Integrated Longitudinal Database for Supporting Policy Studies

by Sia Ziyi, Chan Wen Chang and Malcolm Cai
Longitudinal Data Analytics Division
Singapore Department of Statistics

Introduction

The Singapore Government adopts a data-driven approach in shaping policies using data analytics to examine issues, identify effective strategies, and assess the impact of policy interventions. This is done across diverse domains, encompassing the economy, social compact, security, and urban and cultural development. To support this, the Singapore Department of Statistics (DOS) has developed an integrated Longitudinal Administrative Database (LAD).

This article provides a brief overview of longitudinal data and explains why it is a valuable resource for research and policy studies. It then shows the global landscape of such datasets and provides an in-depth look at DOS's LAD. DOS's considerations and measures taken to optimise the dataset's utility while upholding the confidentiality of data are also discussed.

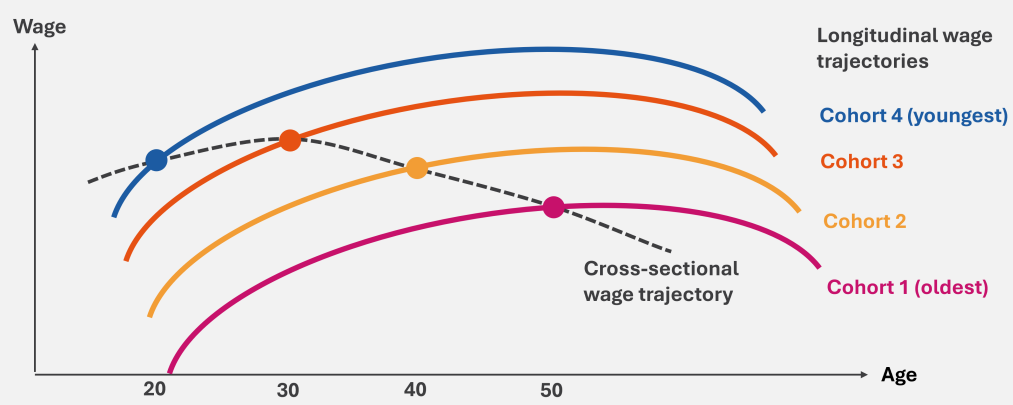
Importance of Longitudinal Data

Longitudinal data refers to data collected on the same research unit (e.g., individual or firm) repeatedly over time. This allows researchers to distil information that may be masked by cross-sectional data.

For example, when studying the lifetime wage trajectory of workers, using cross-sectional data at a particular point in time only captures the wages of different workers at various ages, as depicted by the dotted line in Figure 1, leading to inaccurate conclusions. This is misleading because as a younger worker ages, his wage will not be the same as the wage of a current older worker.

Repeated cross-sectional data can further distort the picture due to compositional changes in the cross-sectional dataset, such as individuals dropping out of older cohorts and the inclusion of new cohorts.

Figure 1: Cross-Sectional vs. Longitudinal Wage Trajectories



Trajectory measured by cross-sectional data can be different from one that is measured by longitudinal data.

In contrast, longitudinal data tracks the same individuals over time, allowing the observation of their actual wage trajectories, as represented by the solid lines in Figure 1, providing a more accurate picture of how wages evolve over a worker's career and better insights into the shifts in wage trajectories across different cohorts.

Using longitudinal data allows Government researchers to better estimate the causal impact of policy levers on outcomes, as the nature of the dataset enables the use of more sophisticated modelling techniques.

- For example, when measuring the economic returns (proxied by wages) to education (a policy lever), other factors such as an individual's innate ability which may also affect the level of wages would need to be accounted for. Individuals with higher ability tend to have both higher education and higher wages, and not accounting for ability would overstate the causal impact of education on wages. With longitudinal data, statistical techniques such as fixed-effects regression modelling can be used to remove confounding effects of time-invariant unobserved characteristics such as ability, providing a better estimate of the effects of education on wages.

International Landscape

International National Statistical Offices (NSOs) construct longitudinal databases for various statistical and research purposes. For example, **Statistics Canada's Longitudinal Administrative Databank** is mainly used by the Canadian Government as a research tool to evaluate programs and support policy recommendations. The Databank can track **low-income persistence** and year-over-year low-income entry and exit rates. The New Zealand Government tapped on its **Integrated Data Infrastructure** to study the performance of their welfare system and social outcomes.

To collect the data required to construct longitudinal databases, NSOs typically tap on either survey data or administrative sources (Table 1). Survey data can be collected through longitudinal surveys, which involve repeatedly surveying the same individuals or firms over time, or by linking survey data from multiple censuses conducted at different points in time. Administrative sources involve linking existing administrative data collected from various sources over time to build a register.

