Asia-Pacific
Economic Cooperation

Advancing Free Trade
for Asia-Pacific Prosperity

Toolkit on using Data Science for Competition Assessment

APEC Economic Committee

July 2022

Asia-Pacific
Economic Cooperation

Toolkit on using Data Science for Competition Assessment

APEC Economic Committee

July 2022

APEC Project: EC 05 2020A

Produced by
Philippines Project Team
Frontier Economics Pty Ltd
+61 3 9620 4488

For
Asia-Pacific Economic Cooperation Secretariat
35 Heng Mui Keng Terrace
Singapore 119616
Tel: (65) 68919 600
Fax: (65) 68919 690
Email: info@apec.org
Website: www.apec.org

APEC#222-EC-03.1

Contents

Tables

Figures

Boxes

1 Introduction

1.1 Background

Digital platforms are applications that service multiple groups of users at once through the internet, providing value to each group based on the presence of other users. This includes search engines, social media, digital content aggregators and E-commerce services.

Around the world, digital platforms have been proliferating. This growth has been driven by technological innovations that have increased online connectivity, improved the availability of supporting infrastructure such as cloud computing and data storage, and expanded data processing capabilities. Digital platforms have leveraged network effects to grow their customer base and expand their service offering. The COVID-19 pandemic has also contributed to the growth of digital platforms, with a large number of businesses and customers moving towards online transactions to comply with physical distancing measures introduced by Governments.

The growth in digital platforms has seen a commensurate increase in the data generated by these platforms. Platforms collect large amounts of data from platform users. This may include, for example:

- *Personal data* – this is data required to confirm the identity of users on the platform and is typically collected during the signup process, such as name, phone number, address, etc.
- *Operational data* – this is data which the platform requires to provide its matching and connecting service, such as product listings, service offerings and financial information including credit cards and e-payment details
- *Search data* – this is data on what products and services customers are searching for on the platform, and the location and timing of the searches
- *Transactional data* – this is data on what products and services customers have purchased on the platform, when they were purchased, and how much was paid by the customers.

Platforms collect data to provide insights on the level, structure and trends of demand and supply for products and services. A platform may use data to benchmark its growth and performance, improve the quality of its service, reduce customer acquisition costs, offer additional services, and expand into other markets.

Regulators must ensure they have the necessary knowledge and tools to assess and monitor markets effectively. With the rapid growth in digital platforms, regulators must develop robust and systematic methods for gathering and analysing large and complex data sets. This will allow regulators to monitor key developments in digital markets and understand the impact of these developments on competition in domestic markets.

1.2 Purpose of this toolkit

This Toolkit provides a practical guide for competition authorities and regulators on the quantitative tools and techniques available to gather and analyse data in competition cases, studies and impact assessments. The Toolkit includes discussion of the strengths and limitations of the tools and techniques to assist competition authorities and regulators in deciding when particular tools and techniques should be applied.

The tools and techniques set out in this Toolkit are not new. They have been (and continue to be) used by competition authorities and regulators around the world. The theory underpinning the analysis of data is grounded in established principles of economic theory, and econometric and statistical analysis. While many of the tools and techniques that are discussed in the Toolkit can be applied across different sectors and industries, the Toolkit includes specific discussion and examples covering their application in the context of digital platforms.

1.3 Scope and structure

The remainder of this Toolkit is structured as follows:

- **Section 3** considers the tools and techniques that can used to gather data, including through information request templates, web scrapping techniques, accessing government information, market studies and surveys.
- **Section 4** considers the use and development of tools, procedures and processes for cleaning datasets to remove errors and inconsistencies and to address missing values.
- **Section 5** considers the tools and techniques that can be used to analyse data for different types of competition analysis, including identifying collusion, defining markets, assessing market power, predicting outcomes of horizontal mergers and determining appropriate penalties.
- **Section 6** examines how the organisational structure of competition authorities and regulators may impact the development and use of data science techniques.

2 Description of data files

To assist in the understanding and implementation of the techniques discussed in the Toolkit, a collection of datasets has been provided so that practitioners may complete exercises as they work their way through this Toolkit.

The datasets contain hypothetical data that is similar to the kind of data that may be collected by regulators over the course of an investigation. The dataset contains:

- Weekly sales data for various products at several retail stores over a sample period. This could be provided by the stores or retail chains as part of a data request.
- Store location information, along with the group to which the store belongs. This could be provided by the stores, or by desktop research.
- Information on the location of a sample of customers of each store. This could be obtained if the stores have loyalty card information, or by regulators sampling customers at random. The locations are provided as latitude/longitude co-ordinates; in practice we might expect addresses or postcode/suburb information.

The retail data covers weekly sales for three products (Q, R and S) common to nine stores over the period January 2019 to December 2020. Total units sold, revenue and the implied average price are provided for each product for each store for each week. The locations of the stores are based on a collection of automotive fuel retailers in Melbourne, Australia; the three products are intended to represent various grades/types of automotive fuel. All data is fictional and is not intended to represent the conduct of fuel retailers in Melbourne.

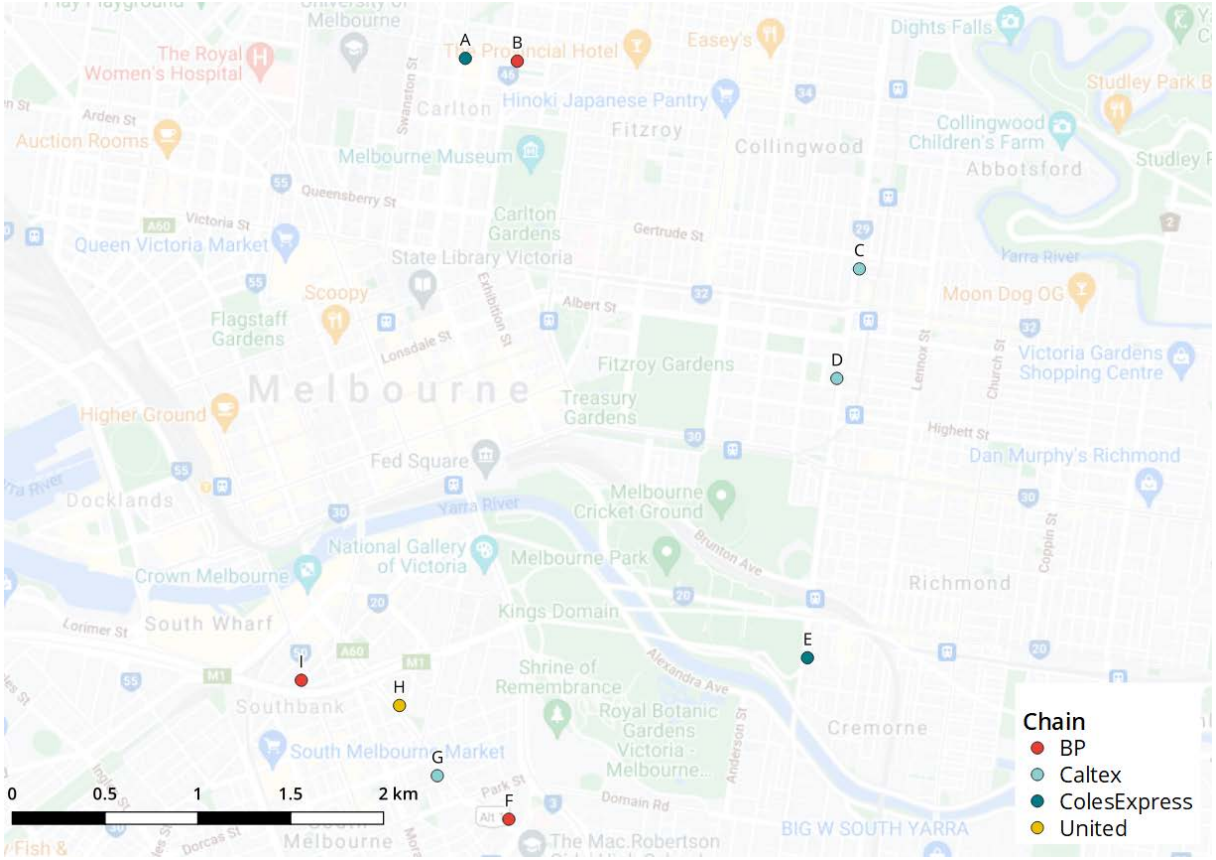
In addition, to reflect that pricing may often reflect external factors, we have included price index information published by the Australian Bureau of Statistics (ABS). We have constructed the retail data in such a way that the “Automotive fuel” price index in the ABS dataset provided is influential in setting the retail prices.

To reflect the realities of investigations, two versions of the datasets are provided. the first version contains “dirty” data before cleaning, and the second version of the datasets contains the data after cleaning. The pre-cleaning datasets have several errors in them, and the weekly retail data has been provided by each of the four retail chains in an inconsistent format. The ABS data, though free from errors, will need to be extracted and applied to the retail price data before undertaking any analysis.

By applying cleaning techniques to the original “dirty” data, it should be possible to reconstruct the cleaned datasets provided.

