Chapter 2
Secondary, Longitudinal, and Panel Data in Social Science Research

Ilaria Primerano, Nicolò Marchesini, Francesco Santelli, Luciana Taddei, and Loredana Cerbara

2.1 Introduction

In the field of social science, secondary research plays a pivotal role. It allows for the analysis of existing data to generate new interpretations, deepen the understanding of complex phenomena, and support the development of evidence-based policies (Biolcati-Rinaldi & Vezzoni, 2012). Secondary research differs from primary research, which is based on the collection of new data through the design and implementation of a study related to specific research questions. Primary research can be conducted using various methods, such as surveys, interviews, and focus groups. Unlike primary research, secondary research relies on the analysis of existing data collected from previous studies. Usually, social researchers start from such secondary data to identify potential gaps in the literature on the investigated

I. Primerano (✉)
Institute for Research on Population and Social Policies (CNR-IRPPS), Fisciano, SA, Italy
e-mail: ilaria.primerano@cnr.it

N. Marchesini
Italian National Institute of Statistics (Istat), Rome, Italy
e-mail: nicolo.marchesini@istat.it

F. Santelli
Department of Political and Social Sciences, University of Trieste, Trieste, Italy
e-mail: fsantelli@units.it

L. Taddei
Institute for Research on Population and Social Policies, National Research Council, Fisciano (SA), Italy
e-mail: luciana.taddei@cnr.it

L. Cerbara
Institute for Research on Population and Social Policies (CNR-IRPPS), Rome, Italy
e-mail: loredana.cerbara@cnr.it

L. Taddei, M. Paolucci (eds.), *Longitudinal Data Infrastructures in Europe*,
https://doi.org/10.1007/978-3-032-07005-0_2

topic, and by interpreting, organizing, and analyzing data from prior studies, they are able to generate new knowledge.

Among the many forms of secondary data, longitudinal and panel data are particularly valuable for understanding several aspects of human life, since they capture changes over time. When used in secondary research, longitudinal and panel data offer unique methodological advantages, as they allow the study of changes, causality, and life-course events. Although originally collected as primary data by public institutions, research bodies, or international organizations within the framework of structured surveys and specific objectives, such data take on the nature of secondary data when reused by other researchers for analyses different from those envisaged in the original research design. This shift from primary to secondary data is based on the principle that the nature of data depends not only on how they were collected, but also on the context in which they are used (Vartanian, 2011). Access to these data archives enables the development of new research on emerging topics, making use of the informational potential of existing data, while also promoting the efficient use of scientific resources (Johnston, 2014).

In fact, longitudinal and panel data are especially useful for research aiming to understand the evolution of opinions, behaviors, values, and socio-economic conditions (Ruspini, 2002; Agnoli, 2008). Specifically, these data are indispensable in various domains of social science because of their ability to account for temporal dynamics and individual-level variation. In political science, for instance, they facilitate the tracking of changes in political attitudes and behaviors over electoral cycles (Bartels, 2006; Neundorf & Smets, 2017). In labor economics, they support the analysis of employment transitions, wage trajectories, and job mobility (Arulampalam, 2001). Sociologists use these data to study intergenerational mobility, educational attainment, and demographic life events such as marriage and parenthood (Blossfeld et al., 2005; Bukodi & Goldthorpe, 2013). Furthermore, longitudinal frameworks allow researchers to investigate how macro-level events, such as economic recessions, public health crises, or policy reforms, affect individuals differently over time. For example, the COVID-19 pandemic was studied using longitudinal surveys to assess its impacts on mental health, employment, and inequality (Pierce et al., 2020; Daly et al., 2020). These capabilities enable robust hypothesis testing using fixed-effects models, growth curve modeling, and structural equation modeling. The inclusion of time as a dimension adds analytical depth that cross-sectional designs inherently lack, providing insights into causality and the long-term effects of social processes.

Given the increasing analytical value of longitudinal and panel data, there is a growing and sustained demand among researchers across disciplines for access to such data. As a result, the development and long-term sustainability of high-quality data sources has become a strategic priority for the scientific community. Within this context, the rising importance of FAIR principles (Findable, Accessible, Interoperable, and Reusable) and the broader Open Science movement have fostered the emergence of Research Infrastructures (RIs) specifically dedicated to longitudinal and panel data. This shift reflects a growing recognition of the limitations of traditional data sources in meeting the complex and evolving needs of social research. In fact, unlike traditional data sources such as censuses and official

registers, which do not offer the flexibility or frequency required for social research, RIs are designed to meet these demands more effectively. By leveraging the flexible and pre-planned design of longitudinal and panel surveys, RIs enable a regular, more frequent data production and dissemination of high-quality data that capture dynamic social processes over time.

This chapter examines the role of secondary data in Social Science research, the development of social RIs, some methodological challenges, and the main statistical methods associated with the analysis of longitudinal and panel data. Specifically, Sect. 2.2 explores the role of secondary data in social research, clarifying its advantages and limitations, with particular attention given to longitudinal and panel data. Section 2.3 focuses on RIs in the context of social sciences, which serve as the cornerstone for data access, preservation, and sharing, facilitating the implementation of FAIR principles and promoting cross-national comparability. Section 2.4 addresses the methodological issues of longitudinal and panel data. Particular attention is paid to strategies aimed at ensuring the inclusion of marginalized populations and reducing sampling bias. Section 2.5 presents the main statistical approaches for analyzing longitudinal and panel data, highlighting their analytical potential. Section 2.6 concludes the chapter.

