the recent project developed in Korea to use granular

[7] A full description of measures taken can be found at www.bcentral.cl/en/web/banco-central/exceptional-measures.

scanner data – that is, information collected by barcode scanners at supermarkets – to reconstruct a CPI index that better captures changes in aggregate consumer behaviour. The new index appears to provide leading insights compared with official figures and is also available at a higher (weekly) frequency. Similarly, a recent bottom-up analysis conducted at the Bank of Canada generated more detailed insights into the behaviour of Canada's unit labour cost, through the development of a coherent framework reconciling granular data with aggregate information. A last example is the growing interest central banks have in developing granular security-by-security databases that can be used to derive meaningful macro aggregates (Dilip and Tissot (2024)).

More timely and resilient sources of information

In addition to offering a level of detail and insight otherwise unattainable, **granular data allow for faster production and utilisation of statistical information**. A key reason is that many micro-level data sources now originate as a by-product of technology innovation, digitalisation and the IoT. Examples of these "financial big data" include digital footprint indicators such as e-commerce sales, Google search trends, social media interactions and mobile phone usage; signals from electronic connected "smart" devices such as mobile phones, thermostats, physical sensors and grids; financial operations such as credit card transactions, bank transfers and payments; and supply chain logistics data such as inventory levels, shipping activities and transportation networks (IFC (2021b)). Certainly, and reflecting the large differences in how these various types of data are produced and collected, the information obtained has several shortcomings compared with traditional aggregate statistics compiled under well established methodologies. Yet, this information can be both more timely and more resilient in specific circumstances, such as during the Covid-19 pandemic (Jahangir-Abdoelrahman and Tissot (2023)).

An **appealing aspect of granular data relates to timeliness**. In contrast to aggregate statistics – such as national accounts data which are computed at regular intervals (eg quarterly or annually) – micro data can be collected and compiled at a much higher frequency, in some cases on a practically continuous basis. Further, the period between collecting and processing can be nearly instantaneous, meaning that one can observe changing conditions or emerging trends in almost real time and respond to situations with greater speed and accuracy. As illustrated in the case of Banco de Portugal, this represents a key opportunity to move from traditional, "ready-made" statistical reports to dynamic, micro data-driven insights.

Another **key feature of granular data relates to resiliency**. Since in many cases the information originates as a by-product of activities, it can be more resilient to disruptions than traditional economic statistics which specifically require adequate compilation exercises (eg censuses, surveys). For example, the US federal government shutdown observed in December 2018–January 2019 because of the inability to reach a political agreement on funding coincided with a period of heightened uncertainty and financial market turbulence. It also led to delayed publication of many official statistics. These statistics were not available until around mid-February 2019, leaving little information to assess the state of the economy. This was the case, in particular, for retail sales reports, one of the most used real-time indicators of consumer spending. In contrast, a retail spending index could be constructed from transaction

data (Aladangady et al (2021)). This alternative measure was unaffected by the shutdown and provided policymakers with a lens into US consumer spending during this period, while official statistics did not. Another example of granular data resiliency was during Covid-19, as the pandemic disrupted the production of official statistics across many countries. Fortunately, statisticians could address these information gaps and related methodological and data quality issues by using a variety of "alternative" micro-level sources, including transaction data, web and social media data, and other digital footprint indicators, to name a few (de Beer and Tissot (2021)).

However, while granular data can be more resilient to disruptions in the flow of information in certain cases, one needs to be cautious about potential changes in their underlying production process and the implications for analysing them. As an illustration, large shifts were observed in payment choices at the onset Covid-19, as consumers opted to use cards much more frequently than cash. This led to a bias in household spending estimates based on card transaction data, with the need to carefully account for the shift in the card-to-cash ratio.

New economic indicators

There has been a **substantial increase in the types of economic indicators available to policymakers**, reflecting the wealth of new data generated by digitalisation as well as the greater use of "traditional" information sets (eg administrative registers, textual content) thanks to improvements in technology and methodological innovation.

Most notably, **central banks now place a large emphasis on compiling and analysing *soft indicators***, for instance agents' perceptions and expectations about the economy (Shapiro and Wilson (2017)). While variants of such sentiment, confidence and uncertainty indicators had been available prior to the recent proliferation of granular data, they were typically survey-based. This can be resource-intensive and may be of limited help in times of abrupt change. In contrast, new approaches – such as the rapidly developing field of natural language processing (NLP) tools – can facilitate the compilation of soft indicators directly from news and media. For example, the Federal Reserve Bank of San Francisco now publishes a high-frequency measure of economic sentiment, constructed by comparing a lexicon (ie a list of pre-defined words) with economics-related text collected from leading newspapers. This approach has proved quite effective in providing timely warning signals. Notably, the index exhibited a sharp decline in January 2020 at the outbreak of the Covid-19 pandemic, nearly two months earlier than survey-based sentiment measures available at that time (Buckman et al (2020)). However, lexicon-based approaches can be limited in their ability to capture context and nuance, and attention is increasingly put on using more sophisticated techniques based on ML to quantify textual information and get more subtle indications and context-specific signals (see Box A).

A second important development has been related to the **new types of indicators originating from so-called *hard granular data***. One example is the Global Supply Chain Pressure Index published by the Federal Reserve Bank of New York, which is derived from a variety of micro data sources, including information on global transportation costs, air freight cost indices and surveys related to supply chains (Benigno et al (2022)). New economic indicators can also be derived from

granular, high-resolution satellite images; for example, the Global Agriculture Monitoring project uses satellite imagery to measure crop conditions, agricultural land use and estimates of future crop yields.

