सत्यमेव जयते

Ministry of Statistics & ProgrammeImplementation

March 2021

Administrative Data: Issues, Concerns and Prospects

An Indian Perspective

Ministry of Statistics &Programme Implementation
Policy Implementation & Monitoring Division (PIMD)
Sardar Patel Bhawan, 4th Floor
SansadMarg, New Delhi – 110001
Ph: (011) 23341867

Discussion Paper #2

Discussion Paper

Administrative Data: Issues, Concerns & Prospects

An Indian Perspective

सत्यमेव जयते

Government of India
Ministry of Statistics &ProgrammeImplementation
Policy Implementation & Monitoring Division (PIMD)
Sardar Patel Bhawan, 4th Floor
SansadMarg, New Delhi – 110001
Ph: (011) 23341867
http://www.mospi.gov.in

This page has been left blank intentionally

Table of Contents

1. Prelude

Administrative data are becoming increasingly important in today's governance framework. They are typically the spinoff of some operational exercise and are often viewed as having significant advantages over alternative sources of data. Although there is no denial to the fact that such data have merits, users should approach the analysis and place reliance on such data with great vigilante. Besides, they should be handled with the same level of prudence and critic as they approach the analysis of data based on the their collection from other sources. The paper identifies some statistical challenges in administrative data, with an objective of invigorating public discourse about the scope of administrative data usages, expanding the sweep of their analysis, and vitalizing stakeholders including researchers to traverse through some of the contemporary statistical problems which essentially comes to the fore with use of such data.

2. Background

In the production of official statistics, for a certain occurrence, the source of availability of data may be either through the statistical surveys or through the administrative processes. In these day and age, the amalgamation of these two sources is a propitious and ingenious blue print which affects the quality and quantity of research and increases the potential of data (Künn, 2015). However, this usage is often accompanied by serious challenges, given the simple fact that the purpose of designing the two sources is different. Administrative data are defined as data sets collected by government institutions or agencies for tax, benefit or public administration purposes (UNECE, 2011). According to Penneck (2007) surveys differ from administrative data in the sense that they are specifically designed for analytical purposes, so coverage of population, definitions, methodology and time can be designed to meet these analytic needs. However, the sample size might be a problem if it is small since large-scale surveys are expensive and small-scale surveys have limited use. Samples are also subject to errors and non-response bias. In addition, observes Penneck, we cannot be sure of the precision of survey responses, compared, for example, with the administrative data collected for tax purposes. Administrative systems also require data from individuals, but the latter time and again is dovetailed as a vital component of the administrative

process rather than as an additional statistical burden.

As has been noted above, Administrative data are data generated during the course of some operation, and then retained in a database. They are becoming increasingly important as the possible insights from such sources of data is being recognized and also since the alternative sources of data become more costly or difficult to use (e.g. because of declining response rates in surveys). Generally, this means that the analysis of administrative data is secondary—the data are being remodeled—although this is not always so. The existence of large, often administrative data sets, offering potential for secondary analysis, was one of the primary drivers behind the development of data mining technology (Handet al., 2000) as well as the modern rise of interest in 'big data'. But the analysis of administrative data poses new statistical challenges. This can be appreciated by a casual scrutiny of the examples in most basic statistics texts, which will involve 'random samples' in almost all experiments: administrative data are, by definition, typically not random samples. The aim of this paper is to explore these statistical challenges and to stimulate discussion. The hope is that it will help to focus attention on what is needed for valid and accurate analysis of administrative data. The need is illustrated by the comment made by Wallgren and Wallgren (2014), on the closely related topic of analysing data from statistical registers:

> *'Although register-based statistics are a common form of statistics used for official statistics and business reports, no well-established theory in the field exists. There are no recognised terms or principles, which makes the development of register-based statistics and register-statistical methodology all the more difficult. As a consequence ad hoc methods are used instead of methods based on a generally accepted theory.'*

There are many definitions of statistics. This is because the discipline has various aspects, including the study of methods for collecting, presenting, interpreting and analysing data, but also because it involves expertise in coping with uncertainty and chance. The diversity in the words of Hand (2008) has been captured as "*Statistics is the technology of extracting meaning from data and of handling uncertainty*". There are fewer definitions of administrative data. The Organisation for Economic Co-operation and Development (Organisation for Economic Co-operation and Development, 2016) defined administrative data as having the following features:

(a) The agent that supplies the data to the statistical agency and the unit to which the data relate are usually

different, in contrast with most statistical surveys;

(b) The data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;

(c) Complete coverage of the target population is the aim;

(d) Control of the methods by which the administrative data are collected and processed rests with the administrative agency.

The definition continues by saying that *'In most cases it is normal to accept (and expect) that the administrative agency will be a government unit that is responsible for implementing an administrative regulation'*.