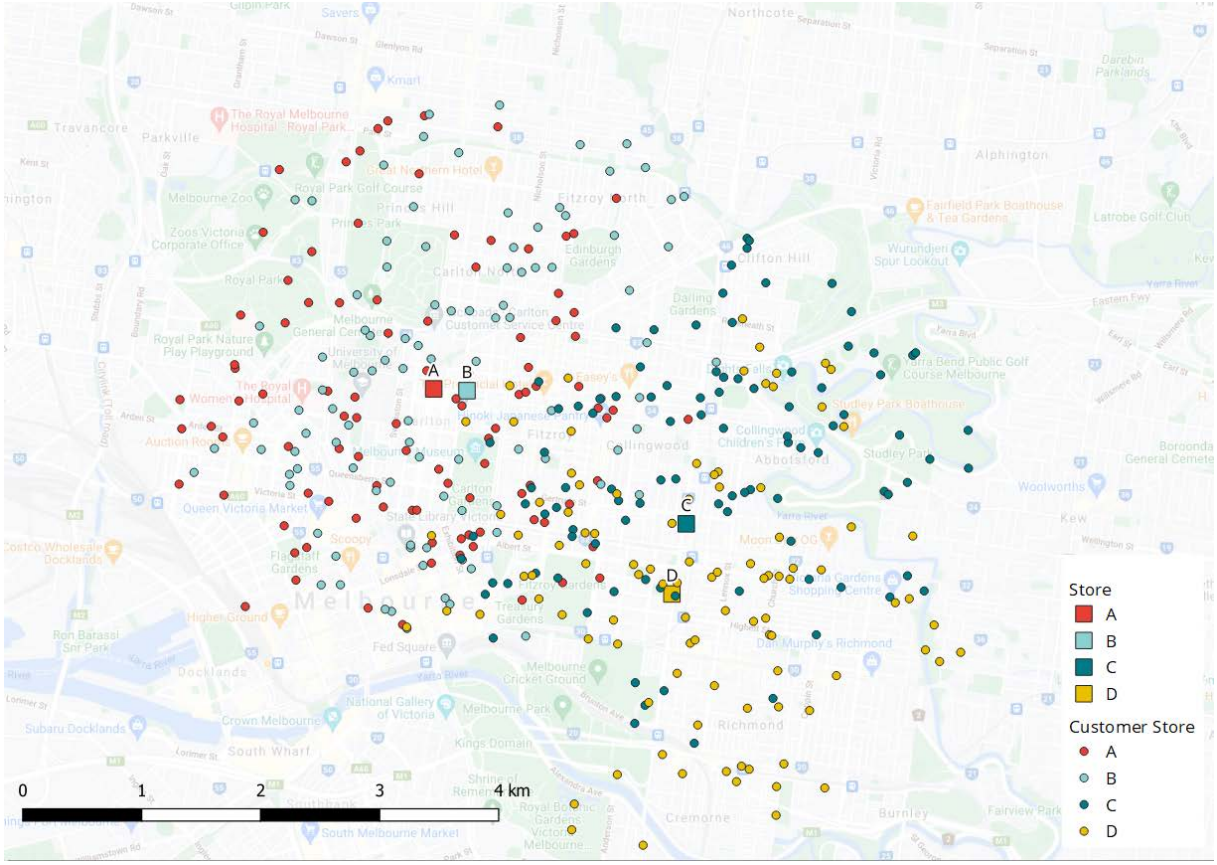
Figure 1 below shows the location of the stores in the datasets, **Figure 2** shows the geographic distribution of customers for four of the stores, revealing that customers tend to choose the store that is close to them. This implies that there may be a strong geographic component to market definition.

Figure 1: Location of stores within Melbourne



Source: Google, Frontier Economics analysis

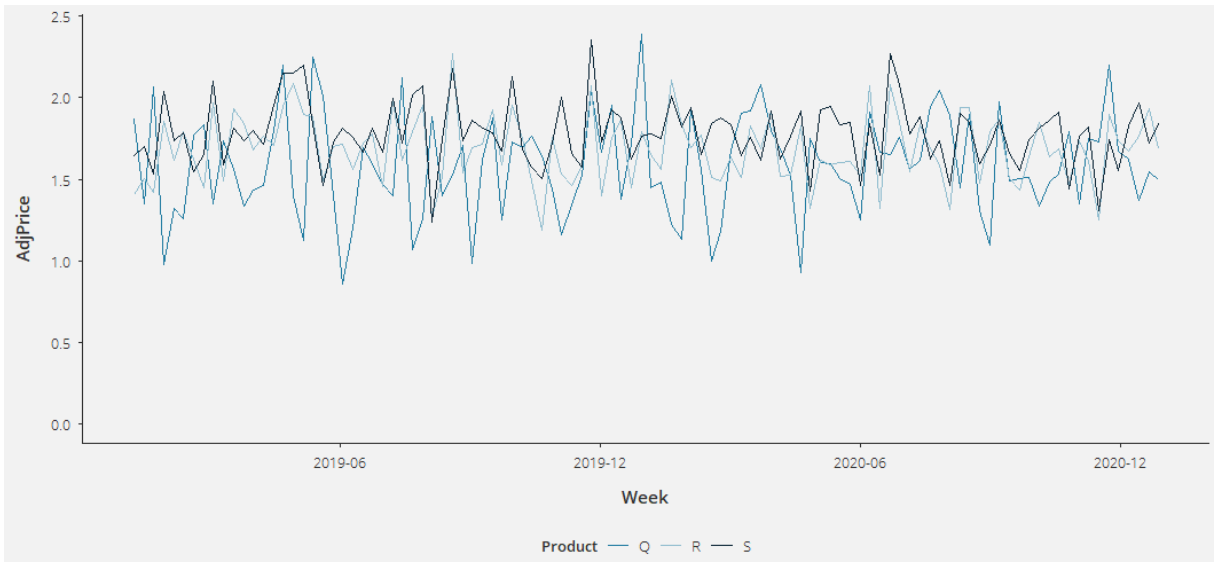
Figure 2: Location of customers of stores A, B, C and D



Source: Google, Frontier Economics analysis

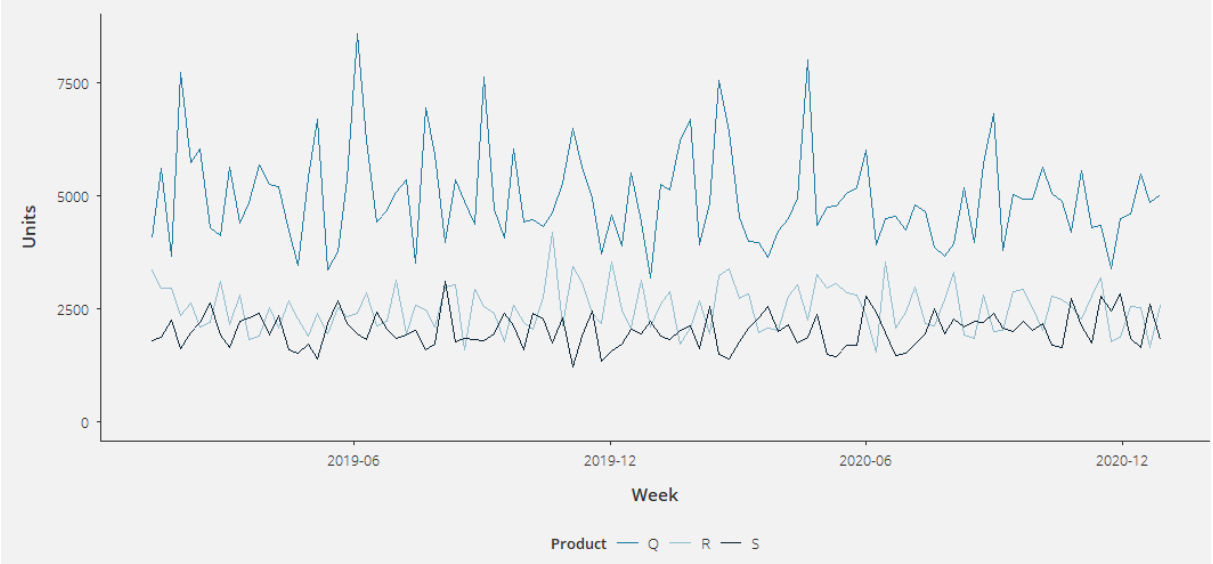
Figure 3 and **Figure 4** present the average prices and units sold of the three products over time, aggregating over all stores in the data. It can be observed that there is reasonable variation in prices over time. This is an important consideration in the ability to estimate demand elasticities when examining product market definition.

Figure 3: Average prices over time



Source: Frontier Economics analysis

Figure 4: Units sold over time



Source: Frontier Economics analysis

We have provided a number of exercises at the ends of **Sections 4, 5.2, 5.3** and **5.4** based on the datasets described above. These exercises give users of the toolkit an opportunity to apply many of the tools discussed in this Toolkit. Worked solutions for the exercises are presented in Appendix B.

3 Data gathering

3.1 Introduction

In this section, we consider the tools and techniques that can used to gather data, including through merger notification templates, web scrapping techniques, accessing government information, market studies and surveys.

3.2 Merger notification templates

3.2.1 What does this involve?

Competition agencies that adjudicate mergers generally learn of forthcoming mergers by being notified by the parties. Notification can be either compulsory (as in the Philippines and the United States) or voluntary (as in Australia and Singapore).

In jurisdictions where notification of mergers is compulsory, the competition agency produces a standard form that sets out the information that is required for the notification to be accepted (this form may be referred to as a merger notification template). The information required in a merger notification template would vary among different jurisdictions but would typically contain the general information about the transaction including the parties involved, scope and value of the transaction, and parties' operations where there might be overlapping businesses or relationships.[1]

The need for a template follows naturally from the requirement to notify: if notification is compulsory, then parties need to know the information they must supply if they are to fulfil the compulsory requirement.

In jurisdictions where notification is voluntary, the practice varies. Some agencies (such as the Competition and Consumer Commission of Singapore) specify the information that is required if the parties choose to notify; and some (such as the Australian Competition and Consumer Commission) leave the form of notification entirely to the discretion of the parties.

3.2.2 How can it be done?

The templates (where they exist) are to be found on the websites of the agencies. They generally require information of the following kinds:

- Details of the parties and how they can be contacted
- The ownership structure of the parties
- The nature of the transaction
- The reasons for the transaction
- The activities of the parties
- The markets relevant to the agency's assessment of the transaction
- The market shares in the relevant markets

[1] PCC-OP-MAO-001-F: PCC Merger Notification Form as of 27 September 2019. See: https://www.phcc.gov.ph/wp-content/uploads/2019/10/PCC-MAO-Revised-Notification-Form_27Sept2019-1.docx

- The likely effect of the transaction on competition, comparing the likely factual with the likely counterfactual
- The efficiencies that are likely to flow from the transaction.[2]

Although it is common to have a standard template for notification of a merger (a Phase 1 investigation), it is less common to have a standard template for an investigation that proceeds to a second phase. The Competition and Consumer Commission of Singapore (CCCS) is an example of an agency with a template that needs to be completed before it will start on its Phase 2 investigation. The CCCS Phase 2 template asks for few extra facts above those required by the Phase 1 notification. Rather, it asks for the applicant's opinion on various matters, such as channels of distribution, customer segments, trends in R&D, details of co-operative agreements by the parties, efficiency gains from the proposed merger and any possible failing-firm justification.

3.2.3 What are the key strengths?

As we suggested above, a template is necessary in circumstances where notification of mergers is compulsory because parties need to know the information must be supplied to fulfil the requirement to notify.

Where notification is not compulsory, there are significant advantages for the agency and for the parties in having a standard template for notification. If the parties choose to notify, they will be keen to learn what basic information the agency requires. Where the regulator is well-established and parties are advised by large law firms, the parties are likely to know what information the agency requires in a notification. However, if the agency is relatively new or parties are not being advised by major law firms, the agency will assist the parties (and itself) by providing a list of information that it requires with a notification.

3.2.4 What are the key limitations?

The key limitation of a template is that requirements for standard types of information will tend to constrain the analysis of the agency. An example is the standard requirement of agencies for data concerning market shares. Economists have known for a long while that market shares may tell us little about market power in markets characterized by a high degree of product differentiation: the more-important question is the closeness of competition of the parties to the proposed merger.[3] The danger of always asking for information concerning market shares is that staff in the agency might concentrate their consideration of the likely effects of the proposed transaction on market shares rather than on the closeness of competition of the parties.

