The 2nd T. N. Srinivasan Memorial Lecture

"Data in Coronavirus Times"

by Dr Pronab Sen

India Policy Forum

July 13, 2020

[Video of the Lecture]

Let me begin by saying how honoured I am to have been asked to deliver the second TN Srinivasan Memorial Lecture. TN was a person with whom my relationship started off on a rather sour note. It happened when I was doing my PhD, and my last chapter, which was the empirical chapter, was essentially looking at the 1966 Devaluation in India. At that time the seminal work on that subject was the book by Bhagwati and Srinivasan, and my chapter essentially was tearing their analysis apart. Now it just so happened that my PhD supervisor, Professor Bela Balassa, who was a close friend of TN, sent him the draft chapter and then over the next couple of months, we had some very acrimonious exchange over correspondence, and remember, those were pre-email days. Nevertheless, the PhD did get through, so it couldn't have been quite as bad as TN made it out to be.

We met face to face the first time in 1996, when I took over as the Advisor to the Perspective Planning Division in the Planning Commission, and TN had come by, and he dropped in to see this person who had taken over this position. He remembered who I was, and said, "Aren't you so and so?" So I said, "Yes I am." And he was angry at that stage, but TN was not a person who could remain angry for very long. He may have had one of the most acerbic tongues in our profession, but he also had one of the warmest hearts. He was a man who was extremely loving, very generous, and over the next two decades plus, we became very close friends and there was nobody I was really quite as close to as I was to him. So this is, in a sense, a tribute to his memory and the detailed discussions that I had with him, both as an economist and even more so as a statistician.

When I took over as the Chief Statistician of India, TN very rightly called me up and said, “What on earth are you doing there? You know nothing about statistics!” I said, “Yes, of course I don't, but then you know nothing about Economics either!” So we were kind of even on that. Nevertheless, this is a monument to his memory and I hope I can do justice to it.

Although the topic is on data, the fact of the matter is that anybody who deals with data knows that you cannot talk about data in the abstract. Data has to be rooted either in a problem or in theory, or preferably in both. So, when one is talking about data in times of Coronavirus, one has to really think about it in terms of what is the requirement of data at times like this, and only then will one be able to get down to the kind of data that one needs, and whether or not this data would be generated under normal circumstances. That, in a sense, is the theme of this lecture; the theme is really about the symbiotic relationship between data and theory.

As far as data itself is concerned, I think people make the error of thinking that data is just a set of numbers, and what I'd like to make very clear is that the data is not just a set of numbers. Data is a set of numbers which have some structure to them so that they can be used in a particular manner, rather than being treated as a purely random set of numbers coming from who knows where. This is unfortunately something that people tend to forget. And what does seem to happen, more often than not, is that because numbers are there, as I think somebody has pointed out during the felicitation of Abhijit and Esther, you run millions of regressions until you find something that breaks, and you say, “Voila! You've got a result and therefore, this data is good.”

The real issue is understanding the data, and this is where a lot of problems exist among the people who generate the data. So, for instance, if you look at the kind of metadata that exists even in India, and India is one of the better ones, the metadata leaves much to be desired, because it gives you only a very basic feel for what the data is meant to be, and what it is capturing, and that frankly is simply not good enough.

There are three kinds of data that one really needs to think about because the data emanates from the need and the use for it. The *first*, of course, is that it's required for certain very specific public purposes. So, for instance, in the Indian case, a lot of the data, indeed the bulk of the data that we have generated over time, have been essentially for two purposes: one is to construct the national income accounts, and the second is to be able to do planning. Now in the course of constructing these two end-products, and therefore collecting the data to be able to do so, the data actually started throwing up patterns, and people started using the same data for other purposes. Therefore, the whole poverty literature really emerged from data which was not collected for poverty at all, but was collected essentially to construct the demand side of the planning model. And these kinds of externalities from data, which is collected for a specific purpose but used for other purposes, are very common, and are really the stuff of what research should be made of.

But there is another dimension, which is when such data is collected for a particular purpose and has been used for a different purpose, it should be remembered that in order to actually meet the need of the new purpose to which the data is being put, the data itself may need to change. That's number one. And if the data needs to change, at some point one has to either take a decision as to whether the original purpose can be served by the new data, or if not, would we have to ditch the old objective and tailor-make the data to the new project? The third way, of course, is to collect supplementary data which would fill those gaps. Now these are

decisions and these are difficult decisions. But what it does require, and this is what I think people are less aware of, is that at the end of the day, when data is being collected, other than the kind of data that is collected by J-PAL, for instance, which is primary data from the field, the people who actually organise the data collection, people who design the data collection processes, have to be more than familiar with the theory and the use and the manner in which the data would be used. Unfortunately a lot of this isn't done.