Box A

Harnessing breakthroughs in language processing for compiling new economic indicators

In recent years **the field of natural language processing (NLP) has seen significant advancements**, most notably with the development of large language models (LLMs), for which a key architecture is the Bidirectional Encoder Representations from Transformers (BERT) (see Araujo et al (2023)). Using deep learning, these state-of-the-art models pre-train on a large amount of textual data with the objective of predicting missing words in sentences based on the surrounding context (for instance the words to both the left and right of the missing ones). To successfully accomplish this task, these models must learn the basic structure, rules and organisation of the language used. A major contributor to their widespread usage is that, once pre-trained, they can be fine-tuned for other purposes with a relatively small amount of task-specific data, a process known as *transfer learning*.

BERT models represent a **promising avenue for constructing soft information measures that can be useful for central banks**. One example is the recent initiative to build a quantitative measure of narrative-based monetary policy uncertainty (NMPU) by drawing on US newspaper articles published over several decades. The approach involves, first, assigning each individual article a score based on the fraction of its content expressing uncertainty and, second, compiling an aggregate measure by averaging scores at desired intervals (eg daily, weekly, monthly). This NMPU index has tended to rise in the days before announcements made at Federal Open Market Committee (FOMC) meetings and to fall afterwards – although it has remained elevated in the days following announcements that were particularly "hawkish". Moreover, the index responded to large changes in macroeconomic fundamentals such as inflation, unemployment and housing prices.

Compared with existing narrative-based uncertainty measures constructed from lexical approaches (cf Baker et al (2016)), which effectively check for the presence of the word "uncertainty" or synonyms or derivatives of it, **one advantage of NMPU-type indicators is that they incorporate the context in which "uncertainty" is expressed**, such as through questions, tentative statements or expressions of doubt. This aspect is particularly salient, as a challenge with static dictionaries is the fact that words expressing uncertainty often vary across time and publication outlets. Additionally, NMPU measures can be computed from relatively few articles and reported at a high frequency (eg daily), in contrast to lexicon-based measures that can require many articles to be aggregated meaningfully and are typically restricted to monthly frequencies.

Lastly, **an interesting aspect of the new approaches is their flexibility**. In addition to the measurement of monetary policy uncertainty, similar soft information measures could be computed in a relatively straightforward way to cover other various areas of interest to central banks, such as energy, housing and inflation.

Third, a wealth of untapped granular data sets can still be used to expand, sometimes on a very experimental basis, **the scope and variety of economic indicators that can address new information demands** depending on circumstances. As evidenced by the Covid-19 pandemic, it is not always *a priori* obvious what data will be useful in rare events such as crises. By allowing a wide range of granular data sets to be integrated into existing analytical frameworks, central banks can have more flexibility to address potential information needs that are hard to predict at the current juncture. For instance, one identified data gap in supporting climate policy analysis relates to geospatial forward-looking physical risk indicators

that are increasingly needed to assess the development of environmental hazards and their potential financial stability implications (Aurouet et al (2023)).

Complementing the offering of macroeconomic statistics

In addition to offering a novel and timely source of information, granular data can be used to **complement official macroeconomic statistics**. For one, initial releases of aggregate economic data can contain some form of measurement error. In the case of variables like GDP, for instance, preliminary releases often reflect incomplete information because of reporting lags for the various data inputs. As these inputs become available, values are corrected to reflect this updated information. But measurement error can also occur for other reasons, such as sampling errors, seasonal variation and changes in methodology, or because the construction of a variable relies on assumptions or estimation methods which are imperfect themselves.[8]

Dealing with these issues can have **important policy implications**. Measurement errors are not simply statistical anomalies which dissipate with time; they can potentially influence more lasting decisions taken based on the initial information released. A well-known example comes from transcripts of FOMC meetings in late 1992, which showed that policymakers were quite concerned about economic weakness. These worries were based on incoming data (eg on industrial production) suggesting little or no GDP growth. But these initial statistical releases contained a significant amount of measurement error, and revised data subsequently showed that the economy in that period was actually quite strong (Croushore and Stark (2001)).

A second way that granular data can be used to complement macroeconomic indicators is **by filling in information gaps**. Such gaps in aggregate official statistics can arise for a variety of reasons, such as when methodological refinements incorporate data that are not available for the entire time series history or because of disruptions in production processes. These gaps may be temporary, as was the case with the US federal government shutdown described above, or more permanent. In this second case, missing data can not only impact contemporary policy decisions but also limit the understanding of historical episodes. This clearly puts a premium on trying to fill information gaps by using the micro data sources available. One example is the recent discontinuation of the collection of accounts receivable data in the US Census Bureau's Annual Retail Trade Survey, which led to gaps in macroeconomic consumer credit indicators published by the Federal Reserve Board. To address this problem, the solution was to use another granular data set available – ie the Consumer Credit Panel data set managed by the Federal Reserve Bank of New York Center for Microeconomic Data. A second example of the use of micro data sources to fill permanent gaps in aggregated indicators relates to the various measures taken to address disruptions in official statistics caused by the Covid-19 pandemic (see Box B).

[8] The importance of these shortcomings can explain why central banks are increasingly dedicating a significant number of resources to nowcasting exercises as a complement to official statistical data releases.

3. Unlocking the value of granular data using data science

Granular data can offer fast and detailed information about economic developments, but they inherently lack the thorough production processes and quality assessments of typical aggregated statistics. Further, for granular data to provide value to central banks they must be situated within a coherent framework and contextualised so that they can ultimately inform policy decisions.

Data science can be essential to addressing these two issues.[9] First, it offers a large set of new tools and techniques (including traditional statistical analysis, various data visualisation and optimisation methods as well as AI/ML techniques) that can operationalise granular data in a systematic way, despite the challenges posed by their quantity, complexity and variety of formats. One example is the recent project using complex language models to understand the impact of economic conditions on the qualitative sentiment of industry leaders, as collected in the Bank of Canada's business outlook surveys.