Instead, although accepting that the features described above do characterize administrative data, it is worthwhile considering Nordbotten (2010) who plainly makes a straightforward divergence between statistical data and administrative data. Statistical data are collected primarily for statistical purposes—e.g. to summarize in order to shed light on the system generating the data, or to make predictions. In contrast, administrative data are initially collected for some administrative purpose—to run an organization, such as a company, government, charity, school, hospital, and so on. Running the organization might require on-going operational analysis of the data but, once collected and stored, the data can later be analysed to shed light on what has happened, to help to predict what might happen in the future, and to evaluate systems and their performance, i.e. the data can later be subjected to statistical analysis. Often statistical data consist of mere samples from the universe of possible values which could have been obtained, and these may have been collected by surveys or experiments for example. In contrast, administrative data will ideally consist of data on all of the cases, records or transactions in some population. This leads to evolving of a metaphysical distinction: sample data are used to obtain estimates of a population parameter. In contrast, administrative data are summarized to obtain a descriptive feature of the population.

Transaction data are an important kind of administrative data concerned with events, typically with sequences of events. Usually the prime operational purpose of collecting the data is to inform the transaction (e.g. to decide how much to charge a customer or to decide how much tax someone should pay), but once collected the data can be retained in a database and analysed to improve understanding of the organization's operations.

At first glance, though appearances can be deceptive, administrative

data appear to have several advantages compared with statistical data.

(a) Since the data have already been collected, no additional cost appears to be incurred in collecting them.

(b) In a sense, we might reasonably expect that 'all' the data are available. After all, the data generators will certainly process and can retain details of all its transactions.

(c) The data might be of high quality, since the effectiveness of the operation of the organization depends on this.

(d) The stored data will certainly be timely and might be regarded as up to date as it is possible to achieve, since they describe the organization as it is, or at least as it was when the last change was made. This advantage is strikingly illustrated in the use of administrative data to derive estimates of population attributes at times that are intermediate between decadal censuses, and in essentially real time estimates of price inflation.

(e) In a real sense administrative data often tell us what people are and what they do, not what they say they are and what they claim to do. We might thus argue that such data get us closer to social reality than do survey data.

(f) Administrative data may provide tighter definitions than alternative sources of data. Wallgren and Wallgren (2014), gave examples of data about income and children in families. Where the time restrictions on eliciting responses to a survey might mean one must simply ask 'what is your yearly income before tax?', administrative data might, depending on the source of the data, specify whether this means 'disposable income, taxable income, earned income or income including unearned income.

Unfortunately, although all of those advantages of administrative data might apply in an ideal world, in practice things are typically not so straightforward. Regarding (a), effort will normally be required to extract the data, to clean them and possibly to link them to other datasets. Moreover, although data may be free for the organization which collected them, other organizations which wish to use these data may have to pay—and the cost must be balanced against that of data from alternative sources, such as surveys or administrative data collected by other organizations. Regarding (b), data will usually enter a database via a complex social process—the sample of records in a database may not be representative of the population to which one wishes to make an inference. An operational database might not have a form which is convenient for statistical analysis exercises. In particular, different parts of an organization might use different database systems—indeed, there is a large

amount of current activity as organizations seek to put all of their data into a single data repository (a data warehouse, for example).

It is worth noting that it is sometimes useful to distinguish between two kinds of administrative data. The first kind is that which is necessarily collected during the course of some operation. Credit card transaction data, for example, necessarily involve the recording of the amount spent, the currency and the business where the transaction occurs, since these items of information are needed to run the credit card operation. The second kind is additional information which is not needed for an operation, but which is helpful for other reasons, and which is collected during the administrative process. The age and gender of a customer might fall into this category: a product might be bought by anyone, but it could be useful to analyse the customer's details later to inform new marketing strategies. In some sense this second kind of data lies between administrative and statistical: they are collected for statistical rather than operational purposes, but they are collected during and as part of the administrative process. There is an important lesson to be taken from this: benefits can be gained from involving the statisticians and data analysts in the data collection stage. This is not a new lesson: we recall Ronald Fisher's comment that

> *'to call in the statistician after the experiment is done may be no more than asking him to perform apost-mortem examination: he may be able to say what the experiment died of'*

This last point leads into the modern world of so-called 'big data'. The term has no universally accepted definition, but for this paper this may be defined as the result of some automatic data collection system. Indeed, many including Hand (2008) have argued elsewhere that the data revolution is not so much a consequence of the size of modern data sets and the ability to store them (big data) but rather of the fact that data are nowadays largely collected automatically without requiring explicit human effort. Examples of automatically collected data are everywhere and include personal health data collected by wrist monitors, automated monitoring of tickets as people travel through a rail network, telemetry of engine functioning and recording of metadata of phone calls. Data arising from the so-called 'Internet of things' would clearly be of this type.

Sen (2009) observed that 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. He notes further that in the Indian context, a very large volume of data is generated by administrative ministries and state governments for each of these purposes.