2.2 The Role of Secondary Data in Social Science Research

Secondary data are used in both quantitative and qualitative research. Their utility spans multiple disciplines, including sociology, political science, economics, demography, and education. These data represent an essential resource for social scientists seeking to study complex social phenomena, as well as the individual and collective behavior of populations. Over time, their increasing demand, fueled by digitalization process, Open Science principles, and the development of Social RIs, has significantly enhanced their relevance, making them increasingly available and accessible.

Unlike primary data, which are collected based on a specific research design, secondary data refer to datasets previously gathered by other researchers or institutions for purposes different from those of the current investigation. However, they can still address new research questions. This is the reason why secondary data encompass a wide range of sources, including administrative records, sample surveys, censuses, electoral rolls, organizational databases, and publicly available datasets, such as those from the World Bank, Eurostat, or national statistical institutes. The diversity of sources, varying by institution and research topic, makes secondary data multidimensional and therefore suitable for addressing a wide array of research questions.

There are various types of secondary data, each with specific features depending on the research objectives. They differ in origin, structure, content, and purpose. For instance, cross-sectional data provide a snapshot of a phenomenon at a specific point in time, while longitudinal and panel data allow researchers to observe changes over time within the same subjects, thus enabling researchers to monitor changes in

behaviors, attitudes, social conditions, or institutional transformations. These data have contributed to reshaping the methodological approaches used to study the evolution of social phenomena over time.

In particular, they allow one to assess both immediate and delayed impacts of social interventions. Thanks to the repeated collection of information on the same statistical units, these datasets allow researchers to reconstruct individuals' life trajectories, following them through different stages of life. This approach enables a deeper understanding of personal transitions such as school-to-work pathways, marriage, retirement, or health decline.

The use of secondary data, in general, and longitudinal and panel data in particular, offers researchers many advantages beyond the wide spectrum of analyzable topics. One of the key advantages often associated with these data sources is the representativeness of the samples, i.e., the extent to which the sample reflects the characteristics of the broader population from which it is drawn. Representativeness is crucial because it allows researchers to generalize findings from the sample to the entire population with greater confidence.

However, not all longitudinal or panel datasets are representative. In fact, while many national longitudinal and panel data, such as those carried out by statistical institutions, seek representativeness through rigorous sampling methods and panel maintenance strategies, others may be designed for different purposes and may not reflect the entire population (see Sect. 2.4).

Another important aspect is related to cost and time savings. By using pre-existing datasets, researchers can avoid the financial and logistical burdens associated with primary data collection, making their research more efficient and less expensive. Furthermore, working with anonymized secondary data can mitigate ethical issues related to direct contact with vulnerable populations, who are often underrepresented or difficult to reach in surveys.

Despite their many advantages, secondary data show some limitations, such as the lack of control researchers have over how the data was collected, how variables were defined, and, particularly in administrative data, the lack of important contextual or theoretical information needed for certain types of analysis. Additionally, secondary data are often unavailable at key levels of territorial aggregation, such as provinces or municipalities in Italy, which limits the scope and granularity of spatial and policy-relevant analysis.

2.3 Research Infrastructures and Social Science

Over the last two decades, large-scale RIs have gained a central role not only in the natural and life sciences, but also within the social sciences, where the complexity of data and the ambition of comparative and longitudinal research have increased substantially. The development of RIs in the social sciences constitutes a critical foundation for advancing scientific knowledge and enabling robust secondary data analysis. These infrastructures are complex, integrated systems that encompass both material resources (such as digital archives, high-performance computing

facilities, and laboratories), and immaterial assets, including collaborative networks, standardized methodologies, and open access to curated datasets. In the context of social science research, they facilitate the systematic collection, harmonization, and dissemination of longitudinal, cross-national, and multi-level data, which are essential for comparative and policy-relevant analyses.

European initiatives, particularly through frameworks like Horizon 2020 and the European Research Infrastructure Consortium (ERIC), have played a pivotal role in institutionalizing such infrastructures. By fostering interoperability between national research systems and avoiding unnecessary duplication of efforts, these infrastructures not only enhance data quality and accessibility but also promote cumulative research and theoretical refinement. Consequently, they contribute decisively to the development of evidence-based knowledge capable of informing public policy and addressing complex societal challenges (European Commission, 2019).

Although longitudinal and panel data are essential in social science research to capture dynamic social processes, individual life-course trajectories, and the long-term effects of policy interventions, researchers often find difficulties in finding, collecting, or analyzing these kinds of data. These data are inherently complex and often require substantial effort in cleaning, harmonization, and transformation before they become suitable for analysis. As underlined by Huber et al. (2021), data preparation phase can consume up to 80% of the total research effort (Wickham, 2014; Press, 2016), thereby limiting the time and resources available for substantive analysis. It is not only time-consuming but also technically demanding, particularly when data must be manually retrieved, processed, and integrated into computational environments.

RIs have emerged as crucial facilitators in this context, especially for the hosting, curation, and long-term preservation of high-quality longitudinal datasets. Critically, RIs have the potential to realize effectively the FAIR data principles and to support the seamless integration of data into modern computational workflows (Wilkinson et al., 2025). While progress has been made, further efforts are needed to adopt common web standards and harmonized metadata practices across RIs.

