American Economic Association

Committee on Economic Statistics: www.aeaweb.org/about-aea/committees/economic-statistics

Updating National Economic Statistics with Real Time Data:

Challenges and Opportunities

Summary of an AEAStat Working Session held January 8, 2022, at the
2022 American Social Science Association meetings

Premise: Digitization has revolutionized how businesses run, and businesses now collect unprecedented amounts of information and data, during the course of ordinary operations. This revolution has created amazing opportunities to modernize our national statistics using real time, fine-grained, digital data.

Objectives: Discussion among 28 experts from data companies, academic departments of economics and data science, U.S. statistical agencies, and nonprofit institutes and organizations was aimed at understanding the opportunities that exist and working towards concrete actions to capitalize on those opportunities. Identifying the advantages, challenges, and alternatives to a strawman proposal to do this was aimed at stimulating useful and creative discussion.

The Working Proposal to Stimulate Discussion: The working session discussed a proposal focused on Retail Trade but its approach could potentially be extended to other sectors. Three observations motivate the proposal:

First, the size distribution of firm in any industry tends to follow a power law. National Retail Federation (NRF) data shows that the top 100 retailers had $2.3 trillion in sales accounting for about 62% of total sales. Collecting data from these companies was seen as a realistic goal.

Second, almost all of these firms maintain their data in data lakes or on the cloud.

Third, using modern distributed computing in the cloud, many statistics of interest to Federal Agencies would be computable at relatively low cost. This could include high frequency (daily, weekly) retail sales estimates on a nation-wide and by geography. Agencies would likely be able to reasonably impute the remaining sales from modeling using lower frequency and less granular sources. Additional statistics (e.g., high frequency inflation rates) could be computed using this method. Over time, the platform could be expanded to include more participants.

There are no major *technical* impediments to this approach, which is described further in the attached proposal summary. Session participants did identify unique opportunities, a number of technical and nontechnical impediments, and issues relating to whether resultant statistics were fit for all uses.

Identified Opportunities, Challenges and Associated Issues:

1. **Opportunities**

    a. There is mounting evidence that high-frequency indicators can be predictive of things like retail sales. For example, the high-frequency retail sales measure CARTS (https://chicagofed.org/carts) maintained at the Chicago Fed outperformed consensus forecasts for the advance retail sales data from the Census Bureau by almost 50% over the last two years.
    b. The proposal is revolutionary in that it would productively connect the companies and statistical agencies.
    c. Highlights of the proposal are that it would increase data frequency, reduce data latency, and facilitate disaggregation of data.
    d. Covering the largest firms will, on average, provide a good view of aggregate spending, particularly in highly concentrated industries; and indeed, even the official statistics suffer from limited coverage of small units (and young units).
    e. Inventory investment is one of our less-well-measured statistics for National Income and Product Account purposes. It appears that this proposal could work on the inventory investment margin, a boon to those who follow inventories.
    f. The more data sources from which we draw, the better our conclusions will be.
    g. The "Notebook" data harvesting model is appealing. Statistical agencies could provide notebooks that include definitions of the data they require, which should be much more efficient and less costly for the firms' providing data than respond to multiple surveys on "paper." It also helps address some concerns regarding security, personally identifiable information and latency.

2. **Measurement Limitations**

    a. Economic turning points may often have a disproportionate role for small or young firms, and also for entering or exiting firms or establishments. Focusing on the largest firms presents a challenge in that the small/entrant/exiter margin is actually a place where the marginal value of improvements to existing official data is relatively high.
    b. Data lakes are over-written, meaning that historical patterns would be missed without some other record keeping.
    c. Making such high frequency statistics compatible with official statistics is challenge.
    d. Evaluating the costs and benefits of daily vs weekly vs monthly data needs to be done.

3. **Costs**

    a. Some suggested that the working proposal would be very costly to implement especially in terms of integrating with official statistics.
    b. Engineers must play an outsized role in scaling up and implementing the proposed approach. Most statistical agencies do not employ a cadre of engineers who could collaborate with private firms to implement the proposal. This raises questions about the cost of implementation.
    c. The statistical agencies already have been employing aggregated private firm data, which has been costly to procure. One limitation is that given limitations on data sharing, there is not a "one stop" acquisition approach currently.

d. There is likely heterogeneity amongst even the top 100 retailers in terms of costs and capabilities.
e. It is unclear how costs would be distributed between the firms and the statistical agencies.

4. Legislative Authority

a. Statistical agencies see nothing in the legal environment that would preclude the approach.
b. It was suggested that the Foundations for Evidence Based Policy Making Act of 2018 (the "Evidence Act") would provide all necessary authority.
c. But private firms report difficulty working with agencies' Offices of the General Counsel to confirm existing legal authority, a time-consuming effort that could become less burdensome with additional experience.
d. A standard legal structure should be put together for application across companies and statistical agencies. This would avoid the need for case-by-case negotiation.