3. Administrative Statistical Systems in India

Administrative Statistics is generally collected by State Governments; consisting of statutory administrative returns and data derived as a by-product of general administration. The Rangarajan Commission (2001), while dealing with the Administrative Statistical System observed as under:

> *The system of administrative information system, whose essential purpose was to aid the Government Departments in the execution of their functions of implementation of different Acts, Rules and Regulations of Governments. Even when such Acts were passed by the Central Government, their implementation was decentralised through the State Government Departments and their district or other sub-offices. The statistics thus had a direct purpose of being not only of interest to but also necessary for the working of the departments. The regularity, quality, and completeness in the collection of these statistics, interwoven with the working of these departments, were thus indirectly ensured. The quality of this system is thus directly related to the interest the administrative departments take in it and the effective use they make of it. It is however a fact that strictness in the administrative functions of several departments of most State Governments is waning, resulting in a virtual neglect of the information system.*

There are other advantages too in the system of collection of statistics through the administrative set up. The collection of data by departmental agencies does not involve special costs. The collection is oriented to definite purposes and the record and verification of information is part of administration. Departmental agencies and officials have not only good knowledge of the subject, but also of local language and local conditions, especially rural. Information collected is relevant and direct, and the respondents do not have to make calculations before answering a query. It is handled by agencies that have special knowledge of the subject. Finally, there is an identifiable purpose in their data collection and they are in the best position to interpret the data. All this has lent a solid foundation to the decentralised administrative statistical system, and in turn, to the Indian Statistical System. An impression is carried by many that data collected by substantive Government departments are likely

to suffer from bias. Therefore, they suggest that an independent agency should collect data to ensure objectivity. But, ignorance should not pass off as objectivity, making the solution worse than the problem. While the impression might be true for certain departments at certain times, it is easy to overstress the point as a justification for the solution suggested.

4. Statistical Usages of Administrative Data in India

Sen (2009) has presented the detailed account of the usage of the administrative data for the statistical purposes in India. He has noted that 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. 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, especially 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.

5. Types of Administrative Sources

The potential range of administrative sources that could be used for statistical purposes is large and growing. The following list is not meant to be exhaustive; instead it aims to show range and types of potential data sources, as the final step towards arriving at an operational definition of administrative sources[1].

- Tax data
  - Personal income tax
  - Value Added Tax (VAT)
  - Business / profits tax
  - Property taxes
  - Import / export duties
- Social security data
  - Contributions
  - Benefits
  - Pensions
- Health / education records
- Registration systems for persons / businesses / property / vehicles
- Identity cards / passports / driving licenses
- Electoral registers
- Register of farms
- Local council registers
- Building permits
- Licensing systems e.g. television, sale of restricted goods
- Published business accounts
- Internal accounting data held by businesses
- Private businesses with data holdings:
  - Credit agencies
  - Business analysts
  - Utility companies
  - Telephone directories
  - Retailers with store cards etc.

6. Issues and Concerns in Administrative Data

Sen (2009) opined that 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

[1] As described in the Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices, UN Document (2011)

introduction of appropriate questions and indicators quite often becomes a victim of the need to keep the 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.

The major issues, concerns and challenges of the administrative data may be listed as below:

6.1. Population Coverage Issues

A system of administrative records defines the population covered by legislation based on the scope of the program intended for registration. This population is often limited by specific demographic or economic characteristics. According to Johnson and Moore (2008), in some cases individuals may need to undertake some actions to become part of the administrative system. It is therefore important, say the authors, to consider what encourages individual units to be part of schemes. There may be some favouring factors for some individuals to avoid registration, especially if their circumstances place them close to a threshold that requires obligatory participation or gets associated with financial costs. Another factor is the change of policies that may fluctuate the population taken in study from year to year.

The study unit needed for statistical purposes often focuses on the characteristics of groups formed by units (e.g. enterprises operating in a particular activity or large enterprises), while the administrative data focus on identifying specific units so that based on their individual characteristics (e.g. full-time

employees) certain actions can be undertaken. Thus, the differences in the entity reported in the tax statements limit the usefulness of the data for some types of research.

6.2. Content Issues

Johnson and Moore list several content issues that need to be considered while working with administrative data. One of them is the purpose for which administrative data are collected, which may have a significant impact on their usefulness for statistical purposes regarding the amount of available data, data definitions, consistency between different time periods and data quality. Many a times the usefulness of administrative registry systems is limited, because for example, only those variables needed to administer the tax and tax schemes are collected. These variables can only be a small part of the data reported in an administrative form of compilation/reporting. In addition, because program requirements are defined by legislation, the concepts and definitions of variables used to meet program needs do not necessarily match those required for social or economic analysis (Brackstone, 1987). An important aspect of the data content is continuity over time of the variables included and their definitions. Coverage and content in administrative data systems may be subject to discontinuities resulting from changes in laws, regulations, administrative practices or the scope of the program (Brackstone, 1987). Administrative data systems also can not ensure perfect data quality. Information that might be important for statisticians, but less important for administrative purposes, is often reported and processed imperfectly, noticed Johnson and Moore (2008). Another issue pointed out by Johnson and Moore (2008) is data reliability which can be affected if the information provided to the tax entity can cause profits or losses for the declaring subject. Moreover, given that the data collected and processed for administrative purposes are generally given priority over what is required for statistical purposes, the amount of processing required to provide administrative data suitable for statistical purposes may affect the time that these data are made available to statisticians.

