Gaurav Bhattcharya, Devanshi Singh

Streamlining Employment Data in India

Through Reliable Survey Methods

I. Introduction

Emerging economies are facing a considerable drop in their consumption and savings pattern worldwide (World Economic Outlook, 2017). This slowdown has affected the South Asian trends too. The average production in manufacturing and the workforce has witnessed a gradual shift in the sector with a greater number of employments generated in the services and less labor-intensive industries. In the context of India, the economy has witnessed a sustained growth of 7% over the last decade, without enough employment creation (World Bank, 2018). We have seen an increased demand for jobs; more people are getting added to the workforce annually without a balanced creation of new economic opportunities. The stagnation in creation of new jobs has majorly impacted the labor-intensive sectors like manufacturing and construction. The employment sector has also witnessed more informalization of the workforce with 93% workforce employed in informal units. Apart from the critical socio-economic challenges of job creation, the employment sector has also faced issues with having robust data collection methods and use of reliable survey instruments. The official statistical agencies are in dire need of restructuring and revamping, given the reliability of data is always questioned.

Any sound policy intervention in the employment sector needs the support of good quality institutions, assigned with the task of collecting and disseminating good quality high-frequency data. If we want to fix our jobs problem, good quality data is essential for devising targeted interventions and smart solution strategies. If we analyze the jobs data in India, there is undoubtedly a need to improve the quality and frequency of data which can resolve the debate around the employment-unemployment metrics through empirical evidence. The jobs data needs to be captured frequently to establish long term trend lines on jobs for better policy intervention. As far as the institutions are concerned, the institutions must be made independent and free from any political interference to conduct unbiased surveys and data reports. However, the recent order of restructuring the statistical system with the National Statistics Office (NSO), National Sample Survey Office (NSSO) and Central Statistics Office (CSO) becoming a part of the ministry and NSSO and CSO merging to form NSO may impact its autonomy and neutrality. To make the claims justified, further evidence of this change is required.

II. Drawbacks in Survey Instruments

Measuring employment is inherently tricky in India majorly due to an absence of uniformity in survey methodologies and definitions. For example, due to different methods followed, the comparability between Employment Unemployment Survey (EUS) and Periodic Labour Force Survey (PLFS) is debated. Moreover, 93% of the employment sector is not adequately represented in these employment surveys because of being operationally informal. The Government of India (GoI) has dedicated institutions to conduct various rounds of surveys like the EUS by the Labor Bureau, PLFS and Quarterly Employment Survey by the NSSO, monthly data on subscriber additions to the Employees' Provident Fund Organization , independent studies on the economy like Centre for Monitoring of Indian Economy, State of working India report by Azim Premji University and the Confederation of Indian Industry. Despite a considerable amount of resources spent on these surveys, the estimations on employment-unemployment have been confusing and diverging. These are mainly because of the absence of non-standardized formula for calculating employment in India, non- uniform definitions on employment metrics, and lack of high-quality data streaming techniques in the employment sector. These gaps are very evident in results of Census 2011, estimating the employment as 11.18%, while the Labor Bureau's EUS 2015-16 puts the unemployment rate at 5%, and Centre for Monitoring of Indian Economy put it at 6.68% (Sep-Dec 2018). Some of the divergences accounted for different methodologies and sample size, but the figures indeed fail to provide a clear picture of the employment sector in the country.

All the existing surveys for employment-unemployment in India have some drawbacks within the current framework like lack of coverage for the informal economy, conventional survey methods being lengthy and resource intensive and sampling and non-sampling errors impacting the quality of data. For example, the NSSO samples the population once in five years for employment information, and there is no way to determine if the year chosen is appropriate or wrong. Hence, there is a chance of sampling error due to faulty selection of the period of the survey. NSSO for its PLFS uses two sets of questions based on Usual Status and Current Weekly Status, not considering the current daily status of the workers. If the data has selection bias, the sample data gathered and prepared for modeling may have characteristics not representative of the right, future population of cases. There are many non-sampling errors that impacts the quality of data like non-response, respondents under-reporting the events, incorrect recording of information by the interviewer, variability in response, errors arising from questionnaire design and skipping of questions to avoid answering a few sections which can be due to both the respondent or the interviewer.