Longitudinal Surveys, such as the Longitudinal and International Study of Adults by Statistics Canada and the National Longitudinal Survey of Youth 1997 by United States Bureau of Labor Statistics, allow for survey questions to be tailored to meet specific data demands. However, conducting large-scale surveys over time can be resource-intensive and impose a burden on respondents, and hence limit their coverage to a portion of the population.

Table 1: Longitudinal Data in National Statistical Offices

NSO/ Country	Longitudinal Database	Period	Scope	Primary [1] Method of Data Collection
Statistics Canada	Longitudinal and International Study of Adults	2011-latest	Sample of households located in ten provinces	Survey (Longitudinal)
United States Bureau of Labor Statistics	National Longitudinal Survey of Youth 1997	1997-latest	Sample of persons born 1980-1984 in 1997	Survey (Longitudinal)
Australian Bureau of Statistics	Australian Census Longitudinal Dataset	2006-latest	5% sample of the population	Survey (Linking Census survey data across years)
Office for National Statistics	Longitudinal Study	1971-latest	1% sample of the population	Survey (Linking Census survey data across years)
Statistics Canada	Social Data Linkage Environment	1926-latest	Generally administrative files to form Derived Record Depository	Administrative
Australian Bureau of Statistics	Person Level Integrated Data Asset	1990-latest	Generally administrative files to form person linkage spine	Administrative
Statistics Canada	Longitudinal Administrative Databank	1982-latest	20% sample of the population	Administrative
Statistics New Zealand	Integrated Data Infrastructure	1840-latest	All New Zealand residents	Administrative
Statistics Sweden	Longitudinal Integration Database for Health Insurance and Labour Market Studies	1990-latest	Persons aged 16 and over	Administrative
United States Census Bureau	Longitudinal Employer-Household Dynamics	1990-latest	95% of the employed population	Administrative

[1] Primary method refers to the main method of data collection. Data can be augmented with other sources (i.e., Administrative source supplemented with survey data, or vice versa).

Agencies such as Statistics Sweden and Statistics New Zealand primarily rely on administrative data to build their longitudinal databases. Administrative sources [2] often provide almost complete coverage of the target population, allowing for studies to focus on specific groups of people (e.g., individuals on social assistance with specific attributes) with less sampling error. The data is also more accurate, as multiple administrative sources allow for cross-checking and validation of the figures. However, the available data is restricted to data collected for administrative purposes, which may limit the range of variables in the longitudinal dataset. Nonetheless, this limitation can be addressed by augmenting the database with additional data (e.g., from surveys).

Maximising the Value of LAD for Policy Analyses

In Singapore, DOS has built the LAD mainly using administrative data augmented with survey data. Administrative processes in Singapore are mostly digital, where data are automatically captured during transactions with the Government, making data collection cheaper, faster, and less burdensome for citizens and firms compared to traditional surveys. Nevertheless, data from surveys are used to augment and update administrative data where necessary.

Context

The LAD was set up to address the policy research needs of the Government and advance the national capability for social science research, particularly in the analysis of social issues such as income mobility and household dynamics. Over time, the LAD has been expanded to support research in a wide range of domains, including analyses of the labour market, education, healthcare, and the corporate landscape.

The LAD is centrally built and managed by DOS, leveraging DOS's expertise in collecting, cleaning, merging, and processing data, as well as its experience in data governance and security. In addition, DOS is familiar with data concepts and definitions of various primary administrative databases within the Government. Considering economies of scale, a centralised LAD is more cost-effective than setting up separate databases in individual agencies.

Over the years, the LAD has supported many research studies within the public sector. In the economic domain, the LAD was used to study the **corporate landscape of Singapore**, aiding Singapore's Central Bank to gauge the financial and productivity performance of firms. Similarly, DOS tapped on the LAD to profile **high growth firms in Singapore**. In the social domain, the Ministry of Finance has leveraged the LAD to estimate **intergenerational mobility** by examining the correlation between the incomes of fathers and their sons.

The LAD has also supported numerous policy evaluation studies. For example, it was used to access the impact of **Workforce Skills Qualifications (WSQ) training on trainees' wages and employment prospects** and to study the effects of Enterprise Singapore's **grants on firms' revenue and exports.** The longitudinal nature of the datasets allows researchers to control for unobserved individual- or firm-specific factors that are time-invariant during the period of analysis.

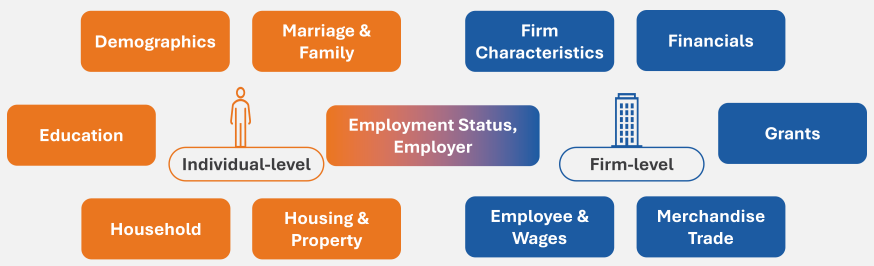
Clean and Rich Data

To ensure that the LAD is logically consistent and coherent over time, DOS developed longitudinal processing capabilities to identify longitudinal anomalies such as inconsistency in records over time. Such anomalies include changes in time-invariant characteristics (e.g., ethnicity), decreases in values that should not decline over time (e.g., highest educational qualification), or cases where deceased individuals erroneously appear in the 'living persons' dataset. Such records would be flagged, investigated, and rectified before storing in the database as clean data.