This limitation of templates is particularly pronounced if templates are used at the start of a Phase 2 investigation. Before it proceeds to a Phase 2 investigation, the agency should have a coherent understanding of its concerns and of the information it requires before it can decide whether those concerns are valid. At the start of a Phase 2 investigation, the agency and the parties will benefit if the agency focuses its requests for information on the information it requires to decide on the key issues it has identified during its Phase 1 investigation.

3.3 Accessing public information

Public information useful to competition agencies may be published by businesses, organisations or consumers on websites or apps. This public information can be collected, collated and used by competition agencies to make evidence-based assessments of issues in competition matters.

---

[2] For examples, see: https://www.cccs.gov.sg/approach-cccs/notifying-a-merger/filing-a-merger-notification-with-cccs; https://www.phcc.gov.ph/notification-form/

[3] See, for example, Joseph Farrell and Carl Shapiro, "Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition", *The B. E. Journal of Theoretical Economics*, Vol 10 (2010) pp 1-39.

3.3.1 What does this involve?

Accessing public information on the internet involves locating relevant data, typically via a search engine or business/app website, and then collecting and collating the data from the public source.

Locating public information is best done either through known websites (e.g., from the website of a business of interest) and/or through search results from an online search engine such as Google. Some search engines enable sophisticated queries that enable searching for document types (e.g., Excel files), combinations of words or exact phrases which may assist in locating specific information or types of data.

Collecting and collating data from online sources can be simple or complex depending on how the information is published and the quantity and type of information available. Public information on the internet is published in a wide variety of forms depending on how the publisher intends the information to be consumed. Some information is designed to be consumed by human users, some information is designed to be consumed by automated processes, and some information is designed specifically to *not* be consumed by automated processes, which makes collection and collation difficult.

For the purpose of collecting data from the internet, online content can be categorised into three different forms:

- Structured data designed for programmatic access: Some content publishers make information available to the public through an interface designed to be accessed by automated systems. This is commonly known as an "Application Programming Interface" or API. APIs can make automated collection of data easy because they are designed for this purpose.[4]
- Structured data designed for human users: Many content publishers make information available to the public through web pages, which sometimes include downloadable documents. While these web pages and documents may be designed for humans to read, these can also be collected and processed automatically as long as the pages or documents in question have a common structure. This includes most web pages and common database and spreadsheet formats.
- Unstructured data: Many content publishers make information available to the public through downloadable documents that are in difficult formats to process in an automated fashion. The most common example of this is PDF documents.

Automated collection of information from the internet or "scraping" is possible for all of the content categories listed above, but best suited to the first two. In the case of unstructured data or small quantities of data, it may be quicker to process documents manually (i.e., download and extract the relevant information by hand) than to develop an automated collection system.

3.3.2 How can it be done?

Accessing public information on the internet in a manual way is self-explanatory. This section focuses on setting up automated collection or web-scraping processes for online information.

There are two important steps to automated access of public information: collecting relevant documents and processing the content. Because information is not always in a ready-to-use format, processing and collating the data is often as much or more work than the initial collection step.

There are two commonly used ways to automate the collection of relevant documents from public online sources:

[4] For example, the Google Sheets API allows users to read, write and format Google Sheets data. Details on the API and further guidance on how it may be used is available at https://developers.google.com/sheets/api

- Using point-and-click software: There are a number of applications available for user-friendly data collection from the web. This software can be found by entering “web scraping tool” into online search engines.[5]
- Using a programming or scripting language: This is a more complicated but also a more powerful option, generally speaking. There are many scripting languages with specialised scraping add-ons (typically called libraries or packages) that are well suited to this task, for example Python (programming/scripting language) and BeautifulSoup (scraping library).

There are a number of ways to process content, once it has been collected. Commonly used approaches include:

- Using specialised software: For some types of documents, there are specialised programs created specifically for extracting otherwise difficult to extract information. For example, there are a number of programs available to extract tables from PDFs or text from images (called OCR or Optical Character Recognition).[6]
- Using a programming or scripting language: For general information extraction, programming or scripting languages can be used.
- Manually processing extracted content: For example, when unstructured content is collected. This is often the case with e.g. extracting results from annual reports.

The collection and processing step can be combined in some instances, e.g., where a programming/scripting language is used.

3.3.3 What are the key strengths?

Accessing public information online, particularly through automated processes, can provide a wealth of information to competition agencies that would otherwise be unavailable or require formal requests to organisations. Accessing information from public sources can be substantially faster than going through other channels.

The use of automated processes enable access most forms of information publicly available online, even if not presented in “easy” formats such as pre-prepared databases for download.

3.3.4 What are the key limitations?

Not all publicly available information is accurate. Care must be exercised to ensure that information gathered from public sources is robust and reliable. One way to achieve this is to draw on information published by an authoritative source. This would include information from established and reputable international organisations, government agencies, and jurisdictional regulators. It would also include information published by companies and other entities as part of their legal financial reporting obligations.

Some documents are difficult to automatically extract data from. Portable Document Format files (PDFs) are the most common example of this. PDFs are a common way for content publishes to produce documents that can be read by a variety of software. However, PDFs are not structured documents in the same way that some other documents are, for example Excel documents. For this reason, it is often difficult to automatically extract information from PDFs, and in particular from tables in PDF documents.

In some cases, information published online does not have helpful documentation or associated information about the data. This may mean that, for example, the basis of published figures or the

[5] For example, https://simplescraper.io/ is a Google Chrome extension web-scraping tool with a point and click software.

[6] For example, https://pdf.online/pdf-to-excel-converter is a free OCR software that can perform conversions of PDFs to many several different file formats, including to Microsoft Word and Excel.

meaning of certain values is unclear. In this case, clarification or documentation may need to be sought from the content publisher, who may be under no obligation to comply.

Automated collection and collation are only suitable where information is published on a consistent basis. Where data is published on an inconsistent basis, e.g., in formats that change over time, manual verification may be required, negating any benefit of automated collection.

Finally, some websites take steps to prevent or limit the collection of their content. These steps may include:

- Stating that web scraping of content is not permitted (e.g., in Terms of Service).[7]
- Preventing web scraping through technical means, such as rate limiting (limiting the number of requests per minute, for example), and checking agents (ensuring the user is accessing content through a browser)
- Offering both human-readable content as well as an API, in order to prevent automated collection of the human-readable component of the site or app (Twitter is an example of this).

Another limitation of publicly available data is the authoritativeness of the data. Searching data on the internet may provide a lot of useful information, but reliability can become an issue. When such data is used in a case or any analysis, the question of liability will be raised. Thus, it must be highlighted that significant effort should be made to validate the correctness and assure authoritativeness of the data before usage.

To validate the correctness of the data, one can run a logical test on the data if there is no counter intuitive information generated from the data. Secondly a reference data from an authoritative source can be used to validate the information generated from the data. Thirdly, the profile of the data producer can be verified for authoritativeness. Lastly, one can validate the data through a direct request of documents that would authenticate the information sourced from the data producer.

3.4 Accessing government information

Government bodies, sectoral ministries and other agencies often collect data and information to inform their policies and decision-making. In some cases, this information may be useful to competition agencies in assessing competition matters.

3.4.1 What does this involve?

Generally speaking, accessing government data requires the following steps:

- A process of discovery to work out which agencies collect which data. In many cases, the data that government departments or agencies collect is not visible to those outside departments or agencies. This process of discovery will typically involve establishing contact with relevant personnel in departments likely to collect or be aware of relevant data and making inquiries. Often, even if the contacts do not collect the data themselves, they may be able to redirect inquiries or have alternative suggestions.
- Establishing a process for sharing this information. Data collected by governments is often confidential in nature and can be large in size. For these reasons, it is advisable (and departments and/or agencies may insist) that secure file-sharing platforms are used in transferring data. Furthermore, particularly for data of a confidential nature, conditions may be placed on the use and retention of the data. For example, data may need to be aggregated or anonymised before use, or

[7] These rules are often available in a websites Terms of Service, Legal, or Copyright sections. For example, the Australian Bureau of Meteorology explicitly prohibits web-scraping data from its websites (https://ideas.bom.gov.au/copyright/)

competition agencies may be required to delete government data after a certain period, e.g., six months. There are typically readily available, secure file-sharing platforms that agencies can procure for file-sharing needs.[8]

Where government data is made public, for example in government data portals[9], the data is generally easy to find and easy to access. Governments often collect data in various forms and make it available publicly, including in spreadsheets, databases, written documents, and maps.

3.4.2 How can it be done?

Once data requirements are well understood, a competition agency may seek data from government departments or agencies with the following steps.

1) Assess which department or agency, and where possible, which subdivision or organisational unit, may collect or hold relevant data. For example, electricity data may be collected by system operators or departments of energy, and transport data may be collected by public transport agencies and departments of transport.
2) Ensure data requirements can be clearly articulated in a way that the relevant departments or agencies will understand. It may be helpful to explain what the intended use of the data is to the agency to help them decide whether they can provide relevant data.
3) Contact representatives from relevant departments and explain the nature of the data being sought and the intended use of the data. At this point, representatives may not be able to help directly, but may be able to redirect inquiries to someone who can or provide alternative ideas on how to solve a particular problem.
4) Where representatives are able to provide requested data, understand the requirements around the use of this data. These requirements may include restrictions on, *inter alia,* how data is transferred, who is able to view the data, how the data is backed up, where the data is stored, how long the data can be stored, and how the data and results derived from the data can be presented. Check with the IT department that restrictions stipulated can be adhered to.
5) Establish secure file transfer arrangements, if the department or agency doesn't already have a preferred file transfer platform. Provide instructions to the department or agency providing data on how to use the secure file sharing platform chosen.
6) Confirm with representatives from the department or agency how best to deal with questions arising from the data. Ideally, the department or agency would make available a representative who understands the data well to answer questions as they arise.
7) Retrieve the data from the secure file share and establish protocols to ensure that restrictions around the use of data are adhered to.

3.4.3 What are the key strengths?

Government data can be an important source of information for competition agencies that is not available from alternate sources.

Because government departments and agencies collect and use the data themselves, this data is often provided in user-friendly formats with documentation or knowledgeable personnel available to provide assistance in understanding and using the data. For the same reason, government data is typically reliable, and/or limitations of the data are generally well understood.

3.4.4 What are the key limitations?

[8] See, for example, Intralinks (https://www.intralinks.com/), asandra (https://www.ansarada.com/) and DropBox (https://www.dropbox.com/).

