An Introduction to SECONDARY DATA ANALYSIS with IBM SPSS STATISTICS

John MacInnes

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

First published 2017

Editor: Mila Steele
Assistant editor: John Nightingale
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Proofreader: Thea Watson
Marketing manager: Sally Ransom
Cover designer: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

Library of Congress Control Number: 2016944792

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-4462-8576-3
ISBN 978-1-4462-8577-0 (pbk)

1

SECONDARY DATA ANALYSIS

The Evidence Is Out There

introduction

There is a vast and ever growing amount of easily accessible data available for analysis on almost any topic in the social sciences. It can be used to get some sense of the dimension of an issue, or for a more thorough and careful analysis that will take a good deal more preparation and time.

The internet has made secondary data analysis much easier, and the basic skills needed to get started are simple. However, there also are all kinds of challenges to getting the most out of empirical evidence, so that there will always be new and more powerful data analysis skills to learn.

1.1 WHAT IS 'SECONDARY' DATA ANALYSIS?

A simple philosophy lies behind this book: that data analysis is something best learned by doing it. Curiosity and a capacity to be intrigued by empirical evidence are your most important resources. You'll build up your knowledge and expertise as exploring the data leads you to ask new questions and discover the technical skills you need to answer them.

Secondary data analysis simply means using evidence, usually quantitative, that someone else has collected and compiled. Many people imagine that secondary data analysis requires technical expertise that they don't have, that it takes time and skill to discover and access the relevant data or that the potential results don't justify the time invested in obtaining them. All these notions may once have had some truth in them, but the development of the internet, computing power and software, alongside a tremendous rise in the number and diversity of social surveys, has revolutionised not only the amount and range of data available, but also how easy it is to access and use. You can now become a secondary data analyst, and access useful and powerful data, in a matter of minutes, as I invite you to do in Chapter 3.

These skills are fundamental ones for all social scientists, because without such skills they are rather like a traveller who cannot read a map. The mapless tourist may happen upon interesting features of a landscape but they can get no real sense of how they might fit into the wider geography of the land they explore. The social sciences' only claim to be scientific rests on the way they use empirical evidence. Given the sheer scale of contemporary society, most of that evidence has to be quantitative. This is no criticism of qualitative work: it's just that without some quantitative context even the most perceptive ethnography is of limited use. Unfortunately, the social sciences tend to be heavy on theory and light on evidence. Theories are expounded more often than they are tested against the evidence. This is not a sustainable position for credible scientific work. Moreover, it is not necessary: the range and quality of secondary data available for social scientists to work with and use to test and elaborate their theory are growing all the time.

Collecting sound quantitative evidence is resource-intensive, technical, skilled work. It is best done by government statistical institutes, survey organisations, and consortia of experienced academics who know not only the theory but also the practice of doing it well, and who've got the resources to do so. That helps explain why most high-quality social survey research is of recent origin. There are remarkably few high-quality surveys from the period before the 1970s. It is only in the last forty years that we've seen the explosion of good survey data, together with the development of data archives to store and catalogue it, and only in the last decade has this been followed, thanks to the internet, with truly easy access to that data, access that enables everyone, with a minimum of expertise, to use it in powerful ways. There is

now a wealth of high-quality data that allows anyone to map the social world around them in unprecedented detail, so that no one unable to do this can really claim to be a social scientist.

The term 'secondary data analysis' is an unfortunate one as it implies that such analysis is somehow second best. The term is a hangover from an era in which an individual or team of social scientists themselves designed the surveys and sometimes also collected much of the data that they worked with. Surveys would usually be bespoke affairs, designed to collect data relevant to a specific study. The scientists' analysis of the results was the 'primary' activity. However, were others to discover another use for the same data and use it for this different purpose, it became 'secondary' data analysis. While the days of the bespoke survey are not over, most surveys are now omnibus affairs, collecting data on a range of subjects and explicitly designed for 'secondary' use by others. Governments who need to collect data for all kinds of purposes now feel obliged, rightly, to make data that has been collected with public resources available for others to use and explore. In a sense almost all data analysis now is 'secondary'.

As well as social surveys based on face-to-face, telephone or self-completion interviews, new sources of data, whether captured through administrative processes, social media or other methods, are growing in importance. The volume of data is growing exponentially. It has been claimed that the world now creates as much new data every two years as existed in all previous history. Like many such claims this is almost certainly an exaggeration (for example, much of that 'data' comprises spam email) and impossible to verify, but it does capture the phenomenal rate of growth of data available to contemporary social scientists if they have the imagination, energy and skill to use it.