Box B

Using granular data to address gaps in aggregate statistics compiled during the Covid-19 pandemic

In the course of 2020 many central banks and statistical agencies encountered disruptions to the production of official statistics, as the spread of the virus and associated containment measures made it difficult to collect data through usual channels, such as surveys and interviews. Additionally, pandemic-induced shifts in consumption behaviour led to heightened uncertainty about the reliability of the information that agencies could compile. This left authorities in the uncomfortable position of having to conduct policy absent many of their usual information inputs.

For instance, the Job Vacancy and Wage Survey (JVWS) in Canada was temporarily suspended during the second and third quarters of 2020. This survey, conducted by Statistics Canada, provides detailed information about employment conditions (eg job vacancies, wage rates and job characteristics across different industries and regions). Due to the suspension, important data gaps arose at a time when the labour market was undergoing significant changes. It was thus **essential to find complementary sources to mitigate these gaps**.

Regarding for instance information on job vacancies, one can utilise granular data from Indeed.com, a popular online job market platform in Canada that allows individuals to search for employment and employers to post job listings. Data from this platform are available in near real time at a high frequency. But one difficulty in using this source to estimate missing JVWS data is that Indeed uses its own sectoral classification standard which differs from the North American Industry Classification System (NAICS) used in the JVWS.

To overcome this issue, the approach followed by Bank of Canada researchers started with matching Indeed data with another granular data set from Advan, a company that produces geospatial data of businesses, and which also happens to contain corresponding NAICS codes. This matching exercise was achieved using an algorithm for identifying similarity between the names of the companies when they belong in the same area. The resulting data set

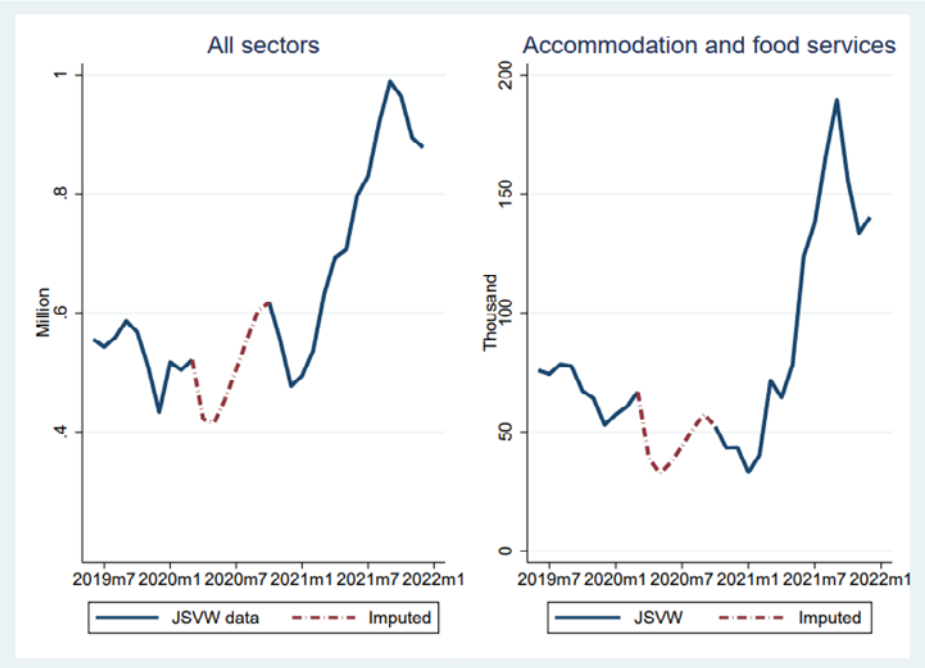
[9] The IFC has been organising recurrent data science workshops with the Bank of Italy to review developments in the big data ecosystem and the ongoing adoption of data analytics. These have provided the opportunity to review the contribution of ML applications to a wide range of use cases in central banks (IFC (2022)) and the development of related data science applications and tools, including NLP techniques and LLMs (IFC (2023a)).

yields a highly detailed measure of job vacancies at the city, provincial and national levels, and with a sectoral classification consistent with NAICS.

As a validation exercise, the job vacancies estimated from the new data set were compared with reported figures in the JVWS for periods for which data from both sources are available. The two measures appeared highly correlated, allowing for estimating a stable relationship between them and imputing missing values for job vacancies in the JVWS during the period when the survey was suspended (Graph B1).

Addressing permanent statistical gaps: imputation of missing data on job vacancies in the Canadian Job Vacancy and Wage Survey during Covid-19

Graph B1



Source: T Dahlhaus, R Ellwanger, G Galassi and P-Y Yanni (2024): "Classifying job postings into NAICS Codes", *IFC Bulletin*, no 61.

Second, data science provides new and alternative ways to derive insights that are needed to conduct evidence-based policies. From this perspective, the most recent advances in AI techniques (eg generative AI), combined with ever-increasing computing power, are widening the range of possibilities to support dealing with granular data (IFC (2020)). Since these methods are being increasingly adopted across industries and offer the potential for significant shifts in many sectors, it is important for central banks to remain agile and aware of ongoing innovations.

Data collection, management and operation

Data science provides tools and techniques which can operationalise dealing with granular data and in particular help to address the challenges they pose. These challenges include not only the large size of micro data sets, usually available in great

quantities with unchecked quality, but also their complexity due to disparate sources and a high level of details. A further problem is the variety of their formats, from large, structured databases formatted with standardised criteria to unstructured big data sets (Schubert (2021)). Examples of the former include detailed administrative databases, such as credit registers or security-by-security databases, while the latter include web and social media data, sensor data, and logs generated from software, applications and devices. And an additional issue is the difficult inferring from micro-level observations developments that are pertinent at the aggregate level.[10]

These challenges illustrate that, before granular data can be utilised, they **typically require a number of pre-processing steps**. These steps include making transformations of the raw data, correcting typos, dropping redundant information, removing duplicates and so forth (ie "cleaning" the data). Additionally, in many cases a granular data set is only one piece of a larger integrated information set, and the data need to be merged with other sources of information or aggregated to a desired level.[11] Historically, these pre-processing steps could be managed manually, but with the diversity, quantity and frequency of granular data, it has become essential to be able to automate them. Needless to say, this effort is best suited for data scientists, because of their expertise in working with large information sets across a variety of formats.