[9] See, for example, https://data.gov.au

The key limitations associated with the use of government data include:

- It may be difficult to establish the existence of relevant data held by government departments or agencies.
- Data collected by government departments or agencies may be limited or narrow in focus for the government department or agency's use case.
- Data provided may be subject to stringent usage limitations for privacy, confidentiality, or other reasons.

3.5 Using market studies

3.5.1 What does this involve?

A market study is a wide-ranging investigation to determine why a market may not be working well. It can be distinguished from the investigation of actual or proposed conduct (such as a merger or instance of predatory pricing) whereby the agency seeks information to determine whether the conduct is likely to lessen competition.

Market studies can lead an agency to undertake (or to recommend that government undertake) a range of actions such as:

- do nothing
- improve access of consumers to better information
- encourage systems of self-regulation
- change legislation
- take enforcement action

3.5.2 How can it be done?

If an agency is to conduct worthwhile market studies, it must have access to information. Some agencies have no power to compel the production of information if the information is required merely for the purpose of a market study. However, some agencies have the power to compel the production of information for the purpose of a market study. For example, the Treasurer of Australia has asked the ACCC to hold inquiries in petrol pricing (in 2007), grocery prices (in 2008) and digital platform services (in 2020). For the purpose of such inquiries, the ACCC may send a notice to a person to produce information or documents. If the person fails to comply with the notice, they commit an offence. Without powers of this kind, agencies have been restricted to gathering information from companies that submit applications (for example, for a merger) or from companies which suspected of infringing the law.

3.5.3 What are the key strengths?

Market studies enable an agency to gain an in-depth understanding of patterns of competition in particular markets. This may lead to actions which range well beyond the traditional actions of mergers and enforcement. However, the studies may alert the agency to illegal conduct. Even if the study fails to disclose any illegal conduct, the study may provide the agency with information which will assist it in identifying illegal conduct which may occur in the future.

Box 1: Case study: Outcomes from the ACCC's Digital Platforms Inquiry

The ACCC was directed to consider the impact of online search engines, social media and digital content aggregators (digital platforms) on competition in the media and advertising services markets in December 2017. The final report from this inquiry was released in July 2019. The report looked at the effect that digital search engines, social media platforms and other digital content aggregation platforms have on competition in media and advertising services markets. In particular, it looked at the impact of digital platforms on the supply of news and journalistic content and the implications of this for media content creators, advertisers and consumers.

The ACCC found that bargaining power imbalances between Australian news media businesses and digital platforms, specifically Google and Facebook, prevented Google and Facebook paying properly for their use of the content of the news media. Following consultation, the ACCC made recommendations to Government. The Government considered these recommendations and developed legislation forcing Google and Facebook to negotiate with the Australian news media. Google and Facebook then reached commercial arrangements with the Australian news media.

Evidence collected during the Digital Market Inquiry also alerted the ACCC to the likelihood that Google misled Android users about the security of their personal information. After further investigation, the ACCC brought proceedings against Google and in April 2021 the Federal Court of Australia found in favour of the ACCC. The Court ruled that when consumers created a new Google Account during the initial set-up process of their Android device, Google misrepresented that the 'Location History' setting was the only Google Account setting that affected whether Google collected, kept or used personally identifiable data about their location. In fact, another Google Account setting titled 'Web & App Activity' also enabled Google to collect, store and use personally identifiable location data when it was turned on, and that setting was turned on by default.

Following this initial inquiry on digital platforms, the ACCC began a five-year program of market inquiries into digital platform services in February 2020, with particular focus on online retail marketplaces to consumers in Australia.

3.5.4 What are the key limitations?

The key limitation of market studies is their cost. These costs are borne both by the firms involved in the market and by the agency that conducts the study. There is an issue of fairness in the costs that are borne by the firms in the industry. When an agency investigates an infringement, it does so because it has reason to believe that there may have been an infringement. However, when an agency undertakes a market study, businesses may be compelled to incur costs of supplying often detailed information even though there may be no suggestion that they have infringed the law. For this reason, if the agency has no power to compel the production of information for market studies, agencies are often unable to gain access to the information that would be important for their production of useful market studies.

3.6 Using surveys

This section explains the circumstances in which surveys can assist in competition analysis and sets out principles of best practice in conducting surveys. We also discuss potential pitfalls to avoid that may undermine the results and provide a practical example of how survey data can be used.

3.6.1 What does this involve?

In simple terms, conducting a survey is the act of questioning a sample of individuals to extract information about a specific service, product or process. Data generated from surveys can assist in measuring how consumers value particular goods or services and how they make their purchasing decisions. In merger cases, surveys are often used to provide information to assist in defining the

relevant market and/or estimating the responsiveness of customers to price changes. Surveys of businesses can also assist in elucidating relationships in an industry. This can provide valuable information when a competition authority is assessing whether a transaction is likely to lead to a reduction in consumer welfare. The main issues where survey data may be helpful are in market definition and in assessing the competitive outcome of a potential merger.

Competition authorities in Europe and the U.S. frequently rely on survey data in coming to a conclusion about a particular transaction. For example, between 2004 and 2008, the Competition Commission in the U.K. used survey data in about half the merger cases when coming to a decision on whether a merger would lead to a substantial lessening of competition.[10]

However, to provide robust results, the survey questionnaire needs to be appropriately designed and analysed. This takes time and specialist skills.

3.6.2 How can it be done?

Conducting a survey usually involves the following steps:

- determine what information is required
- design the questionnaire
- carry out pre-testing using focus groups, and use the information obtained from the pre-testing to refine the questionnaire
- select a representative sample of the population
- administer the survey to the selected sample
- process the results, and
- analyse the data.

Once a decision has been made on what information is required, the way the decision situation is explained to the respondents, the wording of the questions and the layout of the questionnaire can have a considerable impact on the responses. Hence the design of the questionnaire is usually done by a specialist survey company in consultation with the competition authority. Typically, the merging parties are also offered an opportunity to provide feedback on the questionnaire, and they may try to influence the questions to achieve a particular outcome.

Pre-testing of a draft questionnaire with focus groups plays an important role in ironing out problems with the questionnaire. The wording may be obscure to some of the respondents, or the flow of questions that depend on previous answers may not seem logical.

The next step is to select the sample. At this stage it is important to clearly specify the target population relevant to the issue being investigated. The sample should be selected to reflect the target population in terms of demographic characteristics – e.g., sex, age group, geographic location. This is usually achieved by specifying segments of the population and setting targets for the number of sample respondents required in each segment. However, often it is not possible to reach the targets for some segments. To correct for this, the results of the survey need to be appropriately weighted to obtain estimates that are representative of the target population. This requires specialist statistical expertise.

It is common to administer the survey over the internet. The sample is typically selected from a large panel of people who have agreed to take part in surveys. Many survey companies have access to such panels. Sometimes the internet survey is supplemented by a survey over the phone, or in person, to represent people who do not access to the internet. Processing and analysis of the survey responses requires specialist statistical skills and is usually carried out by the survey company.

---

[10] Hurley S. (2011), The use of surveys in merger and competition analysis, *Journal of Competition Law & Economics* 7(1), 45-68.

Typically, carrying out all these steps takes several months. Hence, if a competition authority decides to commission a survey, it should allow sufficient time for all these steps to be carried out.

Box 2: Case study: The UK Competition and Markets Authority's analysis of the Cineworld / Empire merger

To inform its review of Cineworld's acquisition of five Empire cinemas in 2016, the UK Competition and Markets Authority (CMA) undertook a diversion ratio analysis to gain insight into the likely impact of the merger on ticket sales for cinemas that might be affected by the merger.[11] The information was designed to assist in determining the relevant market for the merger under consideration.

To estimate the diversion ratios, the CMA carried out a survey of customers who had visited the Cineworld Luton cinema in the last 6 months. The survey asked customers what they would have done if Cineworld Luton had been closed for refurbishment for a period of one year at the time they last visited it (i.e., a **forced diversion** question).

Respondents were given three alternative options:

(1) choose not to go to the cinema

(2) go to another cinema, or

(3) don't know.

Customers who responded that they would have gone to another cinema were then asked which other cinema they would have visited. Customers were also asked how many visits they had made to Cineworld Luton.

Analysis of survey data

The diversion ratios were estimated as the proportion of Cineworld Luton's ticket sales that would have been diverted to other cinemas in the case of the extended closure. For each alternative cinema selected by respondents, the CMA calculated a range for the diversion ratio, e.g., they estimated that 10 to 20 percent of Cineworld Luton's ticket sales would divert to Empire (see table below).[12]

Conclusions drawn from analysis of the survey data

The CMA assessed that diversion ratios for all the alternative cinemas were relatively low, due to the fact that a large proportion of respondents stated that they would not have gone to the cinema if the Cineworld Luton cinema had been closed for refurbishment for one year.

Moreover, while the CMA placed little evidential weight on the diversion ratios in its overall decision given low response rates, it concluded that the diversion ratios were consistent with other evidence that Cineworld Luton and Empire do not compete closely.

With evidence from the survey data, along with its broader assessment, the CMA ultimately concluded that the Merger did not give rise to a realistic prospect of lessening competition. As such, the merger was not referred by the CMA for further investigation.[13]

11 UK Competition and Markets Authority (Dec 2016), Decision ME/6633/16.

12 To calculate the diversion ratios, the weighted the response to the forced diversion question by the number of times that respondent had visited Cineworld Luton in the last six months. Responses from customers who indicated that they would have gone to another cinema but didn't know which cinema they would have visited, were allocated in proportion to the responses of people who did know which cinema they would go to. Diversions to other Cineworld cinemas were excluded from the analysis.

13 The CMA noted two potential limitations to its analysis of the survey data: (1) the potential for responses to the hypothetical diversion question to be different to behaviour in the event of an actual closure (known as hypothetical bias); and (2) the survey asked respondents for their responses to the closure of the relevant cinema rather than to a small but significant price increase (a SSNIP test). However, the CMA concluded that hypothetical bias is unlikely to

Table 1: Diversion from Cineworld Luton

Alternative cinema	Total diversion from Cineworld Luton
Odeon Milton Keynes	[10–20] %
Empire Hemel Hempstead	***[10–20] %***
Vue Watford	[0–5] %
Odeon Hatfield	[0–5] %
Broadway Cinema Letchworth Garden City	[0–5] %
Other third parties	[0–5] %
The Odyssey St. Albans	[0–5] %
Welwyn Garden City Cinema	[0–5] %
Not gone to the cinema	[50–60] %

Source: CMA (Dec 2016), Decision ME/6633/16

Note: Cineworld Luton and Empire Hemel Hampstead are the cinema owned by the merged entity

3.6.3 What are the key strengths?

The key strength of a survey is that it can provide data that is useful in competition cases, and which is not available from other sources. For example, how far people are prepared to travel to buy certain goods can assist in defining a geographical market. Such data can usually only be obtained through a survey.