Florio and Sirtori (2016) underline that although RIs can generate significant social value, their impact is highly dependent on the institutional, technical, and economic frameworks within which they operate. In the case of longitudinal and panel data in the social sciences, these challenges are particularly pronounced. First, the long time horizons required for longitudinal data collection often exceed the life cycles of research funding schemes, creating discontinuities in data collection efforts and threatening the sustainability of time-series datasets. Second, while many RIs provide access to large and complex datasets, these are frequently fragmented across national boundaries, lack harmonized documentation, or are governed by heterogeneous legal and ethical standards that complicate cross-country comparability and integration.

Moreover, although the FAIR principles offer a valuable blueprint for data stewardship, their implementation in the social sciences remains uneven (Kalinin & Skvortsov, 2023). There is a great effort to adhere to metadata standards and

apply persistent identifiers, making them available and not only linked to landing pages that require manual navigation, rather than enabling machine-readable and automated data discovery (Huber et al., 2021). This could hinder not only reproducibility but also the potential for innovative computational methods such as automated secondary analysis, federated learning, or dynamic modeling across distributed datasets.

Another pressing challenge is the gap between data repositories and computational environments (FAIR-IMPACT, 2025). As current infrastructures rarely provide integrated environments for in-situ analysis, researchers are often forced to export large datasets and replicate complex pre-processing routines in isolated local systems; raising barriers to both efficiency and transparency. Without modular, reusable software interfaces and harmonized APIs, it remains difficult to scale the use of high-quality panel data in large-scale comparative studies or to fully exploit them for policy simulation models. Recent FAIR-IMPACT case studies confirm that many RIs still lack seamless computational integration, and emphasize the need for interoperable tools that enable FAIR data to be used directly within analysis platforms.

Furthermore, ethical and legal constraints, such as those related to privacy, informed consent, and data sovereignty, add an additional layer of complexity, especially when dealing with sensitive individual-level data over extended periods. These constraints, while necessary, can obstruct data linkage or reuse, particularly when infrastructures lack clear and interoperable access protocols that balance data protection with research needs.

Addressing these challenges requires not only continued investment in the technical components of RIs but also coordinated policy action to align legal frameworks, develop sustainable funding models, and promote the co-creation of standards with user communities. Only through such an integrated approach can RIs in the social sciences fulfill their transformative potential, enabling robust, timely, and policy-relevant research grounded in rich, longitudinal data.

Enhanced interoperability would enable the automated transformation of archived longitudinal data into analysis-ready formats, fostering not only reproducible and efficient research but also facilitating advanced machine-assisted data discovery and computational analysis within the social sciences.

2.3.1 European Research Infrastructures

Among the most consolidated and strategically significant RIs in the European social science landscape is CESSDA ERIC (Consortium of European Social Science Data Archives). As a distributed infrastructure involving more than 20 member states, CESSDA serves as the primary archival backbone for social science data across Europe. It provides not only secure and sustainable long-term preservation of datasets, but also access to a vast collection of curated data via the CESSDA Data Catalogue, which includes thousands of studies in multiple

languages. Critically, CESSDA promotes metadata harmonization, persistent identifiers, and interoperability standards, thereby supporting the operationalization of FAIR principles across national borders. Through specifical tools, CESSDA actively enables researchers to discover, compare, and reuse longitudinal and cross-sectional datasets, fostering cumulative and comparative research in the social sciences.

Another emblematic example of an ERIC-compliant infrastructure that provides FAIR data is the European Social Survey (ESS), which has become a flagship resource for comparative and longitudinal research in the social sciences across Europe. Conducted biennially in over 30 countries, the ESS gathers repeated cross-sectional data using rigorous methodological standards in sampling design, questionnaire translation, and data quality control. Although not a panel survey in the strict sense, the ESS produces a harmonized time series that enables robust analysis of social change over time (European Social Survey ERIC, 2024).

In addition to these pan-European initiatives, a number of national probability-based online panels have emerged as important infrastructures for longitudinal and panel social research (e.g., LISS Panel in the Netherlands, the ELIPSS Panel in France, the GESIS Panel and GIP in Germany, the Swedish and the Norwegian Probability Panels and so on). These panels provide true longitudinal designs with high-frequency data collection thanks to the online administration, and rich contextual variables, enabling robust within-subject analyses over time. Many of these panels adopt open science practices, facilitate data linkage, and make extensive use of experimental and adaptive survey designs through open calls to the scientific community. Despite being nationally anchored, they are highly relevant to European integration efforts, especially when developed within broader consortia or comparative research programs.

A significant step toward cross-national harmonization in longitudinal infrastructure is represented by the CRONOS panel (CROss-National Online Survey Panel), a transnational project built upon the ESS. CRONOS demonstrated the feasibility of developing harmonized online panels across different national contexts, using probability-based recruitment and shared survey content. It has contributed substantially to methodological knowledge on panel retention, cross-cultural equivalence, and digital survey deployment.

In the domain of demographic and life-course research, the Generations and Gender Programme (GGP), through its Generations and Gender Survey (GGS), provides cross-national panel data on family formation, fertility, partnership dynamics, and intergenerational support. GGP is currently undergoing consolidation as a European RI, building a central hub for harmonized longitudinal data on demographic behaviors across more than 20 countries. In parallel, the GUIDE project (Growing Up In Digital Europe) aspires to become a European RI focused on children's lives. Designed as a cross-national birth cohort study, GUIDE will track children's well-being, digital engagement, and social mobility from early childhood through adolescence. Both initiatives address major policy priorities related to population ageing, education, family change, and social inequality.