6.3. Privacy Issues

In their work Johnson and Moore (2008) consider data privacy as a very important issue. The authors explain that any use of administrative data for research purposes should take into account the laws that protect the privacy of the data. The research of administrative data is often limited to uses within the scope of an agency mission and should be carried out only by persons working for the agency as employees, contractors or under Memorandum of

Understanding that allow employees of various institutions to exchange the data. The way the public percept privacy protection of their data has a direct impact on the continuity of the levels of declarations. Often, because of these factors, the available data does not contain identifying variables. For example, in the case of individual data from the administrative source, variables which directly identify the subject are missing. Of course in another scenario the availability of these variables could lead to wider statistical use and a combination of data from different sources.

However, data confidentiality is of great importance to the current and future success of any administrative observation and registration. If the subjects do not believe that their data is sufficiently protected, response rates and overall data quality will be subject to deformation.

6.4. Classification Issues

The classification systems used within administrative sources may be different to those used in the statistical world. Even if they are the same, they may be applied differently depending on the primary purpose of the administrative source, perhaps focusing on specific attributes of the unit. For example, an administrative source concerned with licensing, health and safety or environmental protection may be more interested in the economic activities of a business that are of most concern to that source, rather than the main economic activity of a business, which is required for statistical purposes. In other cases, classifications in administrative sources may not be applied at the level of detail required for statistical purposes, or the classification may simply not be a priority variable for the administrative source, resulting in quality deficiencies. Where classification systems or versions are different, the usual solution is to construct conversion matrices to map the codes in the administrative classification to those in the statistical classification. Such mappings may be one to one, many to one, one to many or many to many. In the latter two cases, some sort of probabilistic allocation may be required. In addition to the use of common coding tools, the provision of coding expertise and training to administrative data suppliers can help to improve coding consistency. At the same time, it is always helpful for the statistician to stress the advantages of using a common classification system. It also helps to give early notice of any revisions to the classification system, and to provide as much help as possible to administrative data suppliers during the implementation of the resulting changes.

6.5. Timeliness Issues

There are three separate issues relating to timeliness that affect the usefulness of administrative data for statistical purposes:

- Administrative data may not be available in time to meet statistical needs
- Administrative data may relate to a period that does not coincide with the statistical reference period
- Administrative data may be measured over a period, whilst the statistical requirement is for a specific point in time (or vice-versa).

Considering the first issue, there will generally be some sort of lag between an event happening in the real world, and it being recorded by an administrative source, this is then followed by a further lag before the data are made available to the national statistical organisation.

The second issue related to timeliness is that of differing periods, for example data from annual tax returns are often only available several months after the end of the tax year, so are probably not suitable for monthly or quarterly statistics. In some cases, however, annual administrative data can be used for shorter-period statistics, particularly if they are collected on a rolling annual basis. This can happen if there is a requirement to spread the workload of collecting and processing these data by the administrative source throughout the year. As long as the distribution of the units for which data are collected during the year is sufficiently random, it may be possible to derive meaningful monthly or quarterly statistical trend data from such sources.

The third issue concerns the difference between data referring to a specific point of time and data relating to a period (e.g. an annual or monthly average). For example, there may be a statistical requirement for employment data on a specific reference date, whereas administrative data may only give periodical averages.

6.6. Issues Relating to Inconsistency between Sources

A specific problem where multiple sources are used concerns inconsistencies between those sources. Data from one source may appear to contradict those from another. This may be due to different definitions or classifications, differences in timing, or simply to an error in one source. This can happen when comparing administrative data with statistical data, or when comparing two administrative (or two statistical) sources.

To resolve such conflicts it is necessary to establish priority rules, by deciding which source is most reliable for a particular variable. Once a priority order of sources has been determined for a variable, it should then be possible to ensure that data from a high priority source are not overwritten by a lower priority source. This process is made much easier if source codes are stored alongside variables for which several sources are available. The use and storage of dates can also be helpful, as even when one source is thought to be more reliable than another, data from that source that are ten years old may not be of higher quality than data for the most recent period from the less reliable source. A simpler method that may be appropriate in some cases is to load data in reverse priority order, allowing data of higher quality to overwrite those of lower quality.

The more data sources that are used, the more complex this comparison process becomes, but having multiple sources often helps to validate data quality. In some cases, certain sources may not be used directly for statistical production, but purely for benchmarking purposes as part of a quality assurance process.[2]

[2]An example of benchmarking, using maps to compare the coverage of a statistical business register with that of a commercial telephone directory can be found in the paper "The Development of Small-area Business Statistics in the United Kingdom"

6.7. Issues relating to Missing Data

The problem of missing data is not unique to administrative sources. It can also be due to full or partial non-response to statistical surveys, or even to the removal of data values during the editing process. However, with administrative sources, the issues can sometimes be different, particularly as the problem of missing data can often be more systematic. The main reasons for this are that a particular variable may not be collected at all by the administrative source, or it may only be collected for certain categories of units where there is a specific administrative requirement.