Having spent a fair amount of time now, nearly a decade, being at the head of the statistical system in India, it was fairly apparent to me that the degree of communication between the data users and the data generators was less than adequate. Yes, for every major data set in India, we do set up expert committees in which there are people who are working in the field. But the problem, of course, is that committees of this kind actually simply design the data collection format but not necessarily transfer that knowledge to people who will have to take it forward within the statistical system itself. This is a perennial problem, and it's a problem which I bring up very specifically because it was something that TN kept coming back to over and over again. TN was a very active participant in a lot of these discussions, and one of the points he kept making is that "You know you guys have forgotten that situations have changed, the economic environment has changed, the political environment has changed, but your data is unchanged, and that change can only happen when the people who are charged with collecting the data are given enough information about the theory and the application."

The *second* issue is: What does data in the time of COVID really mean? If you think about it, a lot of this kind of stuff uses COVID as an exemplar of all crises, and the question that is asked is: Is the data appropriate for crises in general? But the fact of the matter is that crises are not similar. There's a whole range of crises that happen; COVID is only one particular type, although it's an extreme example. But we need to be aware that when we are designing data systems, we are designing data collection institutions. They have to be configured in a manner in which they can provide the information that is required for crises as and when they occur. For instance, one of the things that we talk about quite a lot are shocks, which are either external shocks like the Global Financial Crisis of 2008, or they are internal shocks like famines, droughts, or COVID. Do we have systems which can pick up the effects of these different kinds of shocks?

But then there are other shocks, and sometimes they become much more difficult to identify. Again, an obvious question is the nature of the shock. I mean, is the shock coming from the demand side or is the shock coming from the supply side, because it does make a difference. There is a feeling that you know at the end of the day, a supply shock will eventually morph into a demand problem, and therefore, we can ignore the supply part of it in its entirety and just focus on the demand component. That is simply wrong. So, it is important to be able to identify the nature of the shock and the consequences, and this is where the real development of data systems becomes important, and I'll come to that.

The *third* issue is the degree of uncertainty. There are some shocks where the uncertainty is actually quite low. So if you think about things like monsoon failures or floods or a major earthquake, they're almost one-off events and they rarely have lingering effects. You can actually measure it quite precisely and you can actually predict what will happen down the road.

There are other shocks where the degrees of uncertainty are very large. COVID is an exemplar of that. The uncertainty that COVID poses for us today is that the progress of the economy is linked to the progress of a disease about which we know very little. So, we had a 40-day lockdown, then we started to unlock, but various States started locking down again. So what we have is a country in which bits and pieces of the country are at various stages of lockdown, and therefore, the nature of the economic problem that one is facing is changing, and changing very fast. With these kinds of uncertainty, the data collection system has to be able to address this uncertainty, and that again is something we will talk about. But let's get to something which is a little more fundamental, because I think this really arises from the nature of discourse that we are seeing in the country. Three points seem important.

The *first* is: What are the data on which we should be focusing? There is for some reason an undue focus on growth rates. The fact of the matter is that while growth rates are certainly important, what is happening to levels sometimes is far more important. One may say "but growth rates can always be used to recalculate what the levels are". But the way one thinks about progress, the dynamics of a system, matters very much whether you're measuring changes or you're measuring levels.

The *second* is the ability to identify turning points. There is again, I think, a rather pervasive view that economic systems exhibit only a single turning point. So you have a shock and the economy goes downwards, then you see a small turnaround and then the economy just continues to grow from thereon. But that's wrong, economic systems quite often, indeed maybe more often than not, display multiple turning points, and one should be able to not just measure but should be able to accommodate multiple turning points when one is really looking at the relationship between data and turning points.

This then brings us to the *third*, critical question, and that is the nature of the data which would inform these two major issues that I've been talking about, which is: Are we talking about levels or growth rates; and are we talking about single turning points or multiple turning points? And that is really about assessing the relative importance of cross-sectional versus panel data. In the Indian context, because of the needs that were driving the Indian statistical system, we have been excessively focused on cross-sectional data. Most of the data that we collect is really about the measurement of levels. There are some non-cross-sectional data, and most of these really come from the various indices that we have, such as the Wholesale Price Index, or the Consumer Price Index, or the Index of Industrial Production, and so on. These are actually panel data, although we don't describe them as such. The question is: Do we have the balance right, or do we have the balance wrong?