The next phase is to have an efficient pipeline to **extract from the raw data actionable knowledge that can be used to inform policy**. The goal is to establish a systematic approach, in a well documented way, to scale with the ever-growing quantity of granular data and also to be able to constantly evaluate and make improvements to production pipelines. In the event of shortcomings or problems related to the data, having such a framework allows for conducting ex post analysis of errors and taking corrective measures. Here also, data science expertise can help to develop processes in a modular way and efficiently manage a variety of granular data sources.

Finally, to provide policy insights in real time there **must be essentially zero delay** between obtaining the raw granular data and the point where they are transformed and ready for analysis. Data science expertise in code and hardware optimisation, parallel processing and other techniques promoting efficiency and speed can be critical here.[12]

Quality assessment

In addition to making granular data and their insights available to policymakers in a timely fashion, data science tools and techniques can be instrumental in **assessing**

[10] See Chodorow-Reich (2020) for an example of the difficulties faced when inferring from regional developments the aggregate impact of a macroeconomic shock, due in particular to "micro" spillovers.

[11] The need for such multi-purpose data integration is illustrated by the Eurosystem's handling of granular data on securities, loans and entities. Similarly, a recent Bank of Italy study underscores the merits and "the urgency" of integrating granular trade data and business statistics.

[12] See Cui et al (2023) for an example of the benefits of setting up a big data lake and using cloud computing tools for processing large data sets at the Bank of Canada.

the quality of the data and addressing any shortcomings. The fact that granular data are often an "organic" by-product of specific activities implies that important issues can arise with data completeness, consistency and integrity. Further, as public-facing institutions, central banks must rigorously evaluate the data and be aware of their drawbacks or limitations before integrating them as a part of their decision-making processes.

At first, data science tools and techniques have been especially useful in addressing **issues around data completeness** at central banks. For example, the use of ML at the Bank of Israel has supported the efficient classification of daily foreign exchange market transactions. The data collected can omit important details such as whether non-residents are financial or non-financial entities, or whether business sector transactions are from importers or exporters. Supervised classification algorithms were used to impute this missing information, with a reported high degree of accuracy. A similar project conducted at the Central Bank of Chile helped to address gaps in the data collected for compiling a house price index. The underlying source was the administrative data set of the tax office, which does not always distinguish between houses and apartments. After training on historical data, the tool was able to correctly classify residences with very high accuracy. A further example of the use of ML techniques relates to the Central Balance Sheet Data Office accounting data compiled by the Bank of Spain (see Box C).

In addition to data completeness, data science tools and techniques can also be used to **address problems of consistency and integrity**. Examples of such issues are selection, response or survivorship bias, measurement error, data anomalies, etc. For example, a recent project used an isolation forest ML algorithm to detect anomalies in transaction-level settlement data from Lynx, Canada's high-value payment system. These anomalies can hint at cyber events, market stress, systemic stress, fraud or operational issues, all of which are of key interest to central banks in pursuing their mandates. Consequently, detecting them as early as possible is critical. Yet one problem is that there are a large number of transactions, most of which are normal, implying that finding anomalies is tantamount to finding needles in a haystack. A layered approach was adopted to separate typical payments and analyse them for anomaly detection, with a detection rate well above existing, commonly used models.

Box C

Using machine learning to address missing data: dealing with firms' balance sheet information at the Bank of Spain

Each year the Bank of Spain receives nearly one million accounting questionnaires filled by non-financial firms. These reports contain information on firms' balance sheets, income statements, employment and wages, and other relevant financial statistics. The collected data are published in the annual Central Balance Sheet Data Office (CBSO) results. During the production process staff need to manually debug reported information, and ultimately some questionnaires have to be discarded due to missing information.

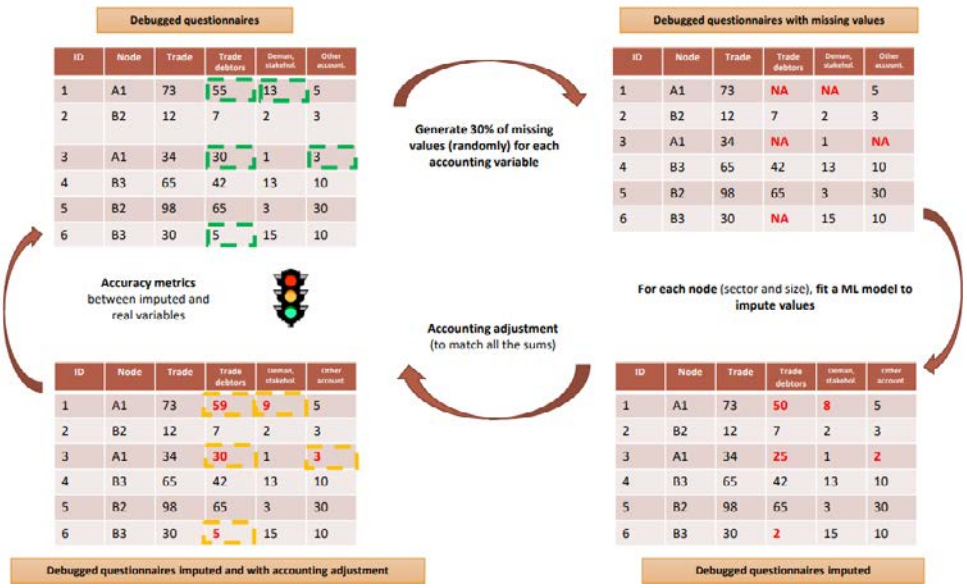
In this context, a specific project was launched **to investigate whether ML can be used to impute missing information**, allowing the Bank to enrich the CBSO database and publish more accurate statistics. The emphasis was

on sub-categories of balance sheet variables – total assets, liabilities and equity – for small and medium-sized enterprises, which are reported in questionnaires but not always with finer disaggregation, such as how total assets are distributed across inventories, cash, financial investments, etc. To impute missing information, the Multiple Imputation by Chained Equations (MICE) algorithm, a technique based on groups of random forest algorithms, was trained on received questionnaires and used to predict data from 2020. Yet one important challenge when dealing with accounting data is to ensure that subcategories can be summed to match coarser aggregates. To address this issue and ensure that the accounting decomposition holds, a scaling factor was computed based on the ratio of non-imputed to imputed values (Graph C1).

