PART I Data Exploration

CAMBRIDGE

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Excerpt
More Information

1 Origins of Data

What data is, how to collect it, and how to assess its quality

Motivation

You want to understand whether and by how much online and offline prices differ. To that end you need data on the online and offline prices of the same products. How would you collect such data? In particular, how would you select for which products to collect the data, and how could you make sure that the online and offline prices are for the same products?

The quality of management of companies may be an important determinant of their performance, and it may be affected by a host of important factors, such as ownership or the characteristics of the managers. How would you collect data on the management practices of companies, and how would you measure the quality of those practices? In addition, how would you collect data on other features of the companies?

Part I of our textbook introduces how to think about what kind of data would help answer a question, how to collect such data, and how to start working with data. It also includes chapters that introduce important concepts and tools that are fundamental building blocks of methods that we'll introduce in the rest of the textbook.

We start our textbook by discussing how data is collected, what the most important aspects of data quality are, and how we can assess those aspects. First we introduce data collection methods and data quality because of their prime importance. Data doesn't grow on trees but needs to be collected with a lot of effort, and it's essential to have high-quality data to get meaningful answers to our questions. In the end, data quality is determined by how the data was collected. Thus, it's fundamental for data analysts to understand various data collection methods, how they affect data quality in general, and what the details of the actual collection of their data imply for its quality.

The chapter starts by introducing key concepts of data. It then describes the most important methods of data collection used in business, economics, and policy analysis, such as web scraping, using administrative sources, and conducting surveys. We introduce aspects of data quality, such as validity and reliability of variables and coverage of observations. We discuss how to assess and link data quality to how the data was collected. We devote a section to Big Data to understand what it is and how it may differ from more traditional data. This chapter also covers sampling, ethical issues, and some good practices in data collection.

This chapter includes three case studies. The case study **Finding a good deal among hotels: data collection** looks at hotel prices in a European city, using data collected from a price comparison website, to help find a good deal: a hotel that is inexpensive relative to its features. It describes the collection of the `hotels-vienna` dataset. This case study illustrates data collection from online information by web scraping. The second case study, **Comparing online and**

offline prices: data collection, describes the `billion-prices` dataset. The ultimate goal of this case study is comparing online prices and offline prices of the same products, and we'll return to that question later in the textbook. In this chapter we discuss how the data was collected, with an emphasis on what products it covered and how it measured prices. The third case study, **Management quality and firm size: data collection**, is about measuring the quality of management in many organizations in many countries. It describes the `wms-management-survey` dataset. We'll use this data in subsequent case studies, too. In this chapter we describe this survey, focusing on sampling and the measurement of the abstract concept of management quality. The three case studies illustrate the choices and trade-offs data collection involves, practical issues that may arise during implementation, and how all that may affect data quality.

Learning outcomes

After working through this chapter, you should be able to:

- understand the basic aspects of data;
- understand the most important data collection methods;
- assess various aspects of data quality based on how the data was collected;
- understand some of the trade-offs in the design and implementation of data collection;
- carry out a small-scale data collection exercise from the web or through a survey.

1.1 What Is Data?

A good definition of data is "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation" (Merriam-Webster dictionary). According to this definition, information is considered data if its content is based on some measurement ("factual") and if it may be used to support some "reasoning or discussion" either by itself or after structuring, cleaning, and analysis. There is a lot of data out there, and the amount of data, or information that can be turned into data, is growing rapidly. Some of it is easier to get and use for meaningful analysis, some of it requires a lot of work, and some of it may turn out to be useless for answering interesting questions.

An almost universal feature of data is that it rarely comes in a form that can directly help answer our questions. Instead, data analysts need to work a lot with data: structuring, cleaning, and analyzing it. Even after a lot of work, the information and the quality of information contained in the original data determines what conclusions analysts can draw in the end. That's why in this chapter, after introducing the most important elements of data, we focus on data quality and methods of data collection.

Data is most straightforward to analyze if it forms a single **data table**. A data table consists of **observations** and **variables**. Observations are also known as cases. Variables are also called features. When using the mathematical name for tables, the data table is called the data matrix. A **dataset** is a broader concept that includes, potentially, multiple data tables with different kinds of information to be used in the same analysis. We'll return to working with multiple data tables in Chapter 2.