Secondary data analysis is relatively easy; the survey designers and data collectors have done most of the hard and difficult work. However, like any skill, it takes a little effort to learn to do it well. Above all, it takes a little time to develop the experience needed to use data critically. Paradoxically, one of the most important skills a researcher can develop is not technical expertise in the location, management and analysis of data – important though that is – but the ability to keep a good grasp of its limitations. Even the best map is a drastic simplification of the terrain it represents. So it is with data. The best analysts develop a good sense of what the data does not, and cannot, reveal. They also keep in mind the data that is *not* there. That enables them to make much more powerful use of what the evidence can genuinely support.

1 2 QUICK AND DIRTY OR CAREFUL AND CAUTIOUS?

In some ways secondary data analysis is *too* easy. You can rustle up some basic information based on secondary data on almost any topic in a few minutes. This is the quantitative equivalent of looking up *Wikipedia*: it's enough to give you a rough idea of what knowledge might be out there, and if it's worth pursuing the investigation further. I am a fan of 'quick and dirty'. One of the most useful skills you can develop is to quickly scan a data source either to see if it contains the kind of information you are after and is therefore worth investigating in more depth, or to get a quick sense of whether a rough idea 'flies' and could be developed further. It is also a wonderful way to reality-check more abstract and theoretical ideas in social science. If the theory is accurate, what kind of empirical data would be consistent or inconsistent with it? Do we find any such patterns? Be sceptical of theories which do not or cannot suggest empirical results or make broad predictions. Perhaps not every theory can be tested empirically, but if it cannot be so tested then it also has to be admitted that the theory is not really a theory but

something else. Of course there is ample room for discussion about what constitutes testing. A good habit to develop is to ask of any piece of work: what is its evidence base?

Quick and dirty is fine for a first and very preliminary look. It is, however, only the very first stage of a scientific approach; the latter stages are more likely to take ten weeks than ten minutes. The difference lies in the care taken with every stage of the process, from the formulation of an exact research question, to the design of an empirical exploration or test of that question, a review of the possible data sources, careful attention to the measurement of the relevant variables, and consideration of how best to summarise and present the results.

Thorough secondary analysis takes time. Even the execution of a simple idea can require extensive data preparation and management that must be done carefully and checked for errors. Checking definitions may require you to delve deep into the data documentation, including original questionnaires and interviewer protocols, to make sure that a variable in a dataset is actually measuring what you hope it is measuring. You may need to review the sampling procedures to check that the weights supplied in the dataset are appropriate for the analysis you want to undertake, or consider whether any high-profile political events during the fieldwork period might have affected interviewees' responses.

For example, you might be interested in whether there is a relationship between age and religious belief. In ten minutes you could go to the European Social Survey website, and, using Nesstar, calculate the correlation coefficient between age and respondents' answers to the question 'How religious are you?' that were measured on a scale from 1 to 10 for the latest wave of the survey in 2012. If you did so you'd find that Pearson's *r* comes out at 0.14 across the 50,000 responses from the thirty-odd countries covered by that survey. You would thus have some preliminary rough evidence that older people *are* more likely to rate themselves as more religious, but that the relationship is not a particularly strong one.

However, this would be only the very beginning of a more thorough analysis. First, you might be interested in how the concept 'religious belief' ought to be defined and operationalised. Is it best thought of as a categorical question (either one believes in religion or one does not) or a matter of degrees of belief? If it is the latter, what might such 'degrees' comprise and what survey questions might uncover them? Would we want some corroboration of claims to belief in terms of action: declaring that one belongs to a particular religion, for example, or going to religious services, or praying? One might want to go even further and think about evidence of how far or in what ways religious belief influences a person's life: do they make decisions based on scripture, ritual or prayer, for example, or follow other ways of thinking and decision making? In other words, is their religious belief central to their social action, or, even in the case of the fervently devout, actually rather marginal to it?

Rather than focusing exclusively on the characteristics of 'believers', you would want to make *comparisons* between people with different degrees of belief or none at all, in terms of other variables such as their marital status, ethnicity, where they live, what jobs (if any) they do or their social attitudes. One might expect a range of factors other than age are correlated with religious belief. This might lead you to reduce your estimate of the impact of age itself in so far as it was correlated with these other factors. Comparison is the bread and butter of much quantitative research, since it most closely approaches the experimental method. Except in rare situations, experiments are rarely either possible or ethical in the social sciences, so that we substitute systematic observation. The basis of all systematic observation is the comparison of groups defined by the variable of interest, such as religious belief.