The variable may also simply be a low priority for administrative purposes, so the owners of that source do not see missing data as a problem. Some of the standard solutions for dealing with non-response in statistical surveys can also be used to solve the problem of missing data in administrative sources. Various imputation methods, such as deductive, 'hot-deck' or 'cold-deck' imputation are often suitable where the problem only affects some of the units. In cases where most or all of the units are affected, a modelling approach may be more appropriate.

6.8. Issues relating to Resistance to Change

One of the main barriers to the more effective use of administrative sources in official statistics, and one of the least recognised, can come from within the organisation. Statisticians may resist the use of administrative data because they do not trust data that they have not collected themselves. They often focus on the negative quality aspects of administrative data, and they have an over-optimistic view of the quality of survey data, often based on the largely untested assumption that survey responses actually comply with statistical norms.

The solution is clearly through better education of statisticians regarding the possibilities offered by administrative sources, encouraging them to take a wider view of all the dimensions of quality, and focus on the impact on data suppliers and users. In this context it is important to determine the real relative quality of survey and administrative data. For example, it is often assumed that data from administrative sources do not meet the requirements of statistical definitions, whereas those from official surveys do. However, there may not be any real difference in practice, particularly if respondents to statistical surveys simply inform the same values from recent administrative returns, without paying attention to the investigators explications at the time of canvassing the Questionnaire or Schedule of enquiry.

6.9. Issues relating to the Adequacy in Decision Support

There can be difficulties in using administrative data to answer specific research questions. This might be because the data were not collected with those questions in mind, because of quality issues that are irrelevant to operations but highly relevant to subsequent statistical analysis, because of changes in definitions of the recorded data items or for other reasons. This brings to the fore the issue of adequacy of administrative data in evolving a decision support system: it can be useful, if it is possible, to have statistical advisors in the Central Ministry/Departments as also in the States involved in the data collection process. They might be able to think ahead and to expand the range of data collected so that they will be more able to anticipate a robust framework for generating the administrative data with a view to buttress a sound and robust decision support system.

Statistical analysis methods are often divided into *descriptive* and *inferential*. *Descriptive* methods are used to summarize a body of data so that the important messages within it can be readily grasped. We might summarize a distribution of values by their mean and standard

deviation, or the results of a census by using a series of counts organized as cross-tabulations. It goes without saying that the summary statistics that are appropriate will depend on the subject matter and on the questions to which answers are sought. Administrative data are often used for purely descriptive purposes—perhaps especially so in official statistics contexts, where we might want to establish the characteristics of some population.

In contrast, *inferential* methods are used to make a statement about unobserved values or underlying mechanisms. We might be trying to infer the disease of a new patient, on the basis of analysis of patients with similar symptoms diagnosed in the past. We might be trying to forecast whether inflation will go up or down next month. We might be trying to elucidate an underlying mechanism, so that we can understand how the data were generated, and perhaps influence things in the future. Much of the statistical theory of inference is based on the notion of random sampling from a (possibly infinite) population of values. Because the sampling is random, solid mathematics (such as the law of large numbers and the central limit theorem) means that sound statements can be made about the characteristics of the population from summary statistics obtained from the sample. Moreover, error bounds can be put on the conclusions. We can say things such as 'on average, 99 out of 100 of our intervals will cover the true population mean', so we can be confident of our results (always subject to data quality issues, of course).

But administrative data are not collected by such a random sampling process. We can certainly calculate descriptive statistics, summarizing the data before us and, if we are willing to assume that the data are perfect, with no missing or distorted values, then this will accurately summarize the population which led to our data. We can make a statement such as 'this is the true population mean'.

Economic and social measures such as gross domestic product, the consumer price index and national wellbeing are what are called *pragmatic* measures (see, for example, Hand (2004)): the definition of the concept and the way that it is measured are two sides of the same coin. Change the measurement procedure and you change the thing being measured, with different measures being suitable for different purposes. It is not a question of any of these being more 'right' than the others, but simply that they measure slightly different things. This means that they have different properties and are suited to answering different questions. Increasingly, interest is turning to the possibility of using administrative data for measuring productivity and price inflation.

Instead of conducting surveys of businesses to obtain data, the data can be automatically transmitted from the transaction to the database.

Scanner data, such as retail purchase data obtained directly from the point-of-sale machine, provide an example, yielding data that are ideal for use in price index calculation. Moreover, such data also give information on the volume of different goods purchased, so that weights can be chosen. But issues of selection bias still apply: not all purchases are made through such routes, and we cannot assume that those purchases which are made in this way represent a random or representative sample of all purchases. A variant of this uses Web-scraped price collection, being explored by various national statistical offices.