Consolidated is the SHARE ERIC (Survey of Health, Ageing and Retirement in Europe), which constitutes the most comprehensive pan-European panel dataset on

individuals aged 50 and over. Conducted across more than 25 countries, SHARE combines economic, health, and social data, and includes life histories, biomarkers, and cognitive tests. Its multidisciplinary design and long-standing panel structure make it an indispensable resource for studying ageing, pension systems, health disparities, and intergenerational transfers.

Taken together, these infrastructures illustrate the maturing architecture of European social science RIs. Their growing institutionalization within the ESFRI framework and their convergence around FAIR and Open Science principles signify an ambitious effort to build an integrated European data space for the social sciences. Yet, their further success will depend on sustained political and financial support, as well as on the ability to harmonize legal, ethical, and technological systems across countries. Only through such integration can the transformative potential of longitudinal and panel data be fully harnessed to inform science, policy, and public understanding in the face of Europe's most pressing societal challenges.

2.4 Methodological Features of Longitudinal and Panel Data

Longitudinal and panel studies face a variety of interconnected challenges that impact their long-term sustainability. Addressing these issues requires continuous methodological innovation, effective engagement of the respondents, and adaptable survey designs to ensure the continuity and reliability of the data over time. On the methodological side, one of the main challenges is preserving the representativeness of the sample over time. Representativeness is crucial in socio-demographic research, as it ensures that findings can be generalised to broader populations. In longitudinal and panel designs, representativeness is not only a function of initial probabilistic sampling procedures but also of long-term participation patterns. Even a perfectly representative baseline sample can drift away from the target population if follow-up waves systematically exclude or lose specific subgroups Lynn (2021).

Socio-demographic shifts (e.g., increased mobility, migration, digital divides) make it particularly difficult to retain the representativeness of younger cohorts, non-citizens, ethnic minorities, or precariously employed individuals (Groves & Couper, 1998; Calderwood & Lessof, 2009). Adjusting weights post hoc may correct some biases, but only if the attrition is fully captured by observed variables (Little & Rubin, 2002; Lynn, 2003; Kreuter et al., 2010). This makes pro-active design strategies—such as oversampling vulnerable populations or employing flexible contact modes—essential from the outset.

Jointly with change in population structure, attrition, i.e., participants dropping out of the study over time due to refusal to continue participation, inability to locate respondents, institutionalisation or health-related nonparticipation, emigration or relocation, or mortality, is one of the most critical threats to the validity of longitudinal and panel research. Attrition is rarely random: it is often correlated with factors such as low socioeconomic status, poor health, unstable housing, or lower trust in institutions (Fitzgerald et al., 1998; Watson & Wooden, 2009). This

can introduce selection bias, undermining causal inference and generalizability. Therefore, addressing attrition is not solely a statistical problem, but also a design and ethical challenge. Retention requires regular communication, respondent incentives, and attention to participants' burden; particularly in long-term studies such as the British Household Panel Survey (BHPS)/Understanding Society or the Panel Study of Income Dynamics (PSID).

To mitigate the effects of attrition and demographic change, many longitudinal and panel studies implement refreshment samples, the deliberate introduction of new sampled respondents into an existing panel. This strategy aims to restore and maintain the representativeness of the sample, particularly as older cohorts age, populations evolve, or new societal dynamics emerge (Lynn, 2009). Refreshment sampling is a common feature in rotating panel designs such as the European Union Statistics on Income and Living Conditions (EU-SILC), where individuals are typically followed for four years before being replaced by a new cohort, ensuring that both longitudinal and cross-sectional objectives are met. Similarly, Understanding Society, in the UK, incorporated a large refreshment sample of approximately 8000 households in Wave 6 (2014–2015) (Carpenter & Deepchand, 2016), with the explicit aim of enhancing representation of ethnic minorities, while 5800 households in Wave 14 (2022–2024) aiming to incorporate new household in Great Britain (Mitchell et al., 2025). In the United States, the Panel Study of Income Dynamics (PSID) added a new immigrant sample in 1997–1999, in response to major demographic shifts and changes in immigration patterns (PSID Staff, 2000). These additions enabled researchers to maintain the national representativeness of the study despite changing population structures.

While effective in countering sample size reduction, this approach introduces methodological complexities. The incorporation of new sample members often necessitates the recalibration of survey weights to ensure consistency with population benchmarks (Sand et al., 2025; Deng et al., 2013). Furthermore, differences in exposure time between original and refreshed respondents can result in asymmetries in response conditioning, longitudinal measurement error, and panel conditioning effects, all of which require careful statistical treatment to avoid analytical bias (Das et al., 2011; Warren & Halpern-Manners, 2012). Eventually, the addition of new participants may also disrupt the continuity of life-course models or within-individual trajectories, particularly when examining long-term outcomes or cumulative exposures. However, when properly designed and integrated, refreshment samples contribute significantly to the sustainability of long-term panels and enhance the inclusion of newly relevant subpopulations, such as recent migrants, digitally engaged youth, or socially mobile individuals—groups that are otherwise difficult to track through legacy samples alone (Das et al., 2011; Warren & Halpern-Manners, 2012).

Moreover, the spacing between waves is a critical design element in longitudinal research, with significant implications for measurement quality, respondent burden, and analytical validity. Short intervals, typically defined as 6–12 months between waves, are effective in reducing recall error and capturing rapid transitions in employment, health, or household composition (Jäckle, 2006; Warren & Halpern-

Manners, 2012). However, frequent data collection can introduce respondent fatigue and increase operational costs. In contrast, long intervals of 2 years or more are standard in large-scale studies, but they raise concerns about retrospective misreporting and missed life events (Lugtig, 2014; Jäckle, 2006).