Similarly, how people would respond to a price increase, which is required for a SSINP test, may be difficult to predict from real market data, since there may not be sufficient price variation to enable robust estimates to be made of price elasticities that measure price responses. Moreover, real market data is often confounded by many other factors that impact on purchase decisions. Appropriately designed surveys (e.g., conjoint analysis) can overcome these limitations of real market data.

3.6.4 What are the key limitations?

The key limitation of a survey is that to provide robust results it needs to be well-designed and competently analysed. This takes time and requires considerable resources and expertise. A poorly conducted survey can lead to biased and unreliable results.

Some of the pitfalls that can compromise the results of a survey include:

- Not clearly specifying the relevant target population, or selecting a sample that is not representative of the target population

be a serious issue in this instance. With respect to the second potential limitation, in a survey conducted as part of its assessment of a previous merger of cinemas, the CMA had found that the diversion question and the SSNIP question produced similar responses.

- A low response rate, which could indicate that the sample is not representative of the target population. In this case one would need to be confident that the non-respondents are similar to the respondents in respects that are relevant to the issue being investigated
- Poorly worded questions that may not be clear or meaningful to some respondents
- Inadequate statistical analysis of the survey responses, e.g., not checking for inconsistent or clearly wrong answers, or, if a segmented sample is used, not weighting the responses appropriately
- If the survey questionnaire asks respondents to provide a response to a hypothetical situation, e.g., the withdrawal of a product line or a price increase, then it is crucial that the decision context presented to the respondent is as realistically as possible.

4 Data cleaning

4.1 Introduction

Data that competition authorities may seek to use in empirical analysis may not always be clean in the sense that the data set contains no errors, inconsistencies or missing values. While data published by government agencies may have been subjected to validation procedures, other data sets, such as those provided to competition authorities by merging parties, may not have been subject to the same rigor. Ultimately the competition agencies may need to conduct data cleaning to ensure that analysis is conducted using valid data. This section will consider the use and development of tools, procedures and processes for cleaning datasets to resolve or address these problems.

Most data sets contain errors, inconsistencies, and missing values. Oftentimes, data from different sources may be combined, or merged to obtain meaningful insight. In order to make effective use of data, competition regulators should take steps to ensure that the data is reliable. Data that contains errors may result in incorrect conclusion. This process is referred to as data cleaning.

This section will consider the use and development of tools, procedures and processes for cleaning datasets to resolve or address these problems.

4.2 Approaches to cleaning data

4.2.1 What does this involve?

Cleaning data involves ensuring that data is of sufficient quality to conduct robust and valid analysis. When working with data, the analyst must be confident that the reported values are what they pertain to be, rather than an error. The issue of data cleaning may also arise when different datasets are merged or combined, and variables across the datasets (such as dates) are stored in different formats.

While there are many possible issues that one may encounter, depending on the nature of the dataset, one common issue that arises when first encountering the data is that of data formatting. Data format issues include the incorrect formatting of date: dates may be saved in the data as text, for example as 12/10/21. One must be sure as to which date this refers to in the data, for example 12 October 2021 or December 2021, and ensure that it is converted to an appropriate date format to facilitate analysis, so that all observations in December 2021 are aggregated for example. One might also encounter inconsistent date formatting within the dataset: some observations may only provide the month and year. Similarly, numerical values may be stored as character strings, this is often an indication that some observations are not numbers, for example instead of providing a single number for the number of employees a respondent instead provided a range.

There may be unit of measurement errors: when requesting revenues in thousands of dollars a respondent may simply provide the dollar amount. A product described as being a multipack may have the price of expressed as per unit or per multipack, similarly the number of units sold may be the number of units or number of multipacks.

There may be missing data, possibly due to formatting issues, which may be represented in the data as “0” or “NA” or simply blank. If the value is “0”, and the issue is not identified, the observation may have a strong influence on results due to the incorrect value.

Incorrect categorisation may occur; retail data may have applied the wrong product category to individual products. When examining products within a certain category one should exclude those products incorrectly labelled as the target category and seek to include those incorrectly labelled as not in the target category.

Date may not conform with defined business rules, constraints, or general business practices, for example transactions occurring during non-operating hours, or prices exceeding maximum allowed prices. This is indicative of a data entry error, although non-conformity may not necessarily be an error per se.

Data entry errors may occur, for example the omission of decimal points so that $1.59 is entered as $159, the latitude/longitude of a retail location may be entered incorrectly, the state of a retailer may be spelled incorrectly, or the product description is incorrect.

Finally, the format of the data may not be in a format suitable for the analysis to be performed. As an example, consider retail data where each row contains the date, the product name, the price of the product that week and the quantity of the product that period. This would be referred to as a long format. However, to perform a regression to explain the impact of prices of competing products on quantities, one would want the data in a wide format, where each row contains the date, and for each product the price and quantity that period. Similarly, the data may be in a format where each observation refers to sales at a particular store, but the analysis calls for sales data at the aggregate level.

4.2.2 How can it be done?

Data cleaning is ultimately performed on a case-by-case basis, depending greatly on the nature of the data. For example, the issues which may arise with dates are different to the issues that may arise with categorical variables. The source of the data also determines the extent of data cleaning required. On one extreme is data from highly reputable sources, regularly published and subject to internal quality assurance procedures. Such data, when correctly imported[14], may only require cursory data cleaning. On the other extreme is data entered manually by market participants.

Data cleaning is accomplished by first examining the data carefully, using software to describe the format of variables, and the range of values that are present in the data. In Stata and R for example, one can readily observe the format in which the data is stored, and can summarize numerical variables: the maximum, minimum, standard deviation etc. This may clearly reveal some data entry errors, for example a positive latitude is inconsistent with retail locations in the Southern hemisphere. One can also construct plots quickly: price, for example, is typically straightforward to obtain for retail data, and examination of histograms of prices may reveal observations with extreme prices.

There are several options for dealing with extreme quantitative observations. A rule may be adopted so as to identify which observations likely suffer from some data entry error. In the case of prices, for example, this may be done by adopting some threshold price, guided by the distribution in the data and expectations. Others may require more nuanced rules, for example dropping the observation if the ratio between store revenue and number of employees exceeds a threshold. It may be appropriate to set a conservative threshold so as to reduce the number of observations omitted but recognising the potential for data entry errors by adopting estimation techniques that reduce the impact of outliers.[15]

Reviewing in detail a subset of the data may be worthwhile. Qualitative data such as product descriptions may be subject to errors that cannot be so readily detected. Reviewing the product descriptions and comparing to other variables such as price or product category, may quickly reveal inconsistencies that will require some type of cleaning. Similarly, location data of stores may be checked using maximum/minimum latitude and longitude to ensure that it is not obviously incorrect, but to ensure that appropriate coordinates are being spot checks should be performed, by examining the location using Google Maps for example. The presence of mistakes may require substantial work to remedy.

One can identify missing data by creating a variable that indicates the presence of blank cells (or some other cell value that indicates missing data, which could even be “missing”). In the case of quantitative data, it can be helpful to tabulate the variable: the preparer of the data may have used a

[14] Data cleaning could be considered part of this process.

[15] For example, a procedure whereby a regression is first performed, then dropping the observation if the residual exceeds a threshold, then repeating the regression using the reduced sample.

specific value to indicate that the data is missing, other than blank or "NA".[16] Such observations may ultimately be dropped from the analysis if the key variable is missing.[17]

The approach whereby the format of the data is adjusted, to facilitate analysis, depends on the type of analysis to be performed, and the software used. In Excel, the pivot table functionality can be used to reshape data from a long format to a wide format; in Stata/R there are reshape functions/packages available.[18] Similarly, to aggregate data one can use pivot tables in Excel, "collapse" in Stata or alternative commands in R.

As an example, consider the data in a long format in **Table 2**. The data contains weekly data for two stores (A and B) for two products (X and Y). The data can be aggregated over stores to obtain **Table 3**, summing the quantities and revenues and deriving an aggregate price as revenue divided by quantity. This data is also in a long format. Finally, to facilitate estimating equations explaining the quantity sold of a product using the price of that product and the competing product, we would transform the aggregated data into the wide format, as shown in **Table 4**.[19]

Table 2: Long format

Date	Store	Product	Quantity	Revenue	Price
7-Jan-19	A	X	47	279.18	5.94
7-Jan-19	A	Y	85	429.25	5.05
7-Jan-19	B	X	81	473.04	5.84
7-Jan-19	B	Y	31	205.84	6.64
14-Jan-19	A	X	63	493.92	7.84
14-Jan-19	A	Y	40	220.80	5.52
14-Jan-19	B	X	67	412.05	6.15
14-Jan-19	B	Y	83	594.28	7.16

Source: Frontier Economics

16 For example, -999 or -9999 may have been used to indicate missing data.

17 Alternative solutions involve that the data provider completes the data as requested, or by applying model fitting if missing variables do not occur at random.

18 One can use other functions such as SUMIFS as a workaround to pivot tables in Excel, referring to appropriate index columns.

19 Of course, more than two observations would be required to estimate such a regression equation.

Table 3: Long format, aggregated

Date	Product	Quantity	Revenue	Price
7-Jan-19	X	128	752.22	5.88
7-Jan-19	Y	116	635.09	5.47
14-Jan-19	X	130	905.97	6.97
14-Jan-19	Y	123	815.08	6.63

Source: Frontier Economics

Table 4: Wide format

Date	Q_X	R_X	P_X	Q_Y	R_Y	P_Y
7-Jan-19	128	752.22	5.88	116	635.09	5.47
14-Jan-19	130	905.97	6.97	123	815.08	6.63

Source: Frontier Economics

It is vital to adequately document all stages of data cleaning. This requires making assumptions and decision rules explicit. This serves three purposes: reducing the chance of making mistakes in the data cleaning process by being able to review what steps have been performed and on what basis, it also facilitates quality assurance and replicability. Finally, documentation reduces the time taken to repeat the data cleaning process: it may be appropriate to repeat the data cleaning process to ensure that all errors are resolved, and revised data may be received.

4.2.3 What are the key strengths?

Performing empirical analysis using uncleaned datasets should be avoided: results may be unreliable at best and may potentially lead to incorrect conclusions. Performing data cleaning can therefore provide confidence that the conclusions are valid, and not merely a consequence of erroneous data.

Further, the data cleaning process will often provide analysts with a greater understanding of the data, which may carry over into the analysis stage, refining the approach and facilitating alternative methods of investigation.