5. Incentives for the Private Sector

a. The right incentives need to be in place for data providers to be fully collaborative and committed to accuracy.
b. If it's extremely low cost for companies and the information is not sensitive, they'll do it. If it is costly, it's a real problem.
c. The Future of Privacy Forum did a survey of CEO's and found that. They are willing to contribute data but: a) want legal cover; b) hesitate to be first in their sector; and c) want credit for helping to solve important national problems.
d. Appeals for collaboration might focus on the benefits to business of better information on business cycles.
e. Enthusiasts need to think about articulation of a convincing "give-get" proposition. For example, statistical agencies have universal frames that could be useful to companies. (But, see "Ethical Issues", below)
f. Statistical agencies could construct valuable "sample weights" unique to contributing organizations, helping their ability to understand how to make samples more representative of the broader economy
g. If the companies had some say in the type of data products that come out such that the research data product would be of interest to them, that could be a nice incentive.
h. For some data and some firms, there may be an efficiency/cost reduction argument. The way the Jobs and Employment Data Exchange has aligned incentives with employers providing data is that it reduces current reporting burden (e.g., state UI reporting) by automating data sharing in a more modern way like this. Is there a potential analog of this cost reduction argument for retailers?

6. Data Quality and Consistency

a. Data quality / integrity need more attention in the proposal. One way this has been tackled in other data sharing projects we've seen is by adding data quality checks in the notebook.
b. High frequency data across potential providers is unlikely to be uniform. What adjustments would be needed as a result? And by whom?
c. Approaches are not in place to gauge or standardize data quality across private providers.
d. The more data sources from which we draw, the better our conclusions will be. But it is noted that capabilities differ from firm to firm.
e. We need to be able to keep track of—and fully understand—non-economic movements in a company's data resulting from M&A activity, special company-specific sales incentives, etc., and have a way to adjust for those.
f. It is also important to note that under the proposal, approved statistical agency staff would need to be able to get behind corporate firewalls, test their analytical notebooks, ask questions, and ensure that they are measuring what they think they are measuring.

7. Leadership and Governance

a. Some participants felt that academic leadership was preferred to statistical agency leadership of an overall effort to implement this or a similar proposal. Others felt that since a primary goal is to enhance statistical data, the statistical agencies should be in charge. Still others questioned whether academics or statistical agency leadership had the appropriate engineering background to lead such an effort.
b. In any case, the effort must be multi-agency.
c. Success will depend on support from firms' CEOs.
d. In any case, legal, policy, and data governance minutiae are huge barriers.

8. Ethical Considerations

a. Ethics should be a major concern in considering statistical measurement using free but highly valuable private data.
b. Because much of the data being considered can be market-moving, there is the possibility of collusion among some firms to manipulate critical economic measures for personal gain. (As, for example, the scandal in which bankers at several major financial institutions colluded with each other to manipulate the London InterBank Offered Rate, a benchmark interest rate for global markets.)
c. The privacy and confidentiality of contributors to federal statistical surveys are vigorously protected. But companies might be tempted by PR/marketing opportunities to tout their specific contributions to linked data. How do we balance this?
d. To what people and how (e.g., tiered) data resulting from public-private permutations will be released deserves thought. Would general availability create problematic behavior change?

Potential Next Steps to Advance a Proposal

- **Announcements**

A joint announcement that the firms, statistical agencies, and institutions representing research statisticians/economists (like AEA) have agreed to collaborate would create a social contract of great promise.

A useful promise associated with the announced collaboration would be the importance of the transparency of the effort, and equality over exclusivity.

Issue a press release that identifies a strategy to communicate with CEOs about the benefits of using private data to enhance Federal statistics.

- **A Research Project**

Several hypotheses can be derived from these comments and tested in a transparent manner with active collaboration among public and private sector participants.

Seek potential funders for a well-articulated and well justified research project.

- **A Demonstration Project**

A good starting point for projects like this is a "minimum viable coalition" of companies sufficient to demonstrate the viability of the project. Doesn't have to be large, but it helps to have a variety of company types.

Recent work on alternative data governance models for public-private data sharing (data collaboratives, data trusts, data cooperatives) have shown promise in aligning both the legal and business considerations. Looking into these should be part of the design process. See: https://foundation.mozilla.org/en/data-futures-lab/data-for-empowerment/shifting-power-through-data-governance/#what-is-a-data-cooperative

A key piece needed for providing clear data provenance across contributors, and a bit of pipeline transparency, is adding the use of Open Lineage to project architecture.

- **A Federal Budget Initiative**

It would be useful to bring the Federal Administration into the discussion and proposal of actions. One effective way of doing that is to propose and find enthusiastic Congressional sponsors for a federal budget proposal that would result in effective use of high frequency, private data to enhance Federal statistics for use in evidence-based policy.

- **Enlist the power of the White House to bring CEOs and statistical and policy representatives of the Federal government together. A convening event could be used to obtain buy-in and issue an announcement of plans to move forward with the proposal.**