Again, as I said, these are situational, but what a crisis does is, it brings to the fore the fact that situations can change very quickly when the economy is on the trajectory with little ups and downs along the way, and you're essentially working towards planning for the future. You know, in proper cross-sectional data, measuring levels is really of the highest importance, but during times of crisis, levels are actually fundamentally unknown, what is important is the direction of change, how things are changing and are we being able to pick up the change properly, which is why we are today in a position where the numbers that are being used are essentially of the few panel data points that we have available to us. So we are looking at things like the IIP, we are looking at what's happening in the WPI, corporate sales, and so on. The problem or the issue here, as I said, takes us back to the theory and the application. What is the data being used for, or is meant to be used for?

By and large, one can sub-classify the kind of models, and the theoretical framework that one has, into models that are essential for policy purposes, and those that are used for business institutions. The models for policy purposes in India, oddly enough, for the longest time, have simply not been dynamic models. Of late, there have been efforts at creating dynamic models, but they are few, and they still haven't gained the degree of popularity that they should. I think, mainly because a lot of people simply don't understand what these models do. But the thing is that, as I said, when you're in a crisis, what is really of help is a dynamic system, and in the absence of dynamic models, or dynamic modules, or policymakers who want dynamic models, you will not get the data that will enable you to be able to create an information set that you need in terms of businesses. That's where we seem to be at this particular point.

In a sense, what has also happened is that when we talk about policy, in particular, and policy at times of crisis, we again talk in terms of monetary policy, on one side, and fiscal policy, as if they were completely compartmentalised. And the experience that we have had in terms of practically every crisis that I can think of over the last 20 years is that the response time in terms of monetary policy has been much faster than of fiscal policy. One of the reasons for that is because the kind of information that monetary policy requires is far more amenable to high-frequency data than what fiscal policy does. But when you have a situation as we have today, where a supply shock is now well on its way to morphing into a demand constraint, the importance of policy interventions will have to switch from the monetary to the fiscal.

The real question then is: first, do the fiscal authorities have the information that they need to be able to take decisions? Second, even if the data does exist, do the models exist that will enable the policymakers to decide what to do? If the latter don't exist, then the former will also get reduced in importance, and that I think has already started.

Insofar as data for business decisions are concerned, this becomes far more difficult because companies, businesses of all kinds, regardless of size, need to have a feel for what is happening to the markets. Now, one could say that this is something that companies should do for themselves. Under normal circumstances, this is certainly true; but under extraordinary circumstances, such as when you have a crisis, they have absolutely no idea what is happening to markets, which makes business decisions far more difficult. Therefore, it raises the degree of uncertainty among businesses, and thereby affects the process of adjustment and recovery.

We are seeing this happening today—the data that is available to businesses simply does not give them the information they need to be able to take a reasoned decision: Now that the lockdown has been lifted, should we restart production or no? If the entire business community gets into a 'wait and watch' mode, we will be putting ourselves into a self-fulfilling prophecy situation where, because everybody is behaving as a herd even if circumstances are propitious for a recovery, people are not being able to take appropriate positions, they are holding themselves back.

The question is: How do we meet this kind of requirement? Clearly, the Government cannot go down to the level of granularity that an individual business can; but what one would need is to be able to give them enough granularity to be able to take a call whether those product categories into which they fall are showing signs of recovery or not.

This then brings me to something that has been discussed, and this is a rather critical component of Abhijit and Esther's work, which is the importance of distributions. One of the problems that we have, and Rakesh Mohan talked about it, is in development macroeconomics. Macroeconomics is actually extremely sensitive to distribution and distributional changes. By and large, in the absence of major shocks, these changes are slow, and if you keep re-estimating your macro equations every so often, you'll pick up some of these effects, but during times of shock, these distributions can change rapidly.

We have just been through one particular example of a major shock, which led to significant distributional changes and, therefore, in the macro behaviour of the economy. This was the demonetisation episode. Everybody should know that demonetisation hit the weaker sections of the population and the small businesses disproportionately hard, and that large businesses actually benefited. Now this kind of a distributional change will alter practically every macro parameter in any sensible macroeconomic model, whereas what we tend to do is to keep the parameters constant when we do macro modelling, and we look at only changes in the variables. But when distributions change, the parameters of these models need to be changed, and changed right away before we can get a real feel of how systems react.

This is where I think the type of work that Abhijit and Esther are doing is so important because they are saying that behavioural differences across different categories are important and need to be taken into account. This has been interpreted, by and large, as saying that specific policy interventions, and specific programmatic intervention towards attaining particular objectives such as health objectives or educational objectives, need to take this into account. Let me submit to you that while that is true, it affects the behaviour of the macroeconomy as well, and any major shock, and any major crisis, will change these parameters fairly quickly, and we do not, as of now, know how and the manner in which we can take this into account.