4.2.4 What are the key limitations?

While the authority may make every reasonable effort to clean the data, it is impractical to guarantee that there are no errors in the cleaned data set: there will be the potential for minor errors to remain uncorrected by the cleaning process.[20] However, unless the errors are systematically biasing results,[21] the impact of individual data entry errors may be of little consequence and may considered analogous to the noise terms of regressions.

Further, data that appears to be inconsistent at first glance may in fact be reflecting underlying patterns. For example, one retailer setting a price for a product greater than that of competitors could

[20] During the analysis phase, robustness checks may be applied to reduce the likelihood that a conclusion results from data errors.

[21] Which raises methodological concerns.

result from a data entry error, or perhaps reflect the retailer's incentives to raise prices. A cleaning process that is too aggressive in removing outlier may remove key observations.

There is a limit to the extent to which data may be cleaned – it may be the case that the data cleaning process raises serious concerns with the reliability of the data so that the data cannot be used at all. Alternatively, adequate cleaning of the data may be sufficiently time consuming so as to be impractical.

4.3 Key takeaways

Data cleaning is an important step in the process of empirical analysis and may be required to obtain a dataset that is compatible with quantitative analysis.

The approach to data cleaning depends to a large extent on the nature and source of the data. Data from government agencies may typically be of high quality and can be assumed to be free of data entry errors, though formatting may remain a possible issue. Data requested of retailers however may depend on the accuracy with which the retailers enter data, there are many possible issues which may be checked for.

Finally, it is important that any data cleaning is well-documented.

4.4 Exercise 1: Data-cleaning exercise

Box 3: Data cleaning exercise

The dataset provided contains errors of the following kinds:

- Variable names in the retail data of the four chains are inconsistent
- The date variables in the retail data containing various formatting issues (some of which may be automatically corrected by software such as Excel), moreover some chains use the "Week Ending" convention whereas others use "Week Beginning"
- Some observations in the retail data express units are duplicated
- Some observations in the retail data express units in terms of thousands
- Some observations have incorrect/inconsistent product names
- Some observations have missing units or revenue or price fields
- Some of the stores have incorrect latitude/longitude coordinates
- Some observations in the customer location sample have incorrect latitude/longitude coordinates[22]
- Some observations in the customer location sample have incorrect/inconsistent store names

Moreover, as part of the cleaning process the four retail data datasets must be aligned and combined, similarly the appropriate automotive fuel price index in the ABS data must be extracted and cleaned to obtain values of the index for each week used in the retail data.

Cleaned dataset are provided, it is possible to recreate this dataset by applying cleaning techniques. Note that not all products are observed for all weeks for all stores; this may reflect pricing and customer choice, though one store withdrew one product for a period of time.

[22] This is limited to the sign on the latitude/longitude coordinates. Some of the locations may appear to be implausible, for example being in the middle of a river; this reflects the artificial data generation process.

5 Data analysis

5.1 Introduction

This section considers the empirical techniques that can be used to analyse typical competition issues faced by competition authorities and regulators. The issues considered are identifying collusion, defining markets, assessing market power, predicting outcomes of horizontal mergers, and determining appropriate penalties for anticompetitive conduct.

5.2 Identifying collusion

In this Toolkit, collusion refers to an agreement among competitors to co-ordinate their conduct in an attempt to replicate the outcome that would be produced by a pure monopoly.

Collusion may or may not be successful. If collusion is successful, it will remove the constraints on profits that are imposed by the forces of competition and result in monopoly prices and monopoly levels of output. If collusion is not successful, it may have no impact on prices or output. Because of this, the empirical analysis of collusion is targeted to identifying when an alleged cartel has caused prices to differ from what they otherwise would have been, and the size of the price increase (as an indication of the damage caused by the cartel).

5.2.1 Detecting cartels

Using regression analysis to detect cartels

Pricing data can be used to determine when collusion has succeeded in raising prices above the levels that would otherwise have been obtained.

The analysis should compare prices in the market that were determined outside the period of the alleged cartel conduct with prices that were determined during the period of the alleged conduct, while controlling for changes in prices that are related to changes in costs over time. If prices determined during the period of the alleged cartel were found to differ systematically from prices determined outside this period, this may support a finding that the alleged cartel caused prices to be different from what they otherwise would have been.

The method that is generally used to undertake this analysis is regression analysis. Regression analysis is a statistical methodology for investigating the relationship between an item of primary interest, in this case the price of a product, and other items, such as cost or order size, that might have an influence on the price. The item of interest is referred to as the dependent or response variable, and the items that might influence the dependent variable are referred to as the independent or explanatory variables, or drivers.

Undertaking this analysis involves 4 key steps:

1. Specifying the regression model
2. Gathering and cleaning the data
3. Estimating the parameter coefficients and testing robustness
4. Interpreting the regression results.

Each step is discussed in turn below.

Specifying the regression model

A typical regression model for detecting cartels can be written algebraically as follows:

(5.1) $$P_t = \alpha + \beta D_t + \sum_i \gamma_i (X_{it}) + \varepsilon_t$$

Where:

- P_t = price, or the dependent variable in the regression in time period t
- D_t = a dummy variable, taking a value of 1 in time period t if time period t falls during the period of the alleged cartel, and 0 otherwise
- X_{it} = a set of explanatory variables, such as cost, size of transaction, size of customer, output, location, etc that may have had an effect on the price but were outside the influence of the alleged cartel.

The terms α, β, and γ_i are coefficients which are estimated from the data, while ε_t is referred to as the residual term in the regression model. The coefficient β of the dummy variable D_t captures the change in prices in period t that were determined during the period of the alleged cartel compared to prices determined outside that period.

The key steps in specifying the regression model are as follows:

- **Identify explanatory variables** – The first step is to identify variables that might have influenced prices in the market over the period under investigation.

  It is important to undertake this step prior to undertaking any estimation. This guards against introducing variables into the relationship that appeared to be correlated with prices over the period, but that had no plausible causal relationship with price. Such variables lead to spurious results and may lead to other genuine influences on price being omitted from the relationship.

  This step is particularly important since regression analysis is only able to establish whether, and to what extent, each explanatory variable moved together with price during the period after allowing for the influence of the other explanatory variables. It is not able to establish causal relationships. A prerequisite for drawing meaningful conclusions from the regression analysis is therefore that the selection of the explanatory variables, and the interpretation of the estimates, be undertaken in conjunction with an analysis of how the market operated during the period.

- **Introduce a dummy variable** – The objective of the regression is to isolate the impact of an alleged cartel on prices.

  A dummy variable is added to the model to allow prices determined during the period of the alleged cartel to be systematically different from the prices determined from the relationship between price and the drivers of prices, i.e. the X_{it} variables. The dummy variable would take a value of either 0 or 1 in a given time period depending on whether the alleged cartel was expected to have been operating. If the prices determined during the alleged cartel period increased by more than could be explained by the price drivers included in the regression model, then we would expect the estimated coefficient on the dummy variable to be positive and statistically significant.

- **Variable transformation** – The relationship between a dependent variable and the independent variables in a regression model does not always look like a straight line. For price equations, it has been found that taking logarithms of the prices and of the X_{it} independent variables, and then fitting a straight line to the logarithms, usually produces more satisfactory models than straight line models for the prices themselves. Such models are referred to as loglinear or double-log models. This transformation is <u>not</u> applied to the dummy variable, which is kept unchanged.

  Taking logarithms of the variables enables the coefficients of the 'logged' variables in the equation to be interpreted as elasticities, which is like dealing with percentage changes in the variables, rather than changes in the actual values of the variables, e.g., a 1% change in price rather than a 1 cent change in price. The coefficient on the dummy variable, when multiplied by 100, can be

interpreted as the percentage price difference between the period of the alleged cartel and the period outside the alleged cartel.[23]

Gathering and cleaning the data

The second step in the analysis involves gathering and cleaning the data required to undertake the regression.

Typically, the required data will include:

- **Historical transactions data** – this is time series data on the price, quantity and type of good or service sold by each market participant. The data should include transactions that were undertaken both during the period of the alleged cartel and outside that period and cover a sufficiently long time period to allow trends in prices to be visible. The preferred time period for the data will ultimately depend on the facts of the case, but generally the more data the better. If, for example, the alleged cartel conduct lasted 2 years, then it may be appropriate to gather between 5 to 10 years of historical data, including the period of cartel conduct.
- **Historical cost data** – this is time series data on explanatory variables, such as cost, size of transaction, size of customer, output, location, etc that may have had an effect on the price but were outside the influence of the alleged cartel. The data should cover the same time period for which transactions data is available. In some case, manipulation of the data may be required to obtain the relevant metrics. For instance, if market participants are only willing or able to provide information on their total costs per year, one could divide this by quantities sold per year to obtain an estimate of the per unit cost of the good or service.

The tools and techniques for gathering this data are discussed in **Section 3**. It is likely that the data will be provided by market participants in different documents and different formats. As such, the relevant data will need to be consolidated into a single file using either Excel or a statistical software such as R. Most statistical programs that are used to estimate regression require the data to be laid out in a particular way, usually in a table with separate rows for each observation of the dependent variable (in this case price) and columns for each explanatory variable in the regression (including the dummy variables). The dataset should be cleaned to remove errors, inconsistencies and missing values using the techniques discussed in **Section 4**.

Estimating the parameter coefficients and testing robustness

Estimating the model involves several steps:

- Gather time series data on the dependent variable (price) and each explanatory variable. Methods for gathering data are explained in Section 3.
- Estimate the regression using a statistical software package. There are a number of software packages that can be used including Excel, Stata and R.
- Assess the robustness of the model.

A critical step in estimating the model is assessing the robustness of the model. In addition to providing the estimates of the coefficients in the regression equation, the statistical and econometric software packages used for regression analysis produce a range of supplementary information useful to the analyst in assessing the estimated regression model.

There are three key metrics used to examine the overall robustness of the model.

The **F-statistic** and its probability value or **p-value** provide information for testing a specific hypothesis, namely that all slope coefficients in the model are not different from 0. If this hypothesis, which is referred to as the null hypothesis, is true then the model does not explain the dependent

[23] This interpretation of the coefficient for the dummy variable is accurate for small price differences. For large prices differences the percentage price difference during the period of the alleged cartel is given by $(e^b - 1)$.

variable any better than chance would and should not be taken seriously. A robust model would reject this hypothesis, the implication being that the model is statistically meaningful.

The actual value of the F-statistic needs to be interpreted together with the p-value. The p-value tells us the probability, if the null hypothesis were true, of obtaining an F-statistic value that is equal to or larger than the actual F-statistic of the regression. In other words, it tells us the probability that pure chance could have produced an F-statistic value as large as the actual F-statistic of the regression. The smaller a p-value is, the stronger the case is for rejecting a null hypothesis. Statisticians usually pick a threshold level, called the level of significance, with which they compare a p-value. If the p-value is smaller than the level of significance then the null hypothesis is rejected (at that level of significance), otherwise it is accepted. Commonly used levels of significance are 5% and 1%.