In a data table, the rows are the observations: each row is a different observation, and whatever is in a row is information about that specific observation. Columns are variables, so that column one is variable one, column two is another variable, and so on.

A common file format for data tables is the **csv file** (for "comma separated values"). csv files are text files of a data table, with rows and columns. Rows are separated by end of line signs; columns are separated by a character called a delimiter (often a comma or a semicolon). csv files can be imported in all statistical software.

Variables are identified by names. The data table may have variable names already, and analysts are free to use those names or rename the variables. Personal taste plays a role here: some prefer short names that are easier to work with in code; others prefer long names that are more informative; yet others prefer variable names that refer to something other than their content (such as the question number in a survey questionnaire). It is good practice to include the names of the variables in the first row of a csv data table. The observations start with the second row and go on until the end of the file.

Observations are identified by **identifier** or **ID variables**. An observation is identified by a single ID variable, or by a combination of multiple ID variables. ID variables, or their combinations, should uniquely identify each observation. They may be numeric or text containing letters or other characters. They are usually contained in the first column of data tables.

We use the notation x_i to refer to the value of variable x for observation i, where i typically refers to the position of the observation in the dataset. This way i starts with 1 and goes up to the number of observations in the dataset (often denoted as n or N). In a dataset with n observations, $i = 1, 2, \ldots, n$. (Note that in some programming languages, indexing may start from 0.)

1.2 Data Structures

Observations can have a cross-sectional, time series, or a multi-dimensional structure.

Observations in **cross-sectional data**, often abbreviated as **xsec** data, come from the same time, and they refer to different units such as different individuals, families, firms, and countries. Ideally, all observations in a cross-sectional dataset are observed at the exact same time. In practice this often means a particular time interval. When that interval is narrow, data analysts treat it as if it were a single point in time.

In most cross-sectional data, the ordering of observations in the dataset does not matter: the first data row may be switched with the second data row, and the information content of the data would be the same. Cross-sectional data has the simplest structure. Therefore we introduce most methods and tools of data analysis using cross-sectional data and turn to other data structures later.

Observations in **time series data** refer to a single unit observed multiple times, such as a shop's monthly sales values. In time series data, there is a natural ordering of the observations, which is typically important for the analysis. A common abbreviation used for time series data is **tseries** data. We shall discuss the specific features of time series data in Chapter 12, where we introduce time series analysis.

Multi-dimensional data, as its name suggests, has more than one dimension. It is also called **panel data**. A common type of panel data has many units, each observed multiple times. Such data is called **longitudinal data**, or cross-section time series data, abbreviated as **xt data**. Examples include countries observed repeatedly for several years, data on employees of a firm on a monthly basis, or prices of several company stocks observed on many days.

Multi-dimensional datasets can be represented in table formats in various ways. For xt data, the most convenient format has one observation representing one unit observed at one time (country–year observations, person–month observations, company-day observations) so that one unit (country, employee, company) is represented by multiple observations. In xt data tables, observations are identified by two ID variables: one for the cross-sectional units and one for time. xt data is called **balanced** if all cross-sectional units have observations for the very same time periods. It is called unbalanced if some cross-sectional units are observed more times than others. We shall discuss other specific features of multi-dimensional data in Chapter 23 where we discuss the analysis of panel data in detail.

Another important feature of data is the level of aggregation of observations. Data with information on people may have observations at different levels: age is at the individual level, home location is at the family level, and real estate prices may be available as averages for zip code areas. Data with information on manufacturing firms may have observations at the level of plants, firms as legal entities (possibly with multiple plants), industries with multiple firms, and so on. Time series data on transactions may have observations for each transaction or for transactions aggregated over some time period.

Chapter 2, Section 2.5 will discuss how to structure data that comes with multiple levels of aggregation and how to prepare such data for analysis. As a guiding principle, the analysis is best done using data aggregated at a level that makes most sense for the decisions examined: if we wish to analyze patterns in customer choices, it is best to use customer-level data; if we are analyzing the effect of firms' decisions, it is best to use firm-level data.

Sometimes data is available at a level of aggregation that is different from the ideal level. If data is too disaggregated (i.e., by establishments within firms when decisions are made at the firm level), we may want to aggregate all variables to the preferred level. If, however, the data is too aggregated (i.e., industry-level data when we want firm-level data), there isn't much that can be done. Such data misses potentially important information. Analyzing such data may uncover interesting patterns, but the discrepancy between the ideal level of aggregation and the available level of aggregation may have important consequences for the results and has to be kept in mind throughout the analysis.