While many 'omnibus' surveys will have a few basic questions on religion (such as asking a person's denomination and how frequently they attend services), delving more deeply into religious belief and its correlates would probably require identifying surveys with modules on religion. For each survey it will be important to know such information as who was covered by the survey or its target population (whether children or adolescents were included, for example) and the way the questions were asked (in what order, how 'don't know' responses were dealt with, whether it was a self-completion questionnaire or an interview, whether interviewers prompted or probed). We would almost certainly want to take some account of the impact of the history of religious institutions in different countries through some kind of comparative analysis.

1.3 DATA EXPLORATION AND THEORY TESTING

Finally, the researcher may have two similar but conceptually completely distinct aims for the research. The research may be *exploratory*: reviewing promising patterns in the data with no clearly established theoretical model or hypothesis guiding that exploration. Catherine Marsh (1988) argued that this aspect of research was akin to detectives looking for evidence or clues. The aim is to collect a range of evidence that may be relevant to the subject under investigation. Alternatively, the research might have a clearly defined *hypothesis* to test. Just as a trial in a courtroom, with lawyers for prosecution and defence, uses evidence to reach a judgement about whether one particular event happened or not 'beyond reasonable doubt', to use Marsh's analogy, so too does a test of a hypothesis either fail or, provisionally, succeed. Most research involves both exploration and hypothesis testing.

What is rarely sufficiently appreciated is the danger of relying on the same data to do both activities. We can think of any dataset as a mixture of signal and noise. The signal comprises the true but invisible values of the variables we want to measure. The noise comprises all the error that gets mixed in with these true values in the process of data construction. There is no 'noise-free' data, since it is always compromised to some extent by the challenges of measurement, sampling and response. It follows that any pattern in the data consistent with a given hypothesis could be there *either* because of the signal *or* because of the noise. For example, it might just have been the case that the sample drawn for the European Social Survey in 2012 happened to contain more older religious respondents than there are in the population of Europe overall, or perhaps something in the survey instrument encouraged older people to emphasise their religiosity, or older religious people were more likely to respond than their less religious peers. The only way to deal with this is either to use one set of data for exploration, and another set for hypothesis testing, or to adopt a much higher standard of test before hypotheses are accepted. Otherwise the commendable process of data exploration can degenerate into the undesirable habit of 'data snooping'. I discuss this issue in Chapter 4 when describing significance testing and its many weaknesses if used indiscriminately.

1.4 THE SOCIAL CONSTRUCTION OF DATA

Any careful and comprehensive analysis starts out from understanding how the data it works with has been produced or, to use a popular term, 'socially constructed'. All data is produced

in this way. Surveys neither harvest facts nor automatically produce 'objective knowledge', let alone 'the truth'. However, this does not mean that the results of secondary data analysis are merely a function of the outlook or standpoint of the analyst, who has cherry-picked some 'results' that happen to fit with a pre-established theory. A good theory or claim about some aspect of how societies operate (such as a claim that religious belief is stronger or more widespread among its older members) can be compared with the evidence. Moreover, every stage of how that claim has been tested against the evidence is open to scrutiny by peers, who can replicate the analysis and debate whether the way the data was used, concepts were defined or operationalised and so on was adequate.

None of this means that secondary data analysis produces only 'superficial' knowledge. There are three main objections that have been raised against quantitative data analysis in general and secondary data analysis in particular. The first is that the way in which quantitative analysis collects data 'fragments' the inevitable complexity of social reality into discrete pieces of data, which once torn from their social context cannot reveal the texture of social life. It measures only what it is possible to measure, not what is really important. It seems to me that, on the contrary, it is this criticism that is 'superficial'.

Can something that cannot be measured be said to exist? The most basic 'measurement' that is possible of any phenomenon is categorisation and classification: whether something is an example of a wider class of objects. If something can be classified then its correlates can be measured too. There are undoubtedly social phenomena that comprise many variables and very few cases. But this is a challenge to be taken up by the refinement and elaboration of concepts in such a way that more cases can be brought into the analysis, not by retreating from the axioms of a scientific approach. Most science begins with careful description. Description inevitably requires categorisation and quantification. It degenerates into the regurgitation of trivial 'facts' only if done in the absence of some theoretical framework that establishes its potential relevance. 'Fragmentation' of data is actually a basic foundation of social scientific knowledge of any sort. Only once the data has been reduced to its constituent elements can patterns and structures within it be identified that would be invisible to a casual observer. This is the whole point of social science research.