To address these challenges, many longitudinal studies employ dependent interviewing, a method that incorporates information from previous waves into current interviews. This approach has been shown to reduce reporting inconsistencies and improve data accuracy, particularly when longer intervals increase the cognitive burden on respondents (Jäckle, 2006). Furthermore, recent experimental evidence from a high-frequency German panel shows that increasing survey frequency does not necessarily lead to greater measurement error or panel conditioning. Cornesse et al. (2023) found minimal conditioning effects when surveying was increased to quarterly intervals, mainly when the questionnaire content remained consistent.

Several studies have adopted hybrid wave strategies, combining annual 'core' waves with periodic rotating modules to balance the respondent burden and the depth of information. In cases where longitudinal designs incorporate refreshment samples, these additions help counteract both attrition and the limitations of longer inter-wave gaps. Deng et al. (2013) emphasise that refreshment samples composed of newly recruited, randomly sampled participants, can provide valuable diagnostic leverage and reduce bias by offering benchmarks unexposed to panel conditioning. Together, these design innovations enhance the validity and sustainability of long-term panels, supporting the robust collection of life-course and socio-demographic data over time.

Longitudinal and panel studies have long struggled to adequately include and retain marginalised populations, including migrants, racial and ethnic minorities, homeless individuals, and low-income or digitally excluded groups. These difficulties stem not only from practical obstacles but also from deeper structural inequalities that shape research participation. Individuals in precarious housing or employment situations often experience higher geographic mobility and instability, which makes them harder to track across waves (Duvoisin et al., 2023; Watson & Wooden, 2009). Others face linguistic or cultural barriers, or may lack trust in academic or governmental institutions due to histories of discrimination or surveillance, particularly among undocumented migrants, or racialised minorities (King-Shier et al., 2017; McMichael et al., 2014). Digital divides further complicate participation in increasingly web-based longitudinal panels. People with low digital literacy or limited internet access–disproportionately older, rural, or socioeconomically disadvantaged–are less likely to respond to online surveys or remain engaged in longitudinal panels that rely on digital follow-up modes (Callegaro et al., 2015; Cornesse & Schaurer, 2021).

Gender significantly influences participation and retention in longitudinal and panel studies, with ample evidence indicating that male respondents exhibit higher attrition rates than female respondents. Systematic reviews of cohort studies show that samples with more men often suffer from reduced retention over time (Teague et al., 2018). Data from a Swiss mixed-mode youth panel support this, revealing that women are more likely to stay engaged–especially in web-based modes–due

to differences in perceived benefits, digital confidence, or time availability (Becker, 2022). Furthermore, gender intersects with other socio-demographic factors such as migration background and parental education, resulting in complex attrition patterns among younger respondents (Malschinger et al., 2023). These findings suggest the importance of incorporating gender-sensitive survey design—using flexible scheduling, multimode contacts, and gender-matching of interviewers—to mitigate biases and maintain panel representativeness.

Moreover, standard sampling frames such as population registers or household-based address lists systematically exclude certain groups—such as homeless individuals, institutionalised populations, and undocumented migrants—leading to initial undercoverage (Tourangeau, 2014). These methodological limitations intersect with structural disadvantage, reinforcing a cycle in which the most precarious individuals are both least represented in longitudinal data and most affected by the social phenomena such data aim to study.

Ensuring representativeness in longitudinal and panel studies is an ongoing methodological and ethical challenge, particularly in the context of socio-demographic change and structural inequalities. Attrition, digital exclusion, and the undercoverage of marginalised populations undermine the validity and generalizability of long-term data. While refreshment sampling, hybrid wave designs, and dependent interviewing offer partial solutions, their implementation requires careful calibration to avoid introducing new biases. Ultimately, inclusive and adaptive research strategies are essential for maintaining the analytical integrity and social relevance of longitudinal socio-demographic research.

2.5 Statistical Methods for Longitudinal and Panel Data

While in many applied scientific fields the terms longitudinal data and panel data are often used interchangeably, in the statistical literature, they have distinct meanings. As discussed by Diggle et al. (2002) and Hsiao (2014), longitudinal data broadly refer to data collected over time with the aim of capturing change and temporal dynamics. This category includes both individual-level and aggregate-level data, where time is the key analytical dimension, leading to several formal implications that require specific modeling strategies; some of which have been introduced in previous chapters.

Longitudinal data can thus be seen as an umbrella term encompassing a range of designs. Among them, panel data represent a specific case in which the same statistical units (which can also be higher-level entities such as Countries, schools, or firms) are observed repeatedly over time. In contrast, repeated cross-sectional surveys apply the same data collection instrument (e.g., a questionnaire) across multiple waves, but on different–yet comparable–samples. This design leads to a different sampling error at each wave, and is typically suited for analysing population-level trends rather than individual trajectories.

Time series data can also be framed as a subtype of longitudinal data (Singer & Willett, 2003), in which one or more macro-level variables (e.g., GDP, unemployment rate, oil price) are measured over time for an entity, such as a Country. The analytical focus in time series is generally on trend, seasonality, and autocorrelation, rather than on between- or within-unit variability and heterogeneity.

Overall, from a formal point of view, longitudinal data pose specific statistical challenges and opportunities due to their core structure. The presence, by itself, of repeated observations for each statistical unit over time, as the first argument, implies that observations should be considered non-independent. This is a clear violation of one of the assumptions of the classic statistical regression modeling, that is, the independence among observations (Flatt & Jacobs, 2019).