Review Box 1.1 Structure and elements of data

- Most datasets are best contained in a data table, or several data tables.
- In a data table, observations are the rows; variables are its columns.
- Notation: x_i refers to the value of variable x for observation i. In a dataset with n observations, $i = 1, 2, \ldots, n$.
- Cross-sectional (xsec) data has information on many units observed at the same time.
- Time series (tseries) data has information on a single unit observed many times.
- Panel data has multiple dimensions – often, many cross-sectional units observed many times (this is also called longitudinal or xt data).

1.A1 CASE STUDY – Finding a Good Deal among Hotels: Data Collection

Introducing the hotels-vienna dataset

The ultimate goal of our first case study is to use data on all hotels in a city to find good deals: hotels that are underpriced relative to their location and quality. We'll come back to this question and data in subsequent chapters. In the case study of this chapter, our question is how to collect data that we can then use to answer our question.

Comprehensive data on hotel prices is not available ready made, so we have to collect the data ourselves. The data we'll use was collected from a price comparison website using a web scraping algorithm (see more in Section 1.5).

The `hotels-vienna` dataset contains information on hotels, hostels, and other types of accommodation in one city, Vienna, and one weekday night, November 2017. For each accommodation, the data includes information on the name and address, the price on the night in focus, in US dollars (USD), average customer rating from two sources plus the corresponding number of such ratings, stars, distance to the city center, and distance to the main railway station.

The data includes $N = 428$ accommodations in Vienna. Each row refers to a separate accommodation. All prices refer to the same weekday night in November 2017, and the data was downloaded at the same time (within one minute). Both are important: the price for different nights may be different, and the price for the same night at the same hotel may change if looked up at a different time. Our dataset has both of these time points fixed. It is therefore a cross-section of hotels – the variables with index i denote individual accommodations, and $i = 1...428$.

The data comes in a single data table, in csv format. The data table has 429 rows: the top row for variable names and 428 hotels. After some data cleaning (to be discussed in Chapter 2, Section 2.10), the data table has 25 columns corresponding to 25 variables.

The first column is a hotel_id uniquely identifying the hotel, hostel, or other accommodation in the dataset. This is a technical number without actual meaning. We created this variable to replace names, for confidentiality reasons (see more on this in Section 1.11). Uniqueness of the identifying number is key here: every hotel has a different number. See more about such identifiers in Chapter 2, Section 2.3.

The second column is a variable that describes the type of the accommodation (i.e., hotel, hostel, or bed-and-breakfast), and the following columns are variables with the name of the city (two versions), distance to the city center, stars of the hotel, average customer rating collected by the price comparison website, the number of ratings used for that average, and price. Other variables contain information regarding the night of stay such as a weekday flag, month, and year, and the size of promotional offer if any. The file VARIABLES.xls has all the information on variables.

Table 1.1 shows what the data table looks like. The variables have short names that are meant to convey their content.

Table 1.1 List of observations

hotel_id	accom_type	country	city	city_actual	dist	stars	rating	price
21894	Apartment	Austria	Vienna	Vienna	2.7	4	4.4	81
21897	Hotel	Austria	Vienna	Vienna	1.7	4	3.9	81
21901	Hotel	Austria	Vienna	Vienna	1.4	4	3.7	85
21902	Hotel	Austria	Vienna	Vienna	1.7	3	4	83
21903	Hotel	Austria	Vienna	Vienna	1.2	4	3.9	82

Note: List of five observations with variable values. accom_type is the type of accommodation. city is the city based on the search; city_actual is the municipality.

Source: `hotels-vienna` dataset. Vienna, for a November 2017 weekday. N=428.

1.3 Data Quality

Data analysts should know their data. They should know how the data was born, with all details of measurement that may be relevant for their analysis. They should know their data better than their audience. Few things have more devastating consequences for a data analyst's reputation than someone in the audience pointing out serious measurement issues the analyst didn't consider.

Garbage in – garbage out. This summarizes the prime importance of data quality. The results of an analysis cannot be better than the data it uses. If our data is useless to answer our question, the results of our analysis are bound to be useless, no matter how fancy a method we apply to it. Conversely, with excellent data even the simplest methods may deliver very useful results. Sophisticated data analysis may uncover patterns from complicated and messy data but only if the information is there.