The **Root MSE** or RMSE is the standard deviation of the observed residuals, and it gives us an idea of how far the actual price points deviate from corresponding prices on the fitted line. The smaller the RMSE the better the line fits the observed data. If all actual price points lie exactly on the line, then the RMSE will be zero. If the RMSE is equal to 0.052, this tells us that actual prices, on average, deviate from the fitted line by 5.2 cents.

The **R-squared** and **adjusted R-squared** measures, also known as coefficients of determination, provide similar information to the RMSE, but the R-squared measures have been normalised so that they always lie between 0 and 1, with a larger value now indicating a better fit between the line and the observed price points. In the extreme case, when all the observed prices lie exactly on the fitted line, i.e., the line fits the points perfectly, both R-squared and the adjusted R-squared will be equal to 1, while the RMSE will be equal to zero.

The difference between R-squared and the adjusted R-squared is that the adjusted R-squared overcomes a drawback of the R-squared measure, namely, that it will always become larger if one adds another independent variable to the regression model, regardless of whether the additional variable is relevant or not.[24] Similarly, the RMSE always becomes smaller if an additional independent variable is added to the model. By contrast, the adjusted R-squared will only increase when an additional independent variable added to the model improves the fit of the line to the data beyond a certain threshold.

R-squared and adjusted R-squared values are most useful when comparing different models for the same data. In practice, it may not be immediately clear which explanatory variables have an impact on price, or what type of parameter transformations are required. To overcome this, a competition authority can estimate multiple specifications of the regression, using different sets of explanatory variables and/or different parameter transformations. The model that produces the higher adjusted R-square is a better fit to the data.

There are two key metrics used to examine the significance of individual coefficients. One would only pay attention to this information if the equation as a whole has been found to provide some explanation of the dependent variable, i.e., if the p-value for the F-statistic is smaller than the chosen level of significance.

The **t-value** and **p-value** for each coefficient provide information for testing a specific hypothesis about that coefficient, namely that the coefficient is not different from 0. If this hypothesis is true, then the associated variable does not assist in explaining the dependent variable.

The t-value for any coefficient needs to be interpreted together with the associated p-value. The p-value tells us the probability that, if the null hypothesis were true, we could have obtained a t-value as large or larger (in absolute value). In other words, it gives the probability that the t-value could have been obtained by pure chance. The smaller the p-value, the more confidently we can reject the null hypothesis.

In order to determine whether a coefficient is different from 0, it is common to first decide on a level of significance. As noted above, commonly used levels of significance are 5% and 1%. If the p-value for

[24] More precisely, an additional independent variable cannot decrease the value of R-squared or increase the value of the RMSE, so they could remain unchanged. However, the conditions under which they remain unchanged are extreme and would never occur when modelling real data.

a coefficient is smaller than the chosen level of significance we say that the coefficient is significantly different from 0, or simply 'significant', at the chosen level of significance. To be significant at a particular level of significance, say 5%, means that the probability of obtaining an estimated coefficient as large as this (in absolute value) by chance is less than 5%.

The **standard error** is a measure of the precision with which each coefficient is estimated. The smaller the standard error relative to the coefficient the more precise the estimate, and the narrower the bounds of uncertainty about the estimated coefficient. Consider, for example, a scenario where two different specifications of a regression are estimated, and both include a common Variable A. The standard error of Variable A in Regression 1 is 0.007 and in Regression 2 is 0.009. In this example, the standard error of the variable is considerably smaller in the first regression, indicating that the coefficient is estimated more precisely in that equation.

Interpreting the regression results

The coefficient of interest from the regression is the coefficient β of the dummy variable D_t. If the coefficient on the price change dummy variable is positive and statistically significant, then it can be concluded that the cartel has caused prices to differ from what they otherwise would have been. This can be assessed by considering the p-value.

In order to determine whether a coefficient is different from 0, it is common to first decide on a level of significance. Commonly used levels of significance are 5% and 1%. If the p-value for the β coefficient is smaller than the chosen level of significance then we say that the coefficient is significantly different from 0, or simply 'significant', at the chosen level of significance. To be significant at a particular level of significance, say 5%, means that the probability of obtaining an estimated coefficient as large as this (in absolute value) by chance is less than 5%.

If the p-value for the coefficients β of the dummy variable D_t is positive and smaller than the chosen level of significance (e.g., 1% or 5%), then this suggests that collusion has succeeded in raising prices above the levels that would otherwise have been obtained.

5.2.2 Statistical screens to detect cartel behaviour

Two categories of cartel screening models

The literature divides cartel screening models into two categories.

Structural screens are used to identify markets which exhibit a propensity for collusion and can be used to create an initial list of industries requiring further scrutiny. Structural screens involve an analysis of market characteristics which are known to facilitate cartels, or which have been exhibited in cartelised industries in the past. These include:

- structural factors, e.g., small number of competitors, high entry barriers, frequent interactions and market transparency
- supply related factors, e.g., mature stage of an industry, low pace of innovation, symmetry and commonality of costs, and product homogeneity
- demand related factors, e.g., stable demand conditions, low demand elasticity, buying power and the absence of network effects.

Behavioural screens are used to identify whether collusive behaviour has affected a specific market. Behavioural screens involve an analysis of bidding behaviour across specific tenders to determine whether suspicious behaviour is more consistent with competition or collusion.

A number of competition authorities have designed and applied pro-active screening models to detect cartel behaviour. These include the UK, Singapore, Russia, the Republic of Korea, and Brazil. To-date, these models are focused on behavioural screens to detect bid-rigging behaviour in public procurement processes. Each economy takes a different approach to designing and implementing their software screening tools. However, in general terms, the tools are likely to consider parameters across the following categories: number and pattern of bidders; suspicious pricing patterns; low endeavour submissions; and the tenderer's bidding history.

In practice, it is not possible to study all markets of potential concern. As such, behavioural screens are more likely to be useful when investigating a complaint by customers or by rival firms. If the behaviour is deemed suspicious from the screen, then a more detailed investigation may be warranted.

Case-study: the CMA's bid-rigging detection tool

In 2017, the CMA launched a digital tool for detecting bid-rigging in procurement processes. The tool allows buyers to input data about tender procedures and bids. It will convert this information into a set of indicators of potential cartel conduct, with each indicator receiving a pass or fail mark depending on how it performs against the outcome that would be expected in a competitive market. A fail mark indicates a greater likelihood that the outcome of the tender was affected by bid rigging.

The total scores for each indicator are also weighted and added together to provide a total 'suspicion score.' The suspicion scores highlight which tenders are more likely than others to be suspect though, importantly, the CMA notes it does not prove the existence of a cartel. The screen is only the first stage of a multi-stage process. If the screen identifies suspicious behaviour, the market should be subjected to further analysis using traditional investigation tools to verify the results.

The CMA's tool uses 12 different indicators, as shown in the table below.

Table 5: Bid-rigging indicators used by the CMA

Theme	Indicator	Suggested weighting
Number and pattern of bids	Low number of bidders	20
	Single bid	30
Suspicious pricing patterns	Winning price is an outlier	20
	Similar pricing across bids	20
	Costs appear to be made up	40
Low endeavour submissions	Same authors in two or more bids	200
	Low endeavour losing bids	40
	Similar text in losing bids	200
Combination tests	Similar text and word count in losing bids	50
	Low number of bidders and make up prices	20
	Winning price is outlier and made up prices	10
	Low endeavour losing bids and made up prices	10

Source: CMA

What are the advantages and disadvantages of detection models?

Cartel screens are pro-active tools that can complement reactive detection methods, such as leniency programs. If applied properly, they may increase the rate of detected cartels and provide additional economic evidence to inform a cartel investigation or support the prosecution process.

On the other hand, cartel screenings are extremely costly processes. They may require costly investments in IT equipment and the employment of staff with expertise in programming and computer sciences. Applying a cartel screen also requires a significant amount of data on the market. Each

market to be analysed will require its own data gathering process using some combination of the techniques explained in **Section 3**. This can be costly and time consuming.

Finally, cartel screening is not an exact science. The screens apply a single model specification across many different types of industries and are unable to account for idiosyncratic factors which may affect the results. As such, cartel screens may give rise to false positives. For this reason, cartel screening should only be the first stage of a multi-stage process. If the screen identifies suspicious behaviour, the market should be subjected to further analysis using traditional investigation tools to verify the results.

5.2.3 Exercise 2: Detecting a cartel

Box 4: Detecting a cartel

The dataset provided can be used to detect a cartel. Consider that the agency has received a report that, in in June 2019 through August 2019, Store A discussed a possible price-fixing arrangement with Store B for product Q. Use the data provided to test whether any possible discussions between Store A and Store B had any influence on the price by estimating the model specified in equation (5.1) with appropriate dummy variable.

For the purpose of this exercise use the cleaned dataset. You will need to control for changes in the ABS fuel price index.

5.3 Defining markets

Competition authorities are concerned with the exercise of market power. To assess market power, it is often helpful to define the relevant market. The definition of the relevant market has two key functions: (1) it helps to specify the products and geographical area of concern, and (2) it allows the authority to identify market participants and measure market shares and market concentration.[25]

The analysis of potential market power need not start with a market definition. The analysis itself may inform the market definition. For example, if the analysis finds evidence that the introduction of a new competitor, or the closure of a competitor business, impacts significantly on the sales and/or prices of certain products, then one can deduce that those products are in the same market.

In this section we discuss several empirical techniques that are commonly used by competition authorities to assist in defining the relevant market in an investigation.

5.3.1 Natural experiments

In the context of competition analysis, a natural experiment is an historical event that potentially had an impact on prices and sales of products in a market. Examples of natural experiments are:

- the entry of a new competitor or product into a market, e.g., a new carsharing service
- the exit of a competitor or product from the market, e.g., the closure of a supermarket or cinema
- a shock to the supply line, for example, an industrial accident forcing a plant closure, a natural disaster, or an embargo
- previous mergers or de-mergers.

[25] Note that the concept of the relevant market for competition analysis may differ from the economic concept of the market.

By studying the impact of such natural experiments on prices, sales, margins, profits and other relevant variables, it is possible to gain valuable information on the nature of the market, including consumer responses to changes in prices.