Often, with longitudinal data, other classes of models are proposed, such as Fixed Effects Models (FE) (Mundlak, 1978; Hedges, 1994). This class aims to control for unobserved time-invariant heterogeneity across units (e.g., individuals, households) by using only within-unit variation (Allison, 2005). It introduces indeed several elements to the classical statistical modeling, in order to account for the temporal dimension: unit-specific intercepts, or removing the temporal variable-mean for each unit (this process is called *demeaning*). This class of models has found great success in the econometric field (Wooldridge, 2010), with a wide range of applications in such contexts given the opportunity to address potential time-varying effects, but not necessarily solving problems related to missing data due to attrition, nor controlling for time-varying unobservables predictors. Moreover, non-random missingness can still bias estimates (Fitzmaurice et al., 2012).

In some cases, there are elements that suggest that covariates are not linked, in terms of statistical correlation, to the individual variability of the units. This property is called *exogeneity* (Engle et al., 1983), often tested through the Hausman test (Amini et al., 2012). In addition, if such a condition holds and part of the interest of the analysis is also set on the between-entities variation, and usually this is the case of panel data, an approach encompassing Random Effect (RE) is suitable. It also allows for proper modeling of a nested structure between individuals (e.g., units nested in a city nested in Countries).

Another class of models, i.e. multilevel models, also known as hierarchical linear models, is specifically designed to handle data with clearly defined nested structures, such as repeated observations nested within individuals, or individuals nested within higher-level units (e.g., families, classrooms, or regions) (Snijders & Bosker, 2011). While dealing with longitudinal data, this class of models is particularly useful for clearly and formally decomposing within-unit and between-unit variation, and for modeling random intercepts and slopes (Goldstein, 2011), dealing with specific covariates effects according to groups.

On the other hand, Growth Curve Models (GCM) estimate proper individual trajectories over time, modeling simultaneously the common trend and the variation in growth parameters (e.g., slope, intercept) across units (Bollen & Curran, 2006). Interestingly, those models are often linked to a Structural Equation Model (SEM) approach, dealing with indicators and latent dimensions (Preacher et al., 2018), allowing for a hierarchical structure in the predictors.

Lastly, Longitudinal Linear Mixed Models (LLMM) are an extension of linear mixed models, exhaustively suited to describe repeated measures over time, incorporating both fixed and random effects to account for correlation at two levels: within individuals and heterogeneity across them (Verbeke & Molenberghs, 2000). LLMMs are considered pretty flexible, allowing for time-varying covariates and fixed covariates such as biological sex, irregular measurement intervals, and missing data under Missing at Randoms (MAR) assumptions. Recent applications include longitudinal analyses during the COVID-19 pandemic, for example a Norwegian cohort study of 4936 adults that modeled trajectories of anxiety and depression using linear mixed-effects models with both fixed and random effects (Ebrahimi et al., 2023), and the use of an ad-hoc survey to examine the sustainable effects of chatbot-based formative feedback on learning performance (Yin et al., 2024).

While this review presents quite a few models that are cornerstones in the statistical and econometric literature for longitudinal and panel data, it is necessary to note that for such class of data recent developments open space for hybrid approaches between statistics, machine learning, and data sciences (Babii et al., 2023), and those new methodologies should not be neglected in a longer and comprehensive review. Furthermore, a broad area of statistical approaches embraces the Bayesian framework, both in parametric (Daniels & Pourahmadi, 2002) as well as non-parametric (Quintana et al., 2016) formalization, implementing also extension such as Latent Growth Curve (Zhang et al., 2007). Fundamentally, these approaches utilize the same techniques as described before, but with a Bayesian specification.

2.6 Conclusions

Social research has undergone a profound transformation in recent decades due to the increasing use of longitudinal and panel secondary data. This chapter has explored the epistemological and methodological value of such data, examining their opportunities and challenges, as well as the infrastructural and analytical context that supports their use.

Longitudinal and panel, data allowing for the analysis of individual and collective changes over time, provide valuable tools for understanding life trajectories, the effects of public policies, and long-term social processes. Their methodological strength lies in the ability to disentangle causal relationships, control for unobserved heterogeneity, and capture dynamic processes that would remain invisible in purely cross-sectional designs. By facilitating the observation of within-subject and between-subject variability across multiple waves, these data structures allow researchers to investigate life-course transitions, policy impacts, and social transformations with a level of precision that other data sources cannot offer.

The establishment of solid research infrastructures, particularly at the European level, has acted as a catalyst for the dissemination and improvement of the use of these data, promoting a culture of sharing, interoperability, and methodological

quality. In this framework, the evolution of solid RIs, particularly at the European level, supported by entities like ESFRI and driven by principles of Open Science and FAIR data, has acted as a catalyst for the dissemination and improvement of the use of these data, promoting a culture of sharing, interoperability, and methodological quality. They combine the statistical reliability of probability sampling with the operational advantages of online data collection, ensuring both timeliness and cost-effectiveness. Their growing diffusion is not only a response to logistical and financial constraints but also a reflection of a broader epistemological shift toward more agile, scalable, and responsive research tools. Their expansion marks a methodological turning point, particularly within Europe, where they benefit from strong institutional support.