We list specific aspects of data quality in Table 1.2. Good data collection pays attention to these as much as possible. This list should guide data analysts on what they should know about the data they use. This is our checklist. Other people may add more items, define specific items in different ways, or de-emphasize some items. We think that our version includes the most important aspects of data quality organized in a meaningful way. We shall illustrate the use of this list by applying it in the context of the data collection methods and case studies in this book.

Table 1.2 Key aspects of data quality

Aspect	Explanation
Content	The content of a variable is determined by how it was measured, not by what it was meant to measure. As a consequence, just because a variable is given a particular name, it does not necessarily measure that.
Validity	The content of a variable (actual content) should be as close as possible to what it is meant to measure (intended content).
Reliability	Measurement of a variable should be stable, leading to the same value if measured the same way again.
Comparability	A variable should be measured the same way for all observations.
Coverage	Ideally, observations in the collected dataset should include all of those that were intended to be covered (complete coverage). In practice, they may not (incomplete coverage).
Unbiased selection	If coverage is incomplete, the observations that are included should be similar to all observations that were intended to be covered (and, thus, to those that are left uncovered).

We should note that in real life, there are problems with even the highest-quality datasets. But the existence of data problems should not deter someone from using a dataset. Nothing is perfect. It will be our job to understand the possible problems and how they affect our analysis and the conclusions we can draw from our analysis.

The following two case studies illustrate how data collection may affect data quality. In both cases, analysts carried out the data collection with specific questions in mind. After introducing the data collection projects, we shall, in subsequent sections, discuss the data collection in detail and how its various features may affect data quality. Here we start by describing the aim of each project and discussing the most important questions of data quality it had to address.

Cambridge University Press
978-1-108-48301-8 — Data Analysis for Business, Economics, and Policy
Gábor Békés , Gábor Kézdi
Excerpt
More Information

A final point on quality: as we would expect, high-quality data may well be costly to gather. These case study projects were initiated by analysts who wanted answers to questions that required collecting new data. As data analysts, we often find ourselves in such a situation. Whether collecting our own data is feasible depends on its costs, difficulty, and the resources available to us. Collecting data on hotels from a website is relatively inexpensive and simple (especially for someone with the necessary coding skills). Collecting online and offline prices and collecting data on the quality of management practices are expensive and highly complex projects that required teams of experts to work together for many years. It takes a lot of effort, resources, and luck to be able to collect such complex data; but, as these examples show, it's not impossible.

Review Box 1.2 Data quality

Important aspects of data quality include:

- content of variables: what they truly measure;
- validity of variables: whether they measure what they are supposed to;
- reliability of variables: whether they would lead to the same value if measured the same way again;
- comparability of variables: the extent to which they are measured the same way across different observations;
- coverage is complete if all observations that were intended to be included are in the data;
- data with incomplete coverage may or may not have the problem of selection bias; selection bias means that the observations in the data are systematically different from the total.

1.B1 CASE STUDY – Comparing Online and Offline Prices: Data Collection

Introducing the billion-prices dataset

The second case study is about comparing online prices and offline prices of the same products. Potential differences between online and offline prices are interesting for many reasons, including making better purchase choices, understanding the business practices of retailers, and using online data in approximating offline prices for policy analysis.

The main question is how to collect data that would allow us to compare online and offline (i.e., in-store) prices for the very same product. The hard task is to ensure that we capture many products and that they are actually the same product in both sources.

The data was collected as part of the Billion Prices Project (BPP; www.thebillionpricesproject.com), an umbrella of multiple projects that collect price data for various purposes using various methods. The online–offline project combines several data collection methods, including data collected from the web and data collected "offline" by visiting physical stores.

BPP is about measuring prices for the same products sold through different channels. The two main issues are identifying products (are they really the same?) and recording their prices. The actual content of the price variable is the price as recorded for the product that was identified.

Errors in product identification or in entering the price would lower the validity of the price measures. Recording the prices of two similar products that are not the same would be an issue, and so would be recording the wrong price (e.g., do recorded prices include taxes or temporary sales?).