As a validation exercise, aggregates were computed from the questionnaires after imputation and compared with the initial reported data (this was done in a variety of settings, such as across different variables, firm sizes and sectors). The comparison suggested that the imputations made were highly accurate, underscoring the merit of using ML techniques to address missing information in granular data sets.

ML-based process for imputing missing values in the Bank of Spain database covering firms' financial statements

Graph C1



Source: I Crespo (2024): "Missing values imputation for Central Balance Sheet Data Office (CBSO) accounting data", *IFC Bulletin*, no 61.

A second example is the use of data science tools and techniques to **identify data tampering**. It is now well-known that a significant portion of social media activity can be driven by pre-programmed bots who post, like or re-post with the aim of spreading or amplifying misinformation or manipulating public opinion.[13] This type of activity can lead to inflated engagement metrics, making it difficult to accurately gauge public opinion. Consequently, the quality of economic indicators such as those

[13] Even in the absence of data tampering, a key issue with social media data is their representativeness, since their activities may not represent a random sample from the population of interest (Wibisono et al (2019)). Another issue relates to the (limited) availability of transparent information (eg metadata, process documentation).

measuring sentiment, confidence, expectations or uncertainty derived from social media could be adversely affected. Data science methods have proved useful in filtering out probable bot-generated behaviour, by examining a variety of profile features such as post timing, repetition, frequency and similarity with other posts, along with interaction patterns and content – a task essentially impossible for traditional data analysis. A notable example is the Botometer project from the Observatory on Social Media (OSoMe) at Indiana University. A classification system based on a random forest ML algorithm with more than 1,000 features from available metadata and information extracted from interaction patterns and content was used to identify probable bots.

New approaches and insights

Data science offers many **opportunities to central banks interested in developing a data-driven approach** to extract deeper insights from the data at hand. The main reason is that, relative to traditional techniques, data science is often better suited to incorporate a great wealth of information – by scaling with large amounts of data; handling a diverse set of formats such as text, image, voice etc; and extracting specific insights.

A main benefit is that these methods offer **tremendous ability to synthesise, summarise and present economic facts** – a key issue when dealing with the vast amounts of observations and variables contained in granular data sets. In addition, the possibility to identify complex patterns, trends and relationships is particularly salient for central banks as well as financial and prudential supervisors interested in getting insights at the level of only a small fraction of the units covered by the data – for instance, on those institutions, transactions or instruments deemed of systemic importance (Israël and Tissot (2021)). Moreover, many data science tools are tailored towards predictive analysis, which can be used to forecast key economic variables, such as GDP, inflation and unemployment (all important ingredients for conducting central bank policies). Furthermore, a notable point is that models within the realm of data science are quite heterogeneous, both in the inputs they use and the way they produce outputs. As a result, central banks can employ a variety of different models to inform policy, each of which may have their own respective benefits and drawbacks.[14] These models can also be used together, for instance for producing a single "combined" forecast instead of a series of alternative forecasts.

4. Looking ahead: addressing risks and setting up adequate mitigation measures

While granular data undoubtedly offer central banks key insights into economic and financial developments, **a number of risks and challenges** are also associated with their use. Central banks' experience underlines in particular four important aspects that deserve careful consideration and appropriate mitigation measures looking ahead, namely (i) data security and confidentiality issues, (ii) quality aspects, (iii) data-

[14] To facilitate the sharing of experience and in particular the use of ML models in economic and finance use cases, the free open-source library gingado is available on the BIS website (Araujo (2023)).

sharing and (iv) the usefulness of the information provided, especially for policy purposes.

Data security and confidentiality

A key area is security and the need to protect sensitive information (especially in terms of individual behaviours, preferences and potentially identities) from unauthorised access, corruption or even theft. Dealing with micro data clearly reinforces the risks faced in this area, reflecting four main reasons.

First, granular data sets are attractive targets for **cyber attacks and unauthorised access**. The more detailed the information, the more valuable it becomes to malicious actors seeking to exploit it for various purposes, including identity theft or financial fraud. Robust security measures must thus be implemented by organisations like central banks that collect, manage and disseminate data, not least to support their reputation and public trust and in turn the reliability of the information being reported. Hopefully, several approaches can aid in mitigating these risks, including access management, anonymisation and synthetic data and privacy enhancement techniques. The ultimate objective is to prevent unauthorised access, misuse and breaches throughout the entire data life cycle.

Despite these risks, central banks have been taking active steps to facilitate access to confidential statistical information, especially for research purposes, as documented by the ECB in the European context. One telling example relates to the Bank of Spain's project, in collaboration with the European Commission and a private software provider, to **facilitate access to synthetic versions of non-public data sets**. The approach aims to preserve the characteristics of the original data, including in terms of statistical distributions and relationships among variables, without compromising privacy. But the pilot conducted underscored important challenges, such as the effect of outliers' suppression and the limited ability for synthetic data to preserve certain linear and non-linear relationships. Another initiative is the Research Data and Service Centre (RDSC) of the Deutsche Bundesbank, which offers access to sensitive micro data for non-commercial research. A key lesson is that the set-up of robust control methods can help to reduce the risk of unauthorised disclosure, particularly when merging micro data sets. The project also underlined two important practical rules: (i) results intended for release must be based on at least five different observation units; and (ii) the combined share of the two largest observation units should not exceed 85% of the total value under analysis.