One of the problems with Web-based tools is the rate of change of that technology. Companies appear, grow to a massive size and vanish at a dramatic pace.

7. Sectoral Position in India: Observations of Rangarajan Commission

7.1. Agricultural Statistics

- Statistics of crop production – both area and yield – are based on scientifically designed methodologies.

- Timely Reporting Scheme (TRS) and the Improvement of Crop Statistics (ICS) Scheme.

- At present, the area statistics are generated through complete enumeration in the temporarily settled States while in the permanently settled States these are arrived at through a sample of 20 per cent villages covered by the Establishment of an Agency for Reporting Agricultural Statistics (EARAS) scheme

- The Commission has, therefore, recommended that crop area forecasts and final area estimates issued by the Ministry of Agriculture should be based on the results of the TRS in the temporarily settled States and on those of EARAS in the permanently settled States.

- The Commission is of the view that it is necessary to make an objective forecasting based on timely and detailed information on crop condition, meteorological parameters, water availability, crop damage, etc.

- While the use of Remote Sensing Technology does offer an alternative route for the regular flow of crop statistics, there are a number of issues that require to be sorted out before this can become extensively operational. Meanwhile, the existing programmes of Remote Sensing

Technology must be pursued with active cooperation from the concerned agencies.

• The data collected through Agricultural and Livestock Censuses are required for identifying and formulating policies and programmes for the rural population. However, as the results of these censuses are not available in time, this defeats the very purpose for which these censuses are conducted. To circumvent these problems, there is a need for conducting the censuses not as complete enumeration but as sample censuses.

• Further, no relationship has been worked out based on the data collected through these two censuses because they are conducted independently with different field agencies, reference periods and basic units of enumeration. In view of several operational and substantive gains, the Commission has recommended the integration of the Livestock and Agricultural Censuses.

7.2. Industrial Statistics

• The Annual Survey of Industries (ASI) has been the principal source for most of the basic statistics of the Industrial Sector.

• Estimates of the growth rates of industrial production based upon the Index of Industrial Production (IIP) are extensively used for policy-making at various levels in the Government and also for decision-making in the banking and Corporate Sectors.

• The IIP is compiled and released by the Central Statistical Organisation (CSO) within six weeks as per Special Data Dissemination Standards (SDDS) norms of International Monetary Fund, based on the data received from different agencies.

7.3. Services Sector Statistics

• The Follow-up Enterprise Surveys on the Services and other sub-sectors (excluding manufacturing and repairing sub-sectors), carried out by the Ministry of Statistics and Programme Implementation, take into account all types of enterprises (other than those in the public sector), irrespective of their size, under the same survey year.

7.4. Infrastructure Statistics

• In developing a proper statistical database for the Infrastructure Sector, a major hurdle is the absence of a clear definition of "infrastructure".

• Quantification of the infrastructural activities in the form of an index would help policy makers and researchers. The Commission has therefore recommended the construction of two types of indices in this regard. While the first one,

called "Infrastructure Index", will provide a summary measure of the growth of infrastructure, the second one, namely, "Infrastructure Utilisation Index", will indicate the extent of utilisation of identified infrastructure facilities.

7.5. Socio-economic Statistics

• In the area of Population Statistics, the Population Census is one of the most comprehensive sources of information on the size, distribution, living conditions and demographic characteristics of the population

• In the area of Health and Family Welfare Statistics, the three Departments of Health, Family Welfare and Indian System of Medicines and Homeopathy of the Ministry of Health and Family Welfare have a separate system of data collection in their respective areas, while the Registrar General of India is responsible for collection and dissemination of vital statistics through its system of registration of vital events.

• The country has a well-established system of civil registration through an elaborate machinery right up to the district level and below for registration of vital events under the Registration of Births and Deaths Act. The Civil Registration System, has the potential to generate vital rates for district level and below and form the basis for planning health and family welfare programmes at the local level as required in the 73rd and 74thAmendments.

• Labour and Employment Statistics are generated largely through the implementation of various labour laws and Regulations by the States and Centre. For the unorganised sector, the National Sample Survey Organisation and Central Statistical Organisation are collecting and disseminating labour and employment-related data by conducting periodic sample surveys.

• The Registrar General and Census Commissioner of India is also publishing data decennially on workers and those seeking work through its census operations.

• The data collected by Ministry of Labour through States suffer from very poor response in submission of returns, delays in filing the returns, poor quality, under coverage and time lag in publication of results.

• Ministry of Human Resource Development is the main agency for producing statistics on school education, which are collected through the States.

• The All India Educational Survey conducted by the NCERT is another important source of statistics on school education in the country. In its review of the educational statistics system, the Commission took note of the

deficiencies of quality, reliability, time lag and weak infrastructure in the collection and dissemination of education data

• The Department of Women and Child Development should play a proactive role and strengthen its statistical set up. Indicators of gender disparity in various aspects of education, health and employment are required to be brought out. The CSO should develop a standard methodology for the purpose of generating these indices to reflect the status of women in the country.