The second objection sometimes made is that the collection of data requires an undesirable power relationship between the investigator and their respondents. There is indeed a power relationship, but how far it is undesirable is a question of the nature and purposes of the research. The power relationship is an inevitable part of the scientific process. If the scientist is not in control of this process, or responsible for it, it ceases to be scientific, no matter how positive the process might be in other ways. However, this is also a power relationship to which the interviewee gives what ought to be their informed consent, and from which they can withdraw. Social scientists, like any others, have the obligation to conduct research in an ethical way, and subject to peer review. Another, rather bizarre, variant of this argument is the proposition that structured interviewing of the kind that produces quantitative data is inherently 'masculine'. Its proponents seem blissfully unaware of the implications of the logic of this argument: the profoundly anti-feminist idea that men have a natural facility with numbers. Nor is it clear that non-quantitative forms of research escape power relationships in research, rather than reformulate them in a less formal or visible way.

The third criticism sometimes made is that quantitative data is better at answering 'what' questions rather than 'why'. It can describe social structure or regular patterns of belief or behaviour

and so on, but is less able to generate evidence about the origin of such structures or *why* such patterns of behaviour exist. Again I'm sceptical about this criticism. There is long debate in the philosophy of social science literature about the nature of knowledge, empirical evidence and processes of causation and correlation. However, the idea that quantitative evidence cannot answer 'why' questions is just wrong. There are many 'why' questions it can and does answer, often by using precisely the kind of knowledge that emerges from fragmenting social experience into discrete measurements and collecting these from respondents in highly structured ways.

Let me cite one 'why' question as an example. Why is global fertility falling? Because we have very good data on births (almost every state attempts to keep track of how many new citizens are born each year, and many link this to, for example, data about the parents) we can answer this question in great detail, comparing the strength of the impact in different countries of such factors as trends in infant perinatal and older age mortality, better women's education and employment opportunities, parents' aspirations for the education of their children, the cost of rearing children, work–life balance policies that facilitate the reconciliation of the conflicting demands of parenting and employment, progress in public health provision and the availability of and knowledge about contraception and abortion, belief that 'planning' a family is a genuine alternative to receiving 'God's will' and so on. All these are factors that can be, and have been, estimated from survey data.

1.5 THE STRUCTURE OF THIS BOOK

The focus of this book is on how to locate, access and manage data in order to analyse it effectively. It is neither a primer on social statistics, nor an introduction to SPSS as statistical software, nor a book about sampling and survey methods, nor a comprehensive guide to data analysis, but rather brings together aspects of all of these topics in order to give you the skills needed to do secondary data analysis. While it assumes no prior knowledge, it will be easier to understand if you already have some familiarity with what quantitative data is, with elementary descriptive statistics or with software packages such as Excel or SPSS, and can remember at least a little of school maths. However, it also aims to be a useful reference handbook for those more experienced in secondary data analysis that can be consulted as need be, hence the organisation of the chapters.

1.6 THE CHAPTERS

Chapter 2 is a brief introduction to surveys, quantitative methods and descriptive statistics. If you're already knowledgeable about these areas, skip this chapter. Conversely, if you know nothing about any of these topics you'll find this chapter a steep learning curve on its own; you may find it best to supplement it with some of the other reading listed at the end of the chapter. It is best used as a refresher if you have already studied these topics, or as a point of reference to remind you of the meaning of key terms or procedures as you work through the rest of the book.

Chapter 3 is an introduction to the panorama of some of the best secondary data that can be used with nothing more than a web browser. It introduces Nesstar, a web-based analysis platform that anyone can master in a couple of hours, and which is used by many data providers. It also presents some basic secondary data skills and rules of good practice to follow when accessing, analysing and presenting secondary data.

Chapter 4 introduces you to the SPSS program as a means of storing, managing, analysing and reporting on data. It does so by looking at attitudes to homosexuality in Europe, and at gender and employment. Although we start with its menu-driven interface, and using a 'practice' dataset, we move on to learning and using syntax as a quicker and more effective way of working. The chapter includes 'step by step' instructions for producing summary descriptive and inferential statistics, tables and graphics and exporting them to other applications. It also covers recoding variables and selecting subsets of data for analysis.

Datasets usually come with extensive documentation, often thousands of pages long. It is therefore important to learn how to navigate your way around such documents quickly to get to the information you need to work with a dataset, or to answer a problem you encounter when doing so. Chapter 5 suggests a dozen questions that you should know the answers to in order to analyse any dataset effectively. We then move on to using the full dataset from Round 6 (2012) of the European Social Survey. We get some practice in searching data documentation to answer some of the puzzles that secondary data analysis often throws up by looking at the correlates of depression as measured by a Depression Scale (CES-D 8) constructed from the answers to a series of questions in one of the modules of the survey. Finally we download a data extract from the US General Social Survey to look at how attitudes to mothers' working have shifted over time in the United States and discover how to make a 'codebook' for your secondary data analysis projects.