Second, while innovation can be part of the solution, it can also lead to **new, unexpected challenges, especially regarding re-identification risks** posed to granular data even when anonymised. The development of more advanced techniques and algorithms, especially the greater ability of AI models to incorporate a wide range of information sets (in particular, from additional and sometimes unexpected sources) as well as higher calculation capacity (eg quantum computing) may be a game changer from this perspective. They could for instance facilitate the re-identification of individuals by correlating seemingly anonymised data with other available sources. This can undermine existing privacy protection measures and expose individuals to unexpected scrutiny or harm.

A third, related point is that **security issues should not be solely confined to data-compiling institutions**, as they involve a wide range of stakeholders. In particular, the fact that highly sensitive information might be revealed underscores the importance of obtaining informed consent from the individuals and firms reporting their own data. They should be adequately made aware of the risks involved and also have the ability to control their records. Yet, this is particularly challenging when dealing with granular data, since it is often difficult to fully understand the implications of such detailed information – in particular regarding how it may be used by new, AI-based algorithms. Addressing these issues requires a multifaceted approach, involving state-of-the-art technical solutions and strong data governance frameworks to secure overall trust in data management processes (UNECE (2024a)). Central banks are also focusing on enhancing data literacy in the population of interest, for instance with public awareness campaigns and the promotion of responsible data practices and AI approaches.

Lastly, an essential element is to **ensure compliance with existing regulatory frameworks and ethical guidelines**. Yet, with the need to deal appropriately with granular data, privacy regulations such as the EU General Data Protection Regulation or the California Consumer Privacy Act have become increasingly complex. Organisations such as central banks must navigate a maze of legal requirements concerning data collection, processing, storage and sharing, all while ensuring that their practices align with privacy principles and rights. This has particularly important implications in terms of internal resources and skills.

Data quality

A common issue faced when working with granular information sources is their limited quality, with the term "quality" covering the various characteristics sought for in official statistics, including their accuracy and trustworthiness, integrity and security, and the need to be properly documented and easy to find and access (Križman and Tissot (2022)). In fact, granular data sets may contain errors, inconsistencies or biases that could lead to incorrect policy decisions or economic assessments. Ensuring their accuracy, reliability and robustness is therefore crucial for maintaining the effectiveness and credibility of central bank operations.

More fundamentally, the limited quality of granular sets can lead to **important inconsistencies with macro aggregates**. Discrepancies may arise for a variety of reasons, including double-counting issues or inaccurate reference data, for instance on counterparties.[15] Such inconsistencies may severely hamper the ability to use micro and macro data as interchangeable sources. A typical example is the use of security-by-security data to compile portfolio investment in the balance of payments and international investment position. Despite considerable efforts and progress, the available micro and macro sources can still exhibit some important discrepancies (Dilip and Tissot (2024)).

[15] These micro-level challenges reinforce the discrepancies already existing in macro statistics that reflect various data vintages, revision effects and different data as well as differences in interpretation and practical implementation of methodologies (ECB and Eurostat (2024)).

Mitigating these challenges calls for **setting up strong data management processes**. This is needed to achieve sound data acquisition and, more importantly, data integration processes, especially to ensure the effective linking of information sets across domains and sources – for instance by matching geospatial information with business registers and financial statements (UNECE (2024b)). They also entail the set up of plausibility checks, both internally – to check consistency based on the granular information collected – and externally – for instance to benchmark against other granular or aggregate data sources. Further, on a practical level, robust production pipelines are needed if central banks want to use granular data effectively in their operations. A key aspect relates to the automation and simplification of workflows and deployments across the statistical production chain, for instance in software development ("DevOps") and ML production pipelines ("MLOps"). Fortunately, modern engineering procedures – eg continuous integration and continuous deployment practices – can be effectively mobilised to better integrate granular sources into data management operations.

Lastly, the strengthening of granular data collection processes should be favourable for **reporting agents**, by lowering their reporting burden with more stable and resilient requirements, reducing the need for them to compile aggregated data, and in turn improving overall data quality. Indeed, the experience of the National Bank of the Republic of North Macedonia is that a key element for success is to avoid collecting the same information twice and to ensure good communication and collaboration with the reporting agents.

Data-sharing

One critical aspect of granular data is the limited possibilities for sharing them. A first issue is that, unlike aggregated data, **granular information is hardly harmonised across sectors and/or countries**, not least because entity-level records are often multidimensional and uncoded. As a result, there is no common and universal framework for dealing with granular data that would ensure consistency in definitions, classifications and methodologies. This is in sharp contrast with macroeconomic statistics, which typically rely on a common methodological framework to ensure comparability and consistency. A telling example is the System of National Accounts, which is almost universally adopted and provides a systematic way to analyse economic phenomena.

A related issue is the **limited standardisation of granular data** that prevents their adequate modelling and exchange. This calls for establishing international standards in order to develop common data structures, definitions, classifications and identifiers. More fundamentally, such standards would allow for interoperability across domains and organisations, hence promoting the findability, accessibility and reuse of granular data (UNECE (2024c)). Fortunately, important initiatives are under way to address this point. For instance, the Data Documentation Initiative (DDI) and XBRL (eXtensible Business Reporting Language) are international standards designed to, respectively, describe socioeconomic surveys, censuses and other microdata collection activities and support business reporting. Moreover, one key project endorsed by global authorities after the GFC has been the launch of the Legal Entity Identifier (LEI), a reference code to uniquely identify legally distinct entities that engage in financial transactions. Furthermore, the Statistical Data and Metadata

eXchange (SDMX) standard that supports the compilation of macro statistics has been adapted to facilitate their reconciliation with the underlying micro sources – in particular with the recent introduction of the new SDMX version 3.0 (Nikoloutsos and Sirello (2023)). Yet, the adoption of statistical standards for granular data remains extremely limited and heterogeneous. For example, the use of SDMX for working with micro data is still in infancy among central banks, mostly due to the lack of sufficient international guidelines (Bogdanova and Buffet (2023)). Other examples include the still limited coverage of the LEI and the need for a better mapping of existing micro data standards.