Besides record-level checks, DOS monitors trends of aggregated statistics (i.e., mean and median) generated through the LAD to detect any trend breaks which may occur from changes in administrative processes, and in turn affecting data collection procedures. DOS ensures that data and definitions remain comparable over time. For example, when the Singapore's Standard Industrial Classification [3] is revised, industry data from earlier years are re-mapped to the latest classification to ensure continuity and comparability over time.

Both individual- and firm-level observations are included in the LAD. At the individual level, it contains variables on employment status, employer, demographic profile, education, households, marriages, and family. Family information includes parental linkages, enabling comparisons between child and parent outcomes to study intergenerational mobility. At the firm-level, the LAD captures key characteristics such as financials, grants, merchandise trade, employees, and wages. Figure 2 illustrates how individual- and firm-level data can be linked to form an employee-employer dataset for hierarchical analysis.

Figure 2: Longitudinal Datasets for Individuals and Firms



The LAD can be fused with external datasets on an ad-hoc basis to create rich, domain-specific datasets for Government researchers to leverage for further deep-dive analyses into specific domains.

[2] Learn more about **Using Administrative and Secondary Sources for Official Statistics [The Advantages of Using Administrative Sources, section 2.5, page 10]**

[3] The **Singapore Standard Industrial Classification (SSIC)** is the national standard for classifying economic activities undertaken by economic units and is used in censuses of population, household and establishment surveys, and in administrative databases.

Consultancy

DOS provides analytical consultancy services to Government agencies to maximise the LAD's value via:

1. **Collaborating with policy agencies**
   to translate their policy questions into empirical analytic problems that can be addressed.
2. **Providing data-related advice,**
   including guidance on the availability and suitability of the data items for their projects, highlighting data caveats that may affect the interpretation of findings, scoping the coverage of the project dataset, and defining appropriate criteria for treatment and control groups for policy evaluation.
3. **Offering statistical and methodological advice**
   to ensure that studies are based on robust methodologies.
4. **Training Government officers**
   in data analytics, econometrics, machine learning, and statistical programming.

Ensuring Confidentiality and Security of the Data

To ensure the integrity and security of the LAD and the confidentiality of the entities within, DOS manages the data using a robust **data governance framework** comprising several components, including legislation, data management policies and process safeguards, and Information Technology (IT) systems.

The LAD is governed by the Statistics Act 1973, one of the key legislation governing statistical activities in the public sector. It mandates data collection and safeguards the confidentiality of information provided. For example, key data items such as income are collected under the Act to study income growth and mobility trends in Singapore. Under the legislation, the LAD can only be used for statistical and research purposes within the Government and cannot be used for administering policies directly.

Data management policies are implemented based on the relevant Government Instructional Manuals, and a high-level of process safeguards are imposed. The LAD is accessible only by a selected team of DOS officers to manage the data and conduct research projects. Identifiable data are not disclosed or shared outside of DOS, even within the Government. Only aggregated data that has passed statistical disclosure controls, or anonymised data, can be accessed by Government officers [4]. DOS's practices are relatively more stringent than those of other NSOs, reflecting a prioritisation of privacy of data records over accessibility.

Robust IT security measures are in place to protect the data in storage and in use. A full suite of data loss protection and isolation measures is implemented to prevent exfiltration of data from DOS's systems. Access management systems control identity and access rights ensuring that data can only be accessed by authorised persons; all access and usage of data is consistently managed, logged, and monitored. Data are encrypted at all times, and data lineage records track the usage of sensitive data, the flows between users and systems and the changes made to the data.

Overall, DOS's robust data management processes adhere to, and are more robust than, international standards. This approach balances maximising the use and value of the LAD while maintaining data integrity, security, and confidentiality.

Conclusion

In conclusion, the LAD is a rich, integrated longitudinal dataset, created to advance robust evidence-based policymaking in the Government. DOS ensures strong measures are implemented to safeguard data integrity, security, and confidentiality. The LAD has been used in many research studies across multiple domains, with findings supporting government agencies to formulate or refine their policies to benefit the people of Singapore. DOS remains committed to maintain the LAD to the highest standards, providing reliable and comprehensive data to support policy studies.

[4] Government officers with access to aggregated data may include academics commissioned by Government agencies to provide expertise to the research team.