III. Paradata: Application and Optimization

To accommodate the changing needs of the economy, NSSO introduced few changes to Periodic Labor Force Survey like Computer Assisted Personal Interviews (CAPI). CAPI is using a tablet for interviewing by the interviewer or the respondent. It was done to speed up data collection and processing to reduce time lags. CAPI enables generation of a large amount of real-time, low cost process data which can be optimized to reduce errors in survey methods. Paradata have grown over time because of an increase in technological capabilities

and new devices coming in. There are three types of paradata: Direct like Contact-info, Device-type paradata and, Questionnaire navigation paradata, Indirect like Video and Voice Recording, Behavioral coding, and Eye tracking. However, the paradata generated through CAPI has not been able to be optimized in India.

The paradata can be processed for various applications like making the survey design responsive, reducing the non-response errors, skipping of questions and interviewer bias. These are major non-sampling issues within the current survey methods framework, which distorts the quality of data. The response times and keystroke measures generated can be used to study aspects of the response process. It can be used optimally to locate errors and devise targeted interventions, two phase sampling, and reduce process errors. It can further be used to make changes to the survey design during data collection in order to optimize the trade-off between costs, quality and time (i.e. responsive design). But such applications would require perquisites like paradata collection should be made goal driven, and selection of process variables should be done appropriately. An archive or centralized data cloud should be created for better data management and two-phase sampling. Statistical disclosure control techniques and the use of data enclaves to safeguard the identities of respondents and interviewers should be adopted for the responsive design of surveys.

IV. Shift from CAPI to CATI

Further, a shift can be made to CATI gradually. Labor Bureau conducts surveys through rotational panel design where the same sample is interviewed 3-4 times for data collection. If we switch to CATI eventually, face-to-face surveys for the households can be conducted in the first round followed by CATI for subsequent interactions, at least for a limited set of core questions. In due course, we can also explore the option of completing surveys electronically for a subset of respondents. It will help bring out estimates of some key variables almost in real-time and make the process less resource intensive. For example, employment data collection in the US through the Current Population Survey (monthly survey of 60,000 households). The first interview is conducted face to face using CAPI with subsequent taking place over phones (CATI). Around 70% of the interview is done through phones. However, there is a massive limitation of survey samples being extremely large in India viz-a-viz the USA. The problem of interviewer bias and variability using CAPI and CATI can persist. Such biases can be mitigated through Machine Learning techniques like supervised learning using interview recordings as the labeled data and unsupervised learning through data clustering methods to determine interviewers' bias and variability. With the digital economy gaining momentum in India, gradually introducing the concepts like CAPI and CATI for employment surveys, labeling of phone recordings data and clustering techniques to improve the survey framework, and mitigating sampling and non-sampling errors and biases can become our long-term goal. It would require an increased allocation of funding for an elaborate physical and cyber-security systems (Demand for Grants 2019-2020, Revised Estimates 2018-19 for Census Survey and Statistics: 904 Cr, and Budget Estimates 2019-2020: 1170 Cr).

Case models across the world can be explored to understand how paradata can be generated and optimized using CAPI, CATI, and electronic surveys. For example, Amazon Mechanical Turks incentivize people to take

computer-assisted surveys. It has been revolutionizing the social science research where MTurk has been used for research on various subjects, and respondents consider it as an additional source of income, as reported by a study. Replicating the model to India could be in the form of computer assisted surveys (CAPI, CATI and eventually electronic surveys) for employment units, and financially incentivizing the employees to take the survey and generate data points. The data points thus made can be used for data analysis and constructing various indicators.

However, optimizing the paradata comes with its own sets of challenges like high costs of analyzing and managing paradata, ethical issues of privacy, and cloud being subjected to cyber risks for sensitive data. ML techniques for mitigating biases are resource and infrastructure intensive with again some bias due to limited human interference. The data can get complicated and messy, especially when combined from many systems. Moreover, a lot of time can be spent accessing and manipulating the paradata to produce datasets that are useful for the data users. Machine Learning techniques of supervised and unsupervised learning face challenges like supervised learning works only with labeled data (to be done by a person). One person labeling the data can be biased, which makes group labeling a better option and most cases group labeling is hugely resource intensive.

V. Conclusion

Optimizing the use of technology in survey methods can improve the process of data collection and processing. Particularly in the case of the employment sector, the US through CAPI and CATI has reformed its data collection methods. Learning from such good practices, India can improve the existing surveying framework and improve data collection and coverage, to obtain timely reliable and relevant labor market data to understand our employment situation. The current debate on job crisis and unemployment is majorly burdened with an absence of accurate and reliable data record. For the employment sector, we need to conduct surveys which are more inclusive and effectively utilize the data generated through technology. Further, institutional and legislative changes to ensure autonomy of statistical bodies, physical and digital infrastructure to accommodate latest technological advances and allocation of additional financial resources are necessary for the useful streamlining and optimization of jobs data in India.