In addition, **several barriers prevent the sharing of granular data.** Those barriers can be technical, particularly in low- and middle-income countries, reflecting differences in technological infrastructure, capacity and statistical systems. Concrete examples include faulty repository systems, download issues and lack of harmonised software and file formats (UNSD (2023)). A perhaps more fundamental obstacle relates to the ethical challenges (eg privacy) posed by the dealing with individual records, as noted above, as well as cultural issues (eg "silo approaches"; IFC (2015)).

Policymaking institutions, in particular central banks, are however increasingly aware of the need to make progress on data-sharing, not least to better monitor evolving global trends and the risks and vulnerabilities they represent (IFC (2023b)). At the global level, critical undertakings have been initiated in this regard. First, the recommendations adopted by the G20 in the context of the third phase of the Data Gaps Initiative (DGI-3) call for the development of an international micro data-sharing standard as well as for the promotion of a sound and better access to information, including granular data, through common taxonomies (IMF et al (2023)). Second, a common task force, gathering central banks, statistical offices and international organisations, has been established to deliver guidelines and enhance the use of SDMX for micro data. Notwithstanding this progress, the adoption of a uniform approach to standardise or harmonise granular data-based statistical concepts remains a challenge, requiring strong coordinated efforts and international collaboration.

Explainability and transparency for use in policy

Properly communicating statistics to policymakers is a constant challenge (IFC (2024)). The difficulties are multiplied when dealing with granular data sets, which are typically complex, multidimensional and very large, limiting people's ability to derive clear and actionable insights out of a mass of data points.

These **issues can be reinforced by using sophisticated data techniques**. Examples include ML/AI-based approaches that can sometimes be seen as very opaque systems ("black boxes"), making it challenging to understand how decisions are made or why certain outcomes are produced. Moreover, any lack of transparency can erode trust, especially when insights drawn from the data can significantly impact individuals' lives, such as in healthcare, finance or criminal justice. For central banks, as well as more generally for policy institutions and the public sector, traceability and explainability therefore sit at the core of their decision-making processes. Fortunately, a number of approaches have been developed to demonstrate how granular information has been used to get some insights and support certain decisions. As illustrated by the recent Bank of Spain project related to bitcoin, in particular, the use

of a SHAP (SHapley Additive exPlanations; cf Lundberg and Lee (2017)) value estimation method can help to increase transparency and interpretability of ML models.

A further complication is that **granular data can inadvertently perpetuate biases and discrimination**, for instance if the data used to train AI systems reflect societal biases or contain inaccuracies. Moreover, the ongoing avalanche of (micro-level) information, combined with the limited traceability of data techniques and the increased attention to "alternative facts" driving good statistics out of policy debates, raises the risk of eroding public knowledge (IFC (2024); Garrett (2024)). Maintaining fairness, equity and transparency in decision-making processes is thus essential. The key is to recognise that the new techniques need to be properly managed to avoid these issues, such as algorithmic bias, and prevent unfair outcomes for certain demographic groups and discrimination. One solution put forward by Haghighi and Taillefer is to develop a dedicated data and analytics risk management framework in central banking. More generally, and as argued by Pablo García Silva, Vice-Governor of the Central Bank of Chile, the challenges above call for using the new data available in today's society in a smart way, by exercising discernment when generating and disseminating them to public authorities.

References

Aastveit, K, T Fastbø, E Granziera, K Paulsen and K Torstensen (2020): "Nowcasting Norwegian household consumption with debit card transaction data", *Norges Bank Working Paper*, no 17.

Aladangady, A, S Aron-Dine, W Dunn, L Feiveson, P Lengermann and C Sahm (2021): "From transaction data to economic statistics: constructing real-time, high-frequency, geographic measures of consumer spending", in K Abraham, R Jarmin, B Moyer and M Shaprio (eds), *Big data for twenty-first-century economic statistics,* National Bureau of Economic Research, pp 115–145.

Albagli, E, A Fernández, J Guerra-Salas, F Huneeus and P Muñoz (2023): "Anatomy of firms' margins of adjustment: evidence from the Covid pandemic", *Central Bank of Chile Working Papers*, no 981.

Aprigliano, V, G Ardizzi and L Monteforte (2019): "Using payment system data to forecast economic activity", *International Journal of Central Banking*, vol 15, no 4, pp 55–80.

Araujo, D (2023): "gingado: a machine learning library focused on economics and finance", *BIS Working Papers*, no 1122.

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2023): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59.

Aurouet, D, D Del Giudice, A De Sanctis, J Franke, J Herzberg, M Osiewicz, R Peronaci and C Willeke (2023): "Indicators of granular exposures to climate-related physical risks for central banks' analytical purposes", in S Arslanalp, K Kostial and G Quirós-Romero (eds), *Data for a greener world: a guide for practitioners and policymakers*, IMF, chapter 4, pp 59–77.

Baker, S, N Bloom and S Davis (2016): "Measuring Economic Policy Uncertainty", *Quarterly Journal of Economics*, vol 131, no 4, pp. 1593–636.

Benigno, G, J di Giovanni, J Groen and A Noble (2022): "A new barometer of global supply chain pressures", *Federal Reserve Bank of New York Liberty Street Economics*, 4 January.