The reliability of the price variable also depends on these issues (would a different measurement pick the same product and measure its price the same way?) as well as inherent variability in prices. If prices change very frequently, any particular measurement would have imperfect reliability. The extent to which the price data are comparable across observations is influenced by the extent to which the products are identified the same way and the prices are recorded the same way.

Coverage of products is an important decision of the price comparison project. Conclusions from any analysis would refer to the kinds of products the data covers.

1.C1 CASE STUDY – Management Quality and Firm Performance: Data Collection

Introducing the wms-management-survey dataset

The third case study is about measuring the quality of management in organizations. The quality of management practices are understood to be an important determinant of the success of firms, hospitals, schools, and many other organizations. Yet there is little comparable evidence of such practices across firms, organizations, sectors, or countries.

There are two research questions here: how to collect data on management quality of a firm and how to measure management practices themselves. Similarly to previous case studies, no such dataset existed before the project although management consultancies have had experience in studying management quality at firms they have advised.

The data for this case study is from a large-scale research project aiming to fill this gap. The World Management Survey (WMS; http://worldmanagementsurvey.org) collects data on management practices from many firms and other organizations across various industries and countries. This is a major international survey that combines a traditional survey methodology with other methods; see Sections 1.5 and 1.6 below on data collection methods.

The most important variables in the WMS are the management practice "scores." Eighteen such scores are in the data, each measuring the quality of management practices in an important area, such as tracking and reviewing performance, the time horizon and breadth of targets, or attracting and retaining human capital. The scores range from 1 through 5, with 1 indicating worst practice and 5 indicating best practice. Importantly, this is the intended content of the variable. The actual content is determined by how it is measured: what information is used to construct the score, where that information comes from, how the scores are constructed from that information, whether there is room for error in that process, and so on.

Having a good understanding of the actual content of these measures will inform us about their validity: how close actual content is to intended content. The details of measurement will help us assess their reliability, too: if measured again, would we get the same score or maybe a different one? Similarly, those details would inform us about the extent to which the scores are comparable – i.e., they measure the same thing, across organizations, sectors, and countries.

The goal of the WMS is to measure and compare the quality of management practices across organizations in various sectors and countries. In principle the WMS could have collected data from all organizations in all sectors and countries it targeted. Such complete coverage would have been prohibitively expensive. Instead, the survey covers a sample: a small subset of all organizations. Therefore, we need to assess whether this sample gives a good picture of the management practices of all organizations – or, in other words, if selection is unbiased. For this we need to learn how the organizations covered were selected, a question we'll return to in Section 1.8 below.

1.4 How Data Is Born: The Big Picture

Data can be collected for the purpose of the analysis, or it can be derived from information collected for other purposes.

The structure and content of data purposely collected for the analysis are usually better suited to analysis. Such data is more likely to include variables that are the focus of the analysis, measured in a way that best suits the analysis, and structured in a way that is convenient for the analysis. Frequent methods to collect data include scraping the Web for information (web scraping) or conducting a survey (see Section 1.5 and Section 1.6).

Data collected for other purposes can be also very useful to answer our inquiries. Data collected for the purpose of administering, monitoring, or controlling processes in business, public administration, or other environments are called administrative data ("admin" data). If they are related to transactions, they are also called transaction data. Examples include payment, promotion, and training data of employees of a firm; transactions using credit cards issued by a bank; and personal income tax forms submitted in a country.

Admin data usually cover a complete population: all employees in a firm, all customers of a bank, or all tax filers in a country. A special case is Big Data, to be discussed in more detail in Section 1.9, which may have its specific promises and issues due to its size and other characteristics.

Often, data collected for other purposes is available at low cost for many observations. At the same time, the structure and content of such data are usually further away from the needs of the analysis compared to purposely collected data. This trade-off has consequences that vary across data, methods, and questions to be answered.

Data quality is determined by how the data was born, and data collection affects various aspects of data quality in different ways. For example, validity of the most important variables tends to be higher in purposely collected data, while coverage tends to be more complete in admin data. However, that's not always the case, and even when it is, we shouldn't think in terms of extremes. Instead, it is best to think of these issues as part of a continuum. For example, we rarely have the variables we ideally want even if we collected the data for the purpose of the analysis, and admin data may have variables with high validity for our purposes. Or, purposely collected data may have incomplete coverage but without much selection bias, whereas admin data may be closer to complete coverage but may have severe selection bias.