• Environment Statistics is in its nascent stage in the country and as such there is a need to build up an efficient system for the collection of Environment Statistics and developing environmental indicators based on the international framework provided by the United Nations Statistics Division.

7.6. Financial and External Sector Statistics

• The Reserve Bank of India (RBI) is the principal though not the sole agency for collection and dissemination of statistics in respect of Financial and External Sector Statistics.

• The other major public sector agencies and institutions that collect, compile and disseminate Financial Statistics are the Ministry of Finance, Securities and Exchange Board of India (SEBI), National Bank for Agriculture and Rural Development (NABARD) and Industrial Development Bank of India (IDBI).

• There are continuing discrepancies in merchandise trade data, both exports and imports, between the Directorate General of Commercial Intelligence and Statistics (DGCI&S) and RBI.

7.7. Price Statistics

• Central and State Government agencies collect the primary data on prices for varied purposes. The data on prices, both for the wholesale price index and consumer price indices, are not satisfactory.

7.8. Corporate Sector Statistics

• The Corporate Sector includes not only the domestic corporates but multinational companies of various types as well. In the Department of Company Affairs (DCA), the Registrars of Companies (ROCs) are primarily responsible for provision of the basic information on the Corporate Sector and the statistical machinery in the ROCs is inadequate to deal with this task.

8. Quality & Administrative Data

Concerns about the quality of administrative data are often one of the main barriers to their increased use for statistical purposes. These concerns may or may not be justified, and are often based only on specific aspects of quality, such as timeliness. To properly address these concerns an objective quality management framework is needed; one that considers all relevant aspects of quality, and allows an informed decision to be made.

Many statistical organisations have already put in place some sort of quality framework for data collected via traditional survey methods, but relatively few have extended this approach to cover data from administrative sources.[3]

To fully understand the quality of administrative sources, and their impact on the quality of statistics, the following elements may be considered:

8.1. The quality of incoming data

The incoming data, whether they are from administrative or survey sources, can be judged against set of criteria such as those listed above. The most important criteria are likely to be timeliness, and relevance in terms of the extent to which the coverage and concepts of the source meet requirements. Comparability with other sources can also be important, and some sort of exercise to reconcile data from different sources may be necessary from time to time to get a clear picture of quality. Quality check surveys are sometimes used for this purpose. One point worth bearing in mind is the extent to which the data subject has an interest in the quality of the data. The amount of effort and care put into providing the data will vary according to the perceived value or importance of the data collection, thus data subjects may, in some cases, provide better quality data for administrative purposes than they do for statistical purposes.

8.2. The quality of data processing

Even if the incoming data are perfect, their quality can still be affected by the different processes they go through before they are used for statistical outputs. Ideally processing should improve quality, but unfortunately this is not always the case. Examples of how data processing can affect quality include:

- Data matching and linking – too many false matches will lead to

[3]Examples include approaches developed by Statistics Netherlands and Statistics Sweden

errors in the data; too many false non-matches will lead to duplication, which will overstate the size of the population of interest, and possibly introduce bias.

- Outlier detection and treatment – using outlier detection methods to detect errors can help to improve the quality of the data, and generally the more extreme the outlier, the more likely it is to be an error. However over-zealous treatment of outliers will result in genuine data values being altered and can lead to important trends in the data being missed.

- Quality of data editing – as for outlier detection and treatment, data editing should improve quality, but if not done carefully it can introduce error and bias.

- Quality of imputation – if imputation is used to fill missing values or records it can help to improve coverage, but again the methods used need careful scrutiny to avoid the introduction of bias.

8.3. The quality of statistical outputs

The usual interpretation of the quality definition by statistical agencies is that quality is all about meeting user requirements. The quality of statistical outputs is therefore determined in this context. This means that it is necessary to determine these requirements, to discuss them with users, and to get regular feedback, for example via user satisfaction surveys.

Moving from survey to administrative sources will clearly have an impact on output quality. Typically this impact may be positive for some quality criteria, and negative for others. In all cases, it is necessary to get an overall view of the impact, giving greater weight to those criteria the users consider to be the most important. For example, users may feel that an improvement in timeliness more than compensates for a reduction in accuracy, particularly for short-period economic data. Another consideration should be the impact on time-series data, and whether it is possible to construct a consistent series of sufficient length following the change. It can be particularly important to give at least as much weight to the views of users as to the perceptions of statisticians, which may, in some cases be too heavily focussed on traditional notions of accuracy. Overall, it is vital that any judgment of the impact on statistical outputs is based on objective evidence rather than on supposition, as this is the only way to counter the potential for resistance to change.

