

Challenges of Using Administrative Data for Statistical Purposes : India Country Paper

**Pronab Sen
Chief Statistician of India**

1. Background

The Indian Statistical System is highly decentralized in that much of the responsibility for generating statistics for a wide variety of indicators is vested in the concerned administrative Ministry at the central level and the equivalent department at the State. Consequently, it becomes extremely important to be able to define precisely what is meant by the use of administrative data for statistical purposes. At one level, administrative data can be defined as any data generated by any government agency other than the National Statistical Office and its counterparts at the state level. This would of course be wrong in the Indian case since the statistical responsibility is decentralized and the statistical units within each concerned line Ministry should be considered as an integral component of the National Statistical System.

An alternative way of defining administrative data as distinct from statistical data is: when the specific identity of the respondent or data source is central to the use of the data. Statistical data, on the other hand, would be those where such identity is in itself of no consequence. In other words, the distinction between statistical and administrative data is based on the focus on the general as against the particular.

If administrative data is defined in this manner, there are three broad categories of purposes for which administrative data are collected. These are: (a) monitoring of government programmes and other forms of

government intervention; (b) enabling regulatory activities and audit actions; and (c) targeting outcomes of government interventions.

In the Indian context, a very large volume of data is generated by administrative ministries and state governments for each of these purposes. The purpose of this Paper is to reflect briefly upon the usability of such information for statistical purposes, whether by the National Statistical Office or by the statistical arms of the administrative ministries and state government themselves.

2. Main Uses of Administrative data for Statistical Purposes

As in every other country, India makes extensive use of administrative records and other forms of administrative data for developing the sample frames for a wide range of activities. However, in addition, there is considerable use of administrative data for direct statistical purposes as well. The earliest and perhaps the most important form of administrative record use in Indian statistics is the land-use data that is generated on a regular basis by the land revenue administration of State Governments. In earlier times, the land use records not only captured the area under various crops, but also the output of each crop. Over time, however, there were questions raised regarding the accuracy of the crop estimates obtained from visual inspection carried out by land revenue officials. Consequently, crop cutting experiments were introduced to measure the yield of each crop. Nevertheless, the land use records continue to be central to the whole process of agricultural production estimates. In the first instance, they are used as sample frames in order to determine where crop cutting experiments would take place. But they are also used as a basic statistical input since production estimates are derived as the product of

the yield given by the crop cutting experiments and the area under a particular crop as measured by the land-use records.

Given the centrality of agriculture in the Indian context, administrative record keeping in agriculture goes beyond merely production estimates. For many years there has been a regular system of recording prices of agricultural commodity at the market yards and also the quantum of market arrivals of agricultural produce. These are central to the existing system of recording agriculture output and consumption in value terms for the country.

In the case of the industrial sector, registrations under the Factories Act provide the basic frame for all industrial estimation, whether it is done through the Annual Survey of Industries or the Index of Industrial Production. Even the wholesale price index relies heavily upon this framework. One of the residual legacies of the industrial licensing system in India is the existence of a number of industry specific associations which have for many years carried out the responsibility of collating production and price data for their particular industrial sector. These should be treated as administrative data since they are mandated by industry-specific legislation and they continue to be important sources of data for production estimates, although their price data is never used. This is particularly important for the mining sector and for heavy and basic industries.

In recent years, the emergence of regulators in some areas of industry has led to a new source of administrative data i.e. regulatory records. These have been found particularly useful in areas such as pharmaceuticals and power. The role of industry associations and regulatory authorities has been especially important in a number of commercial service sectors. Indeed, practically all the data for the

banking and financial sectors, including insurance, is derived almost exclusively from regulatory demands. This is also true of practically the entire transport sector including road, air and marine transport. Since the privatization of the telecom sector from mid 90s, the primary data source has shifted from returns from the public sector companies to regulatory data provided by the Telecom Regulatory Authority of India. It is believed that as the process of private participation in a number of other sectors accelerates, new regulatory authorities will have to be created and regulatory records will become increasingly important as source of statistical data.

