The NCRM wayfinder guide to moving to secondary data analysis

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Primary and secondary data

Research work in all forms and across all fields can benefit from making use of both primary and secondary data analysis. While primary data commonly refers to the data collected and used for a specific targeted research study, secondary data is the re-use or re-analysis of primary data collected elsewhere for new purposes - potentially in addition to or extending beyond the original scope of the data.

The importance of secondary data has been highlighted by COVID-19 and related research work. For example, administrative Test and Trace records are used not only for counting the number of positive cases and prevalence of COVID-19 infections, but have further been used in conjunction with other data to study the broader dynamics of the pandemic and help improve insights. We can extend the value of these data through linking and matching different secondary data sources to other external datasets such as local characteristics of neighbourhood demographics, socio-economic conditions, deprivation, or accessibility to key public amenities and health infrastructure.

As we move into the territory of including secondary data to the analysis, it is important to understand the benefits, limitations and best practices associated with it. Learning to effectively incorporate secondary data and analysis into a research project can be an important way in which to build up a robust and comprehensive study.

Relevance of secondary data analysis

Secondary data analysis can be used not only to complement ongoing work focused on a specific primary data source, but fully developed research projects in their own right can be created from a mix of secondary data sources. One of the most evident advantages of using secondary data is the cost effectiveness of not having to spend time, resources or money in collecting primary data where similar secondary data already exist. This can also provide an avenue to explore important research questions where the collection of primary data is not feasible.

In this modern age of big and accessible data, secondary data sources can be found almost anywhere and there are an ever increasing number of datasets and data products being made available to researchers. Additionally, as novel data sources continue to be created, driven by the increasing sophistication and capacity of computational and storage technologies, there are new ways in which they can be used as secondary data to generate new knowledge. There has been a recent emergence of new forms of data available such as aerial photography and remote sensing data, web scraping, or tracked anonymized mobility patterns (among many other types) - all providing new ways to explore human behaviour and answer complex research questions.

Getting creative with secondary sources of data can go a long way towards the framing and delivery of a research idea. In the times of COVID-19 where research projects and studies have been heavily impacted or delayed, exploiting all available secondary data sources can be a useful and important complement to ongoing work.

Secondary data and analyses have also benefited from the recent push towards open science and replicable research as the sharing of code and data resources has opened up the number of places where secondary data can be found. Starting off research projects under the mindset of *open and accessible*, where possible and non-sensitive data allows, contributes to this broader community of making research, data and data products available for the public good. Promoting the availability of and access to any of your own generated data or resources can often help other researchers in related fields working in the same context.

Where to find and access data sources

Secondary data can come from any number of sources, and the type and format of the data used will be determined by the research question of interest. **Census data and small area population estimates** can provide key information on local differences in demographics and socio-economic conditions. Well-designed **surveys can inform on labour force or household dynamics**. Municipalities may make important local (potentially spatial) data available such as the **location of urban infrastructure or public (health, education) amenities**. Land, air, or marine **trade and transport trajectories and flows** can highlight **mobility** or density patterns. Environmental data on **pollution, meteorological conditions, greenery and open spaces** provides important context to better understand health variations. **Social media** may be able to represent **local sentiments and reactions**. The list is extensive.

Learning to effectively work with **library catalogues, online data repositories and research registries** is an important step in finding not only the relevant literature for a research project, but also the potential set of accessible secondary data sources that have been used in similar contexts. **National, regional or municipal government open data platforms and statistical agencies** also host numerous different types of data resources and products.

It is important to consider not just the context of the data, but the underlying source and format. The potential (secondary) analysis that can be done is different depending on whether this data comes from **surveys, comprehensive registries, randomized control trials, administrative data, remotely sensed data, spatial map layers**, or any other. These may further be openly available or require a secured infrastructure (trusted research environment) for access. The scope, population and timing of the data all play a factor in the degree to which they are useful for the project at hand.

Limits and working properly with secondary data

One of the biggest drawbacks of using secondary data and analysis is being limited by the original aims of the data collection and classification - which may not align directly with your research questions. When reviewing secondary data sources it is important to understand and highlight how these differences and limitations may impact the analyses undertaken and therefore the possible findings.

This is one of the most important points of working with secondary data - making sure the data is **relevant and adequate to support the analysis of the research question**. When searching for secondary data sources, they should complement and add value into the research, the study population, and/or broader study context.

It is important to make sure that any secondary data has **sufficient detail and granularity** for the purpose of your work. This includes ensuring that the data have comparable location or population, as well as any classifications, breakdowns or typologies are adequately represented. This takes on increased importance when working with multiple secondary data sources jointly. Outputs or data must complement each other and be relevant and interpretable jointly.

Beyond this question of data-research fit, the **overall quality and (statistical) validity of the data** must also be considered - especially when working with samples or subsets of the data. This is important given that the data is by nature coming from some external source. The standard set of data diagnostics and checks should be undertaken to ensure third party data is sufficiently accurate and valid for your intended use.

Case Study - Local data spaces and COVID-19 secondary data analysis

One example COVID-19 resource drawing on secondary data analysis is the Local Data Spaces Programme. This was a collaborative initiative between the Joint Biosecurity Centre, the Office for National Statistics (ONS), and Administrative Data Research (ADR) UK which brought together researchers to analyse a variety of data for profiling and tracking rapid-response health and economic indicators across local authorities in England.

The project incorporated multiple data types; from surveys to administrative data to novel rapid indicators on mobility. A series of reports for each local authority highlighted data analytics and outcomes from COVID-

19 inequalities across demographics and geographies and broad economic vulnerability. These reports were developed using national secondary data sources focused on the comparability and differences of outcomes at the local authority level.

Working within the ONS trusted research environment, the SRS, several secondary data sources were combined. These included both those related to health outcomes such as the COVID-19 Infection Survey, NHS Test and Trace and Excess Mortalities databases, and those related to economic or socio-demographics such as small area population estimates, the Labour Force Survey, Business Impacts of COVID-19 Survey, Business Structure Database and Business Registry and Employment Survey, among others.

The benefit of secondary data here is the scope of coverage in profiling a local region from different angles to give a broad snapshot of local authorities which can then be compared.

Useful tools and resources

There are several useful tools and resources available related specifically to secondary data analysis.

The UK Data Service highlight secondary data and provide data skills training modules for working with a variety of data types:

https://ukdataservice.ac.uk/learning-hub/new-to-using-data/

ESRC sponsor the *Secondary Data Analysis Initiative* grants specifically targeted for working with secondary data:

https://esrc.ukri.org/research/our-research/secondary-data-analysis-initiative/

Local Data Spaces (LDS) Programme joint collaboration of secondary data analysis for COVID-19 outcomes and related economic challenges and recovery.

https://data.cdrc.ac.uk/dataset/local-data-spaces

References

1. Johnston, M. (2017). Secondary Data Analysis: A Method of which the Time Has Come. *Qualitative And Quantitative Methods In Libraries, 3*(3), 619-626. Retrieved from http://www.qqml-journal.net/index.php/qqml/article/view/169

This guide was produced in 2021 by Jacob Macdonald, Mark Green (University of Liverpool) and Maurizio Gibin (University College London) as part of a series produced for the Changing Research Methods for Covid-19 Research Project. We are grateful to participants in the knowledge exchange workshops for sharing their experiences.

.

National Centre for Research Methods
Social Sciences
University of Southampton
Southampton, SO17 1BJ
United Kingdom.

Web	http://www.ncrm.ac.uk
Email	info@ncrm.ac.uk?
Tel	+44 23 8059 4539
Twitter	@NCRMUK