9. National Data Quality Assurance Framework (NQAF) in India

A National Quality Assurance Framework contributes to the improvement in the National Statistical System by laying down quality parameters and corresponding good practices and elements that need to be in place to facilitate and ensure effective management of quality in the statistical system, processes and products. The NQAF follows and aligns with the UN National Quality Assurance Framework 2019. The NQAF constitute the common quality framework in terms of quality management principles and their corresponding requirements and elements for assessing the Statistics which are produced and disseminated. These principles, requirements and elements have been constructed by carrying out certain customization/modifications on the UN NQAF 2019 based on rounds of consultations and review processes held with relevant stakeholders, viz. Various Divisions of Ministry of Statistics &Programme Implementation (MoSPI) and key Central Ministries/Departments and States/UTs, to ultimately build a National Quality Assurance Framework (NQAF) appropriate to the Indian context. Subsequently, the QAF-ISS was duly approved and recommended for adoption by the Task Force headed by Director General (Statistics), National Statistical Office (NSO), MoSPI.[4]

The UN NQAF 2019 Manual calls for the establishment of the necessary institutional arrangements for the development of an NQAF. This includes establishment establishing a quality unit at the NSO and a quality task force (or working group) MoSPI, vide its Officer Order dated 18.03.2020, has designated the SSP Unit as the Standardization Cell akin to quality unit at NSO.

10. Improving Usability

Sen (2009) while noting in Indian context is of the opinion 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

[4] Task Force was notified vide MOSPI's OM dated 21.08.2020

indicators for the inter-census period.

He further opines that surveys carried out by the National Statistical Office (NSO) 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.

11. Institutional Mechanism

The evaluation of administrative data quality for statistical purposes can be a huge task. One step in this evaluation process is – after dealing with concepts, classifications, timeliness, processing and data treatment, data linkage and matching and other issues – verify if the information that we get from administrative data sources is valid and precise. Recently, MoSPI has put in place a dedicated Division in the Ministry to explore the possibilities of statistical usages of alternative source of data for official statistics. Besides, on the recommendations of the National Statistical Commission (NSC), MOSPI has also constituted a Committee[5] under the Chairmanship of Dr. KiranPandya, Member, NSC *inter alia* to do the following:

a. Identify availability of possible alternative institutional data sources in Centre and State in India;

b. Suggesting an institutional framework enabling usages of alternative data sources in official statistics in India;

c. List out the scope and coverage of alternative data source(s) in Indian context;

[5] MOSPI Order dated 11.02.2021

d. Lay down guidelines for developing survey instruments in view of available alternative data sources; and

e. Deliberate upon any other relevant matter.

12. Conclusion

It is reiterated that the objective of the document is to generate discussion, and deliberation among policy framers, theoriticians and practioners and other stakeholders, necessitating the need for methodological statistical work on administrative data. It is evident that such data are being used increasingly more widely—partly a consequence of the 'big data' revolution. But drawing any reliable inferences from such data encounters problems that are diverse from the sociable and arterial promenade of sampling theory inference. The problems are manifold and contrasting, so it is irresolute that an integrated theory, as elegant as that of sampling theory, can be developed without providing for environment for exclusive research and development in the country. But, nevertheless some fundamental propositions are well settled. These include the need to stand on one's two feet regarding data quality issues, the identification that, despite peripheral emergence, we ordinarily do not have 'all' the data, possible mismatches between the question we want to answer and the information in the available data, challenges arising from the fact that the data are (usually) merely observational, so elucidation of causality is difficult, the need to combine data from multiple rather different sources, and issues of confidentiality, privacy and anonymization which might be rather different from those of survey data.

13. References

Brackstone. (1987). Statistical uses of Administrative Data: Issues and Challenges. Statistical Uses of Administrative Data Proceedings.

Hand, D. J. (2004). Measurement Theory and Practice: the World through Quantification. Chichester: Wiley

Hand, D. J. (2008). Statistics: a Very Short Introduction. Oxford: Oxford University Press.

Johnson, B., & Moore, K. (2008).Comparing Administrative and Survey Data.IRS Statistics of Income Working Paper Series.

Künn, S. (2015).The challenges of linking survey and administrative data.IZA World of Labour, 214.

National Statistical Commission (2001). Report Submitted to the Government of India.

Nordbotten, S. (2010). The use of administrative data in official statistics—past, present and future: with special reference to the Nordic countries. In Official Statistics: Methodology and Applications in Honour of Daniel Thorburn (eds M. Carlson, H. Nyquist and M. Villani), pp. 205–223. Stockholm: Statistics Sweden.

Organisation for Economic Co-operation and Development (2016) Short-term economic statistics (STES) administrative data: two frameworks of papers. Organisation for Economic Co-operation and Development, Paris.

Penneck, S. (2007).Using Administrative Data for Statistical Purposes.ICES-III. Montreal, Quebec, Canada.

PronabSen (2009). Challenges of Using Administrative Data for Statistical Purposes: India Country paper.

UNECE. (2011). Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices, United Nations Economic Commission for Europe. New York and Geneva: UNITED NATIONS.

Wallgren, A. and Wallgren, B. (2014) Register-based Statistics: Statistical Methods for Administrative Data, 2nd edn. Chichester: Wiley.