The assessment of parametric changes requires us to be able to know what is really the distribution of these parameters among different categories of people, whether by income classes or by regions or whatever the distribution variables are that you wish to consider; but income classes and regions would be the bare minimum, and then to be able to take a call on fairly limited data as to how the weighted averages will change given the nature of the shock.

So, think of what we are talking about in terms of COVID. We have particular parts of the country where the disease burden has been more pronounced and other parts of the country where it has not. Therefore, the aggregate, overall weighted average is going to be very different than what it was in prior times, in normal times, and if we did have the information as to what the parameters were by regions, we would be able to have a better sense of what these parametric values are.

Similarly with income distributions. Again we know that because of the nature of the shock, because of the degree of staying power of different production systems, it is the small units, the micro and small units that will be disproportionately affected. We know what the production contributions are, but we really have absolutely no idea as to what effect this is going to have on the consumption function; and the consumption function is at the heart of any macroeconomic model. Similarly, we really do not know how the investment demand function is going to change. Now, to be able to assess this, we need to go far deeper than we do so at present. It is not as if the data doesn't exist, and this is really what brings me to my final set of issues.

The data does exist, it exists up to a point, but the fact of the matter is, access to this data, or even the knowledge that the data exists, is simply not there. So, if you really think, and by the way this is not criticism, I personally have been a party to it, if you really think about how national income accounts are built up, there's a huge volume of data that is needed, is collected, and is compiled. What most of us have access to is only the compiled version. Compiled versions by their very nature give you aggregates and not necessarily the kind of granularity that you may need to be able to make adjustments during the time of a crisis. But the data is there. That's number one.

Number two is on the use of administrative data. Again, because of our historic focus on levels and level estimates, we have, I think, a tendency to decry the use of administrative data sets, and the reason is fairly clear. Most of these data sets are collected by government agencies for essentially monitoring and reporting purposes, and, therefore, are subject to very strong biases. So even the ministries that collect these data don't trust the data themselves, and more often than not, a lot of our survey work is done essentially to validate or invalidate the administrative data.

But on the other hand, what we forget is that administrative data, for all its faults, all its incentive problems, is usually panel data. Now, if you know only 80 per cent of India is covered by our health system, but if you want say that the number of people being covered by the health system has gone up by this percentage point, for this administrative data may actually be quite useful. But the point is that somebody needs to have worked with administrative data on this before the crisis and to have established this measure.

We have been talking quite a bit about the use of administrative data for general statistical purposes and I think, quite rightly, there has been a huge amount of resistance from the statistical system itself, essentially because of the incentive problems that are inherent in most administrative data. But this is a situation where we find that non-administrative data will simply not give us the information that we need to be able to react reasonably efficiently and fast, so we have to use administrative data.

The point is: Will we take this lesson to heart? Will we actually start working on seeing the degree to which administrative data allows us to measure change because that data is, by and large, real-time? If we can establish this, we would be in a much better position to handle a crisis like the current one in future. But if we don't, then another crisis of a similar type or even of much less severity will catch us unawares again, and again we will be flailing around.

In particular, if you're going to focus on economics and economic consequences, there is one data set which is very valuable and is not being used in the manner it should be, and that is the GST database. A lot of people are doing the next best thing though. One thing that is pretty much available is the information on e-way-bills, that's pretty much public data and a lot of people are using it, but the GST data itself gives you all the granularity you want. It is granular in terms of products and activities, it is granular in terms of geography. It is incomplete because there's a lot of things that are not covered. But today, if I were to try and assess the extent to which economic recovery has started in different parts of the country, then the one data set which would give me that information with a high degree of accuracy would be the GST. But unfortunately, and sad to say, that it is not in the public domain, and there doesn't seem to be very much chance that it will be so in the near future, unless

sufficient public pressure builds up so that confidentiality can be maintained without the loss of granularity.

The final word on all of this, what a crisis, and particularly a crisis of this magnitude brings out, is the kind of things that we development economists have been doing for many years, looking at the broad sweep of development experiences, is good, is required. The theory is still not fully established, but developing countries that have low levels of resilience, low levels of being able to cope with crises, require data that will enable governments to take decisions on a 'here and now' basis. That, I would submit to you, would not be best done through cross-sectional information. For such decision making in crises, we will have to blend panels with cross-sectional data, and we will have to use administrative data much more rationally and much more thoroughly than we do now.

Thank you!