Although India has a long history of labour regulation and regular reporting of data from establishments of above a particular size, and it also has a fairly wide network of employment exchanges spread throughout the country, neither of these two sources of administrative data has been found to be particularly useful in monitoring employment trends in the country. This is partly because of the fact that India has a very large proportion of its non-agriculture labour force in the unorganized sector, specially in own account enterprises, which do not come under any labour regulation. Nevertheless, even for the organized sectors of the economy, these data sources have not been found to be particularly accurate.

In so far as the social sectors are concerned, the quality of administrative data in both the health and education sectors have dropped alarmingly. In the past, hospitals and public health care system records and their counterparts in the public education system could provide fairly comprehensive information on these two sectors. However, with increasing private participation, especially in a situation of under-regulation the coverage has dropped to such an extent that such data cannot be relied upon except perhaps as sentinel indicators. More recent

public interventions, such as the Central Government's operation of the Integrated Child Development Scheme and the Government aided local schemes, have opened up new sources of administrative data, but their reliability is yet to be established.

The situation is considerably better in terms of environment statistics, almost all of which are obtained from administrative data with a reasonably high degree of confidence. Nevertheless, there are clear weaknesses present in some key areas such as air and water quality.

3. Issues in Use of Administrative Data

One of the advantages of a decentralized statistical system is that the Ministries and Departments which are important contributors to statistical information are empowered with statistically qualified personnel. As a consequence, there is usually little problem in adopting and conforming to international standards and using statistical methods. The real issue in this context is not so much the understanding of the concepts and definitions, but the appropriate methodologies that need to be used in collecting data in a highly heterogeneous country. This is one of the key roles that is played by the Central Statistical Organisation.

There are two main issues that create problems in more extensive use of administrative records in the Indian context. The first is that there is often a divergence between the nature of data required for administrative purposes, especially when the objective is to monitor programmes, and the nature of the data that would be required for statistical purposes. Since the data collection machinery is generally under administrative control of the programme authorities rather than the statistical authority, the introduction of appropriate questions and indicators quite often becomes a victim of the need to keep the record

keeping process manageable. In recognition of this issue, recently the Indian Government has raised the status of the statistical officers in the line Ministries significantly and, hopefully, over time their voice would be heard more prominently while designing the administrative data collection system. A similar effort is also underway to persuade the State Governments to give more emphasis on statistical components of administrative records.

The second problem relates to the accuracy of the data. Although the completeness of coverage is frequently an issue, there is really no insurmountable problem in using the data if certain statistical corrections can be made. Inaccuracy, however, can render the data completely useless for statistical purposes. By and large, it has been found that in situations where the data is collected for monitoring programmes, the quality of the data becomes highly questionable. This is a particular problem in the social sectors and also in data that is collected by the taxation authorities. On the other hand, when the purpose is mainly for regulatory oversight the quality of data tends to be high.

4. Improving Usability of Administrative Data

In the Indian context, given the size and diversity of the country and the limitations the reach of the Government, administrative records will always tend to be incomplete. The classic case of this is the coverage of the civil registration system, which is so low that it cannot be used for measuring demographic parameters between the census years. Many of these problems can be tackled through cross checks and corrections made through survey data. For instance, in the case of civil registration, India operates a sample registration system which provides reasonable estimates of demographic indicators for the inter-census period.

Similarly, surveys carried out by the National Statistical Survey Organisation also provide important cross checks on a variety of statistical indicators.

Unfortunately, in a number of cases the survey data has supplanted the administrative data as the primary source of statistical information. Although the reasons are obvious, this is in not a particularly desirable state of affairs since it pits the statistical agency against the administrative mechanism, rather than the two working as partners. It is also much more expensive. A more sensible system would involve the use of limited surveys based on strong statistical principles to provide validation and corrective factors for the data generated on a regular basis through administrative accounts. In an important sense this would be akin to implementing a sample audit system, where perhaps the purpose would not be to find fault but to provide information which would be used to correct the inherent biases that may occur in administrative record keeping.

In the final analysis, however, the main factors governing the usability of administrative data for statistical purposes are the legal framework underpinning the data collection activity and the political importance attached to the government interventions concerned. And of course governance too.