Bogdanova, B and B Buffet (2023): "Towards a full integration of SDMX in the data value chain – the present and the future of SDMX tools", presentation at the 2023 SDMX Global Conference, November.

Buckman, S, A Shapiro, M Sudhof and D Wilson (2020): "News sentiment in the time of Covid-19", *Federal Reserve Bank of San Francisco Economic Letter*, no 2020-08, 6 April.

Buda, G, V Carvalho, S Hansen, J Mora, Á Ortiz and T Rodrigo (2022): "National accounts in a world of naturally occurring data: a proof of concept for consumption", *Cambridge Working Papers in Economics*, no 2244.

Chetty, R, J Friedman, M Stepner and Opportunity Insights Team (2024): "The economic impacts of Covid-19: evidence from a new public database built using private sector data", *Quarterly Journal of Economics*, vol 139, no 2, pp 829–89.

Chodorow-Reich, G (2020): "Regional data in macroeconomics: some advice for practitioners", *Journal of Economic Dynamics and Control*, vol 115, 103875.

Croushore, D and T Stark (2001): "A real-time data set for macroeconomists", *Journal of Econometrics,* vol 105, no 1, pp. 111–30.

Cui, M, G Jiang and B Mosweu (2023): "Swimming in the data lake: an application to NielsenIQ Homescan", *IFC Bulletin*, no 59.

De Beer, B and B Tissot (2021): "Official statistics in the wake of the Covid-19 pandemic: a central banking perspective", *Theoretical Economics Letters*, vol 11, no 4, August.

Dilip, A and B Tissot (2024): "Development and maintenance of a security-by-security database", *IFC Guidance Note*, no 5, May.

Draghi, M (2016): "Central bank statistics: moving beyond the aggregates", speech at the Eighth ECB Statistics Conference, Frankfurt am Main, 6 July.

European Central Bank (ECB) (2022): *The Eurosystem Integrated Reporting Framework: an overview*, September.

ECB and Eurostat (2024): *ESS-ESCB Quality assessment report on statistics underlying the Macroeconomic Imbalance Procedure*, July.

Gabaix, X (2011): "The granular origins of aggregate fluctuations", *Econometrica*, vol 79, no 3, pp 733–72.

Garrett, A (2024): "The devil, the detail, and the data", *Journal of the Royal Statistical Society*, Series A: statistics in society, qnae063.

Huynh, K, H Lao, P Sabourin and A Welte (2020): "What do high-frequency expenditure network data reveal about spending and inflation during Covid-19?", *Bank of Canada Staff Analytical Notes*, no 2020-20.

Inter-Agency Group on Economic and Financial Statistics (IAG) (2017): *Update on the Data Gaps Initiative and the outcome of the Workshop on Data Sharing*, March.

International Monetary Fund (IMF), Inter-Agency Group on Economic and Financial Statistics and Financial Stability Board Secretariat (2023): *G20 Data Gaps Initiative 3 workplan: people, planet, economy – delivering insights for action*, March.

Irving Fisher Committee on Central Bank Statistics (IFC) (2015): "Data-sharing: issues and good practices", *IFC Report*, no 1.

——— (2017): "Big data", *IFC Bulletin*, no 44.

——— (2018): "Central banks and trade repositories derivatives data", *IFC Report*, no 7.

——— (2020): "Computing platforms for big data analytics and artificial intelligence", *IFC Report*, no 11.

——— (2021a): "Micro data for the macro world", *IFC Bulletin*, no 53.

——— (2021b): "Use of big data sources and applications at central banks", *IFC Report*, no 13.

——— (2022): "Machine learning in central banking", *IFC Bulletin*, no 57.

——— (2023a): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59.

——— (2023b): "Data-sharing practices", *IFC Guidance Note*, no 3, March.

——— (2024): "Communication on central bank statistics", *IFC Bulletin*, no 60.

Israël, J-M and B Tissot (2021): "Incorporating microdata into macro policy decision-making", *Journal of Digital Banking*, vol 6, no 3, pp 233–50.

Jahangir-Abdoelrahman, S and Tissot, B (2023): "The post-pandemic new normal for central bank statistics", *Statistical Journal of the IAOS*, vol 39, no 3, pp 559–72.

Križman, I and B Tissot (2022): "Data governance frameworks for official statistics and the integration of alternative sources", *Statistical Journal of the IAOS*, vol 38, no 3, pp 947–55.

Lundberg, S and S Lee (2017): "A unified approach to interpreting model predictions", in *Advances in neural information processing systems* 30 (NIPS).

Nikoloutsos, S and O Sirello (2023): "SDMX, an international standard for micro data", presentation at the 2023 SDMX Global Conference, November.

Schubert, A (2021): "Big data for policymaking in economics and finance: the potential and challenges", in P Nymand-Andersen (ed), *Data science in economics and finance for decision makers*, pp 2–3.

Shapiro, A and D Wilson (2017): "What's in the news? A new economic indicator", *Federal Reserve Bank of San Francisco Economic Letter*, no 2017-10, 10 April.

United Nations Economic Commission for Europe (UNECE) (2024a): *Data stewardship and the role of national statistical offices in the new data ecosystem*, ECE/CES/STAT/2023/4.

——— (2024b): *In-depth review of linking data across domains and sources*, Conference of European Statistics, ECE/CES/2024/5.

——— (2024c): *Data governance framework for statistical interoperability*, High-Level Group for the Modernisation of Official Statistics, March.

United Nations Statistics Division (UNSD) (2023): *Microdata dissemination: Summary report*, Microdata Dissemination Task Force of the Inter-Secretariat Working Group on Household Surveys, February.

Wibisono, O, H Ari, B Tissot, A Widjanarti and A Zulen (2019): "Using big data analytics and artificial intelligence: a central banking perspective", *Capco Institute Journal of Financial Transformation*, 50th edition, "Data analytics", pp 70–83.