An excellent way to develop your skills in secondary data analysis is to take some published work based on a publicly available dataset and attempt to replicate the analyses contained in it. We do this with two articles in Chapter 6, on religion, ethnicity and national identity (using the UK Home Office Citizenship Survey) and on helping behaviour and attitudes (using the European Social Survey Round 3). You'll find that doing so gives you a much deeper understanding of the analytical choices faced by the original authors and the decisions they made. It also allows you to explore what the impact of making different choices would have been on the analysis, or to explore other ways of analysing the same data. Such an approach delivers a much sharper critical insight into the articles that even the closest reading of the article text could ever do.

Chapters 7 and 8 deal with data management. By this point in the book you will have come to understand how important this is. Paradoxically, the 'analysis' part of secondary data analysis takes relatively little time and effort, although it is important to choose the right kind of analysis and interpret it correctly. Rather you will find that the more challenging and time-consuming aspect lies in managing and preparing your data so that it is in a format that can be analysed in the way you want. This means more than just selecting cases or variables for analysis. Often you need to deal with weights and missing values, construct new variables using the information from several existing variables together, assemble your dataset from more than one source of data, create a new dataset out of an existing one, or merge a dataset with another one. We look at all these operations and when they need to be undertaken. We use the World Bank site to download data and build an SPSS data file that we'll use in Chapter 9. Then we look at how to handle household roster information and the 'hierarchical' nature of some of the data that you'll encounter. Finally, I stress the importance of keeping an accurate record of your work.

Chapter 9 covers ordinary least squares multiple linear regression: a long name for an analysis technique that is much less intimidating than its name implies and allows us to set up powerful 'control' conditions in observational analysis that are usually as close as social scientists can get to mimicking experimental control. We look at infant mortality and fertility across

the world, and how transforming variables (e.g. by taking their logarithm) often allows us to model associations where we are more interested in relative change than in absolute numbers, and how to deal with categorical variables by producing sets of dummies. We also look at causation and correlation, and why good evidence of the latter is not necessarily evidence of the former. Finally, we look at a range of diagnostic tests that help us to decide if a model we build of some social relationship or process using linear regression is any good.

Chapter 10 looks at one of the most widely used techniques in secondary data analysis, binary logistic regression, where the dependent variable takes only two values. I look at how such regression can be understood as a further development both of the analysis of contingency tables and of linear regression. When analysing the social attitudes or behaviour of individuals, as opposed to institutions or countries, most of the variables we deal with are categorical rather than continuous, which makes logistic regression necessary. First we look at odds, odds ratios and probabilities so that we have a clear understanding of what we are doing, and then work through the components of a logistic regression analysis and its results. Finally, in Chapters 11 and 12, we bring all our skills together to look at political activity and the 'Arab Spring' using data from the World Values Survey and completing our replication of two journal articles that we started in Chapter 6.

Chapter 13 takes stock of what you've learnt in the book, and emphasises perhaps the most important skill a secondary data analyst can nurture: healthy scepticism about the value and quality of the data they work with. As the Polish economist Kalecki once said: 'The most foolish thing to do is not to calculate. The next most foolish is to follow blindly the results of your calculations.' This does not mean that statistics are merely 'damned lies' but rather that if the social production and analysis of data are to be done well they must always be done critically, that is to say, with a sober assessment of the real difficulties of the measurement of social phenomena and a sound understanding of both the potential and inevitable limitations of the kinds of analysis we can carry out on the results of these measurements.

The book is linked to a website which has videos demonstrating all the procedures described in each of the chapters and other resources to help you develop your skills, including further practice exercises, examples of SPSS syntax, practice datasets and links to various other learning and data resources. At the end of each chapter you'll find a summary of the key concepts and skills covered in it. You may find this helps to check that you've understood the most important points from each chapter. However, you'll find that by far the best way to use this book is alongside a computer. The only good way to learn about data analysis is to do it. You could read a library of books about art, but that would be of little help in learning to draw or paint: only practice would develop the skills you need. So it is with data analysis. Like any skill that takes a little time to develop, the rewards grow as you become more proficient, but I hope you'll soon find that becoming a data explorer is just as interesting as investigating unknown corners of the earth. Don't worry about taking wrong turnings or making mistakes. Playing around with data is an excellent way to learn all about it.

A note on presentation

Throughout the text **Helvetica Neue LT Std Medium font** is used to refer to SPSS commands, menus and syntax. Bold typeface is used when referring to **variable names**, while italics are used for *emphasis*.