Looking ahead, the future of empirical social research will increasingly depend on the sustained development of these RIs. It will be essential to invest in initiatives that ensure the sustainability and accessibility of data, as well as the adoption of flexible, inclusive, and interdisciplinary methodological approaches. Projects like FOSSR, developed in the Italian context, represent concrete examples of how the convergence of methodological rigor, technological innovation, and data governance, in order to provide high-quality data useful for a better and empirically-based understanding of social changes.

Disclaimer The opinions expressed in this article by Nicolò Marchesini are his own and do not reflect the view of ISTAT.

References

Agnoli, M. S. (2008). *Il disegno della ricerca sociale*. Carocci.

Allison, P. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary: SAS Institute.

Amini, S., Delgado, M. S., Henderson, D. J., & Parmeter, C. F. (2012). Fixed vs random: The Hausman test four decades later. In *Essays in honor of Jerry Hausman* (pp. 479–513). Emerald Group Publishing Limited.

Arulampalam, W. (2001). Is unemployment really scarring? Effects of unemployment experiences on wages. *The Economic Journal*, *111*(475), F585–F606.

Babii, A., Ball, R. T., Ghysels, E., & Striaukas, J. (2023, Jul). Panel data nowcasting: The case of price-earnings ratios. arXiv preprint.

Bartels, L. M. (2006). Three virtues of panel data for the analysis of campaign effects. In P. E. Sniderman (Ed.), *The logic of comparative social inquiry* (pp. 134–156). Wiley.

Becker, R. (2022). Gender and survey participation: An event history analysis of the gender effects of survey participation in a probability-based multi-wave panel study with a sequential mixed-mode design. *Methods, Data, Analyses*, *16*(1), 3–32.

Biolcati-Rinaldi, F., & Vezzoni, C. (2012). *L'analisi secondaria nella ricerca sociale*. Il Mulino.

Blossfeld, H.-P., Klijzing, E., Mills, M., & Kurz, K. (Eds.). (2005). *Globalization, uncertainty and youth in society: The losers in a globalizing world*. Routledge.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Wiley.

Bukodi, E., & Goldthorpe, J. H. (2013). Decomposing 'social origins': The effects of parents' class, status, and education on the educational attainment of their children. *European Sociological Review*, *29*(5), 1024–1039.

Calderwood, L., & Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data. In P. Lynn (Ed.), *Methodology of longitudinal surveys.* Wiley.

Callegaro, M., Lozar Manfreda, K., & Vehovar, V. (2015). *Web survey methodology*. Sage Publications.

Carpenter, H., & Deepchand, K. (2016). *UK household longitudinal study: Immigrant and ethnic minority boost (IEMB) technical report* (Understanding Society Technical Report No. 2017-11). Institute for Social and Economic Research, University of Essex.

Cornesse, C., Blom, A. G., Sohnius, M.-L., González Ocanto, M., Rettig, T., & Ungefucht, M. (2023). Experimental evidence on panel conditioning effects when increasing the surveying frequency in a probability-based online panel. *Survey Research Methods*, *17*(3), 323–339.

Cornesse, C., & Schaurer, I. (2021). The long-term impact of different offline recruitment strategies on participation in a probability-based online panel. *Journal of Survey Statistics and Methodology*, *9*(3), 402–427.

Daly, M., Sutin, A. R., Robinson, E., & Daly, P. J. (2020). Longitudinal changes in mental health during the covid-19 pandemic in the UK. *Nature Communications*, *11*, 5356.

Daniels, M. J., & Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, *89*(3), 553–566.

Das, M., Toepoel, V., & van Soest, A. (2011). Nonparametric tests of panel conditioning and attrition bias in panel surveys. *Sociological Methods & Research*, *40*(1), 52–95.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., & Zheng, S. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, *28*(2), 238–256.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford University Press.

Duvoisin, A., Refle, J.-E., Burton-Jeangros, C., Consoli, L., Fakhoury, J., & Jackson, Y. (2023). Recruitment and attrition for panel surveys of hard-to-reach populations: Some lessons from a longitudinal study on undocumented migrants. *Field Methods*, *36*(4), 294–310.

Ebrahimi, O. V., Hoffart, A., & Johnson, S. U. (2023). Mechanisms associated with the trajectory of depressive and anxiety symptoms: A linear mixed-effects model during the covid-19 pandemic. *Current Psychology*, *42*(34), 30696–30713.

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica: Journal of the Econometric Society*, *51*, 277–304.

European Commission. (2019). *Research infrastructures make science happen.* Publications Office of the European Union. https://doi.org/10.2777/446084 (Catalogue number KI-03-19-636-EN-N).

European Social Survey ERIC. (2024). *ESS annual activity report 2023–24.* https://www.europeansocialsurvey.org/sites/default/files/2024-12/ESS-annual-activity-report-2023-24.pdf

FAIR-IMPACT. (2025). *Fair-impact use cases & stories: Real-world fair-enabling practices across scientific domains.* European Commission. https://fair-impact.eu/sites/default/files/2025-02/UseCases_Stories_A4_February2025.pdf

Fitzgerald, J., Gottschalk, P., & Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *Journal of Human Resources*, *33*(2), 251–299.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley & Sons.

Flatt, C., & Jacobs, R. L. (2019). Principle assumptions of regression analysis: Testing, techniques, and statistical reporting of imperfect data sets. *Advances in Developing Human Resources*, *21*(4), 484–502.

Florio, M., & Sirtori, E. (2016). Social benefits and costs of large scale research infrastructures. *Technological Forecasting and Social Change*, *112*, 65–78.

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Wiley.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. Wiley.

Hedges, L. V. (1994). Fixed effects models. *The handbook of research synthesis*, *285*, 299.

Hsiao, C. (2014). *Analysis of panel data* (3rd ed.). Cambridge University Press.
Huber, R., Schäffer, B., Weigel, T., Ludwig, J., Vancauwenberghe, G., & Wubbe, M. (2021). Integrating data and analysis technologies within leading environmental research infrastructures: Challenges and approaches. *Ecological Informatics*, *61*, 101245.
Jäckle, A. (2006). *Dependent interviewing: A framework and application to current research* (ISER Working Paper No. 2006-32). University of Essex, Institute for Social and Economic Research.
Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries (QQML)*, *3*(3), 619–626.
Kalinin, N. A., & Skvortsov, N. A. (2023). Difficulties of fair principles implementation in cross-domain research infrastructures. *Lobachevskii Journal of Mathematics*, *44*, 147–156.
King-Shier, K., Lau, A., Fung, S., & LeBlanc, P. (2017). Retention of ethnic participants in longitudinal studies: A systematic review addressing challenges and effective strategies. *Journal of Immigrant and Minority Health*, *19*(6), 1530–1541.
Kreuter, F., Müller, G., & Trappmann, M. (2010). Nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, *74*(5), 880–906.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.
Lugtig, P. (2014). Panel attrition: Separating stayers, fast attriters, gradual attriters, and lurkers. *Sociological Methods & Research*, *43*(4), 699–723.
Lynn, P. (2003). Developing quality standards for cross-national survey research: Five approaches. *International Journal of Social Research Methodology*, *6*(4), 323–336.
Lynn, P. (2009). *Sample design for understanding society* (Understanding Society Working Paper No. 2009-01). University of Essex.
Lynn, P. (Ed.). (2021). *Advances in longitudinal survey methodology*. John Wiley & Sons.
Malschinger, P., Vogl, S., & Schels, B. (2023). Drop in, drop out, or stay on: Patterns and predictors of panel attrition among young people. *Österreichische Zeitschrift für Soziologie*, *48*, 427–450.
McMichael, C., Nunn, C., Gifford, S. M., & Correa-Velez, I. (2014). Studying refugee settlement through longitudinal research: Methodological and ethical insights from the good starts study. *Journal of Refugee Studies*, *28*(2), 238–257.
Mitchell, J., Cabrera Álvarez, P., & Lynn, P. (2025, June 19). *Wave 14 boost sample representativeness* (Understanding Society Working Paper Series No. 2025-09). Institute for Social and Economic Research, University of Essex.
Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: Journal of the Econometric Society*, *46*, 69–85.
Neundorf, A. & Smets, K. (2017). Political socialization and the making of citizens. In *The Oxford handbook of political behavior*. Oxford University Press.
Pierce, M., Hope, H., Ford, T., Hatch, S., Hotopf, M., John, A., Kontopantelis, E., Webb, R., Wessely, S., McManus, S., & Abel, K. M. (2020). Mental health before and during the covid-19 pandemic: A longitudinal probability sample survey of the UK population. *The Lancet Psychiatry*, *7*(10), 883–892.
Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2018). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 497–517). Guilford Press.
Press, G. (2016). *Cleaning big data: Most time-consuming, least enjoyable data science task, survey says.* https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says (Forbes)
PSID Staff. (2000). *Information on the PSID immigrant sample addition of 1997/1999* (Technical Series Paper No. 00-04). Survey Research Center, University of Michigan, Panel Study of Income Dynamics.
Quintana, F. A., Johnson, W. O., Waetjen, L. E., & B. Gold, E. (2016). Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, *111*(515), 1168–1181.
Ruspini, E. (2002). *An introduction to longitudinal research*. Routledge.
Sand, M., Bruch, C., Felderer, B., Schaurer, I., Kolb, J.-P., & Weyandt, K. (2025). Creating design weights for a panel survey with multiple refreshment samples: A general discussion with

an application to a probability-based mixed-mode panel. *Methods, Data, Analyses*, *19*, 1–19 (Special issue).

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Teague, S., Youssef, G. J., Macdonald, J. A., Sciberras, E., Shatte, A., Fuller-Tyszkiewicz, M., et al. (2018). Retention strategies in longitudinal cohort studies: A systematic review and meta-analysis. *BMC Medical Research Methodology*, *18*, 151.

Tourangeau, R. (2014). Defining hard-to-survey populations. In R. Tourangeau, B. Edwards, T. P. Johnson, K. M. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 3–20). Cambridge University Press.

Vartanian, T. P. (2011). *Secondary data analysis*. Oxford University Press.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer.

Warren, J. R., & Halpern-Manners, A. (2012). Panel conditioning in longitudinal social science surveys. *Sociological Methods & Research*, *41*(4), 491–534.

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 157–181). Wiley.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Wilkinson, S. R., Aloqalaa, M., Belhajjame, K., Crusoe, M. R., de Paula Kinoshita, B., Gadelha, L., Garijo, D., Gustafsson, O. J. R., Juty, N., Kanwal, S., & Khan, F. Z. (2025). Applying the fair principles to computational workflows. *Scientific Data*, *12*, 328.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Yin, J., Goh, T.-T., & Hu, Y. (2024). Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Transactions on Learning Technologies, 17*, 1378–1389. https://doi.org/10.1109/TLT.2024.3364015

Zhang, Z., Hamagami, F., Lijuan Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*(4), 374–383.