An example where a natural experiment played a crucial role was in the consideration by the Directorate-General of the European Commission in 2005 of the proposed acquisition of Acetex, an active producer of acetyls and plastics, by Blackstone, a US private merchant bank. Blackstone also controlled a company, Celanese, active in the same product market. Hence, the proposed acquisition was of concern to the Commission.[26]

When examining this case, the Commission undertook an analysis of natural experiments likely to have provided a "shock" to the relevant market. The Commission focussed on unexpected plant outages of plants in Western Europe and Asia that produce vinyl acetate monomer (VAM). The Commission estimated the impact of outages on regional prices in the Western Europe, Asia and North America. It also estimated the impact of outages on imports from the US into Western Europe. The Commission found that the outages led to price increases in all three regions. It also found that the outages led to a significant increase in imports from the US It concluded that Western Europe is not a separate antitrust market for VAM.

The method used for these analyses was an econometric model with dummy variables similar to that discussed in **Section 5.2.1**. However, there are many different kinds of natural experiments that could be helpful in particular cases, and different quantitative methods may be useful in analysing the impact on markets of natural experiments in other cases.

The key strength of using a natural experiment is that it provides a shock to the market whose impact is likely to be much less confounded with other drivers of prices and demand than real market data for normal conditions. Even in cases where the impacts are not sustained, examining which products are affected by the shocks, and in which geographical regions provides valuable information useful in defining the market.

A key potential pitfall in the analysis of natural experiment analysis lies in an inappropriate choice of natural experiments or drawing incorrect conclusions from the analysis. For example, there are typically many intermediate products that go into making a final product. A disruption in the supply of any of these intermediate products could impact on prices of final products. However, the same intermediate product might be used to produce a wide range of products. Consider the case of a shortage in computer chips. Computer chips are used in a wide range of products, including computers and cars. A shortage of computer chips is likely to affect the supply and/or the prices of both computers and cars. A naive analysis of this natural experiment might therefore wrongly conclude that computers and cars are in the same market.

5.3.2 Price correlations

Examining price movements of products of interest across time and/or across regions is one of the most common empirical methods used in defining a market for competition purposes. Graphical analysis of price movements for a group of products can sometimes give quick insights into the relationships between the products. The strength of these relationships can be measured by price correlations. Price correlations are simple to calculate and do not require much data. The price correlation for a pair of products can be calculated either from the series of prices over time, or from prices across different geographical regions.

Correlation analysis is based on the proposition that the prices of goods that are substitutes for each other should move together. If the price of one good rises, demand will shift to its close substitutes and, thereby, cause the prices of the substitutes to increase. A table of pairwise price correlations for a group of products provides an indication of which goods may be substitutes for each other and which are not, and hence which products in the group belong to the same market.

[26] Durand, B. and V. Rabasse (2005), The role of quantitative analysis to delineate antitrust markets: An example, Blackstone / Acetex*, European Commission, Competition Policy Newsletter*, Autumn 2005.

In a market for differentiated products, the prices do not always move closely together. In such cases it is also helpful to undertake graphical analysis of the quantities of sales for the different products. If there is a price increase for one of the products, one would expect a decrease in sales for that product and an increase in the sales of substitute products. This can provide additional evidence on whether or not products are substitutes.

Calculating price correlations

The steps involved in calculating price correlations are as follows:

- obtain price data for the different products over time and aggregated across stores.[27] The prices can be obtained by first calculating the total quantities sold across all stores and the corresponding revenue for each product in each week. The prices are then calculated by dividing the revenue for each product in each week by the corresponding quantity
- arrange the data in wide format, as discussed in **Section 4.2.2**. In **Table 6** we show the price data for the 3 products (Q, R and S) for the first 10 weeks in our hypothetical dataset.
- calculate the price correlations between each pair of products. The correlations can be calculated in Excel, in R or in another econometric or statistical software.

The price correlations for the data in **Table 6** are shown in **Table 7**. This table shows that there is strong positive correlation between the prices for products R and S, but that the other correlations are negative. This indicates that the prices for products R and S move up and down together and suggests that they may be in the same market. In practice, one would want to corroborate this finding with other information to confirm that these two products do belong to the same market.

Table 6: Average prices for first 10 weeks of hypothetical example data

Week ending	Price product Q ($)	Price product R ($)	Price product S (S)
7/01/2019	1.87	1.41	1.65
14/01/2019	1.35	1.50	1.70
21/01/2019	2.07	1.42	1.53
28/01/2019	0.98	1.85	2.04
4/02/2019	1.32	1.61	1.73
11/02/2019	1.25	1.79	1.78
18/02/2019	1.77	1.62	1.54
25/02/2019	1.83	1.45	1.65
4/03/2019	1.35	1.96	2.10
11/03/2019	1.73	1.49	1.59

Source: Frontier Economics

[27] Instead of obtaining prices across time, one can instead obtain prices across different regions or across different stores. The data in **Error! Reference source not found.** would then have a row for each region or store rather for each time period.

Table 7: Price correlations for data in **Table 6**

Pairs of products	Price correlation
Q and R	-0.77
Q and S	-0.78
R and S	0.87

Source: Frontier Economics

Example of use of price correlations in merger investigation

An example of a price correlation analysis submitted to the European Commission in its enquiry into the proposed merger of Nestlé and Perrier in 1992 is provided in

Table 8. The table shows the price correlations for products in the still water (products A to C), sparkling water (product D to F) and soft drinks (products G to I) categories. There are strong positive price correlations between all the still water and sparkling water products, but weak or negative correlations between the water products and the soft drink products. This provides evidence that still water, and sparkling water are part of the same market, but that soft drinks do not belong to this market.

Table 8: Correlations between prices of brands of water and soft drinks

	A	B	C	D	E	F	G	H	I
A	1								
B	0.93	1							
C	0.91	0.94	1						
D	0.91	0.85	0.86	1					
E	0.94	0.97	0.95	0.92	1				
F	0.93	0.99	0.96	0.88	0.99	1			
G	0.11	0.05	-0.01	0.33	-0.02	0.01	1		
H	-0.57	-0.55	0.25	0.16	0.24	0.27	0.17	1	
I	-0.77	-0.75	-0.81	-0.86	-0.86	-0.79	0.33	-0.11	1

Source: Davis and Garces (2009) Quantitative Techniques for Competition Analysis, Table 4.3. Note: Correlations between prices of brands of still water (A–C), sparkling water (D–F), and soft drinks (G–I)

Potential pitfalls

There are a number of pitfalls in the use of correlations for identifying products that are substitutes, and hence may be considered as belonging to the same market. There are factors that can produce positive correlations between prices regardless of whether or not the products are substitutes. The most obvious of these factors is inflation. Prices for many products increase gradually over time due to inflation. This is an external common factor that induces positive correlation between the prices. To avoid this pitfall, it is important to calculate real prices using a price inflator such as the consumer price index before calculating the price correlations.

This is an example of “spurious correlation”. Spurious correlation between two variables arises when each of the two variables has a trend. For example, both the ownership of smart phones in the world and the number of people who die from falling down a set of stairs both have a positive trend and hence have a positive correlation. But it is very unlikely that there is a causal relationship between two phenomena.

There is an extensive literature in econometrics on the topic of spurious correlation and sophisticated statistical models have been developed for analysing data with trends. A simple way to reduce the impact of trends in the data is to analyse price changes from one period to the next rather than the prices themselves. Even if they both have a trend, two unrelated price series are unlikely to have changes in their prices that are similar in terms of their timing, direction and the size of the price changes.

A similar issue arises if products use common inputs in their production – e.g., electricity, oil, steel etc. Even after correcting all prices for inflation, changes in the prices for common inputs are likely to flow through to the prices for the final products and induce some co-movement in the prices for the final products and positive price correlations, even if the products are in totally different product markets. To correct for this problem requires more sophisticated modelling that enables a comparison of prices movements to be made after correcting for the influence of changes in the input prices.

It is worth noting that issues relating to time trends can be avoided if price correlations are calculated from prices across different regions rather than over time.

5.3.3 Demand estimation

A key consideration for competition authorities is how buyers will respond if a business increases the price for one of its products. If there are close substitutes for that product, then it is likely that some consumers would switch to the substitute products. Hence, the substitute products act as a constraint on the ability of the business to raise the price of its product.

To obtain estimates of how consumers respond to a change in price involves estimation of the price elasticities for both the business’ own product and the potential substitute products (cross-price elasticities). Price elasticities measure the impact of a one per cent change in the price of a product on the quantity sold of that product (own-price elasticity) or the quantities sold of other products (cross-price elasticities). One would expect the own-price elasticity to be negative since an increase in the price of a product is likely to result in a reduction in sales. With respect to cross-price elasticities, one would expect these to be positive for substitute products since an increase in the price of one product is likely to make people switch to a substitute product.

Estimates of price elasticities are used in other measures used by competition authorities, such as the SSNIP test and diversion ratios to assist with defining a market. Hence, price elasticities not only provide direct evidence of substitution between products but are also useful as inputs into these other measures.

A simple way to estimate the price elasticity of demand is to calculate the percentage change in the quantity sold of a product due to the price change, divided by the percentage change in the price. For example, if a business increases the price of a product by 10%, and it is estimated that this resulted in an increase in sales for a substitute product of 5%, then the cross-price elasticity is 5%/10% = 0.5.

Estimating price elasticities using an econometric model for demand

In practice there are usually other factors, in addition to prices, that may have had an impact on sales. To account for these other factors, estimates of price elasticities are usually obtained by estimating econometric models of demand.

A commonly used econometric model for estimating prices elasticities is the so-called double-log or linear-in-logs model:

(5.2) $$log(Y_t) = \alpha + \sum_i \beta_i log(P_{it}) + \sum_k \gamma_k log(X_{kt}) + \varepsilon_t$$

where:

- Y_t = the demand for the product under consideration in period t, which is the dependent variable in the regression

- P_{it} = the price of product i in period t, which includes both the price of the product itself, and the prices of potential substitute products
- X_{kt} = a set of other explanatory variables in period t , such as GDP, household disposable income and other variables that could influence the demand for the product. It also possible to add dummy variables to the model for special features such as a change in the number of competitor businesses.
- ε_t = the residual term in the regression.

The parameters β_i are the price elasticities with one of them being the own-price elasticity. The parameters γ_k are the elasticities of demand with respect to the economic variables and other drivers of demand in the model. The parameter α is the intercept in the equation.

Estimation of the double-log model can easily be done in any statistical or econometric software package, such as Stata or R.

The steps involved in estimating the double-log demand equation are:

- obtain quantity and price data for the different products over time[28]
- arrange the data in wide format, as discussed in Section 4.2.2.
- add columns for the values of any other drivers of the quantities sold, such as household income or other economic drivers
- calculate the logarithms of all the data in the table
- estimate the regression model
- interpret the results.

Table 9 show the quantity and price data for the 3 products (Q, R and S) in our hypothetical dataset shown in wide format suitable for estimating demand functions. The prices are the same way as in **Table 6**. The quantities were obtained by adding, for each product, the units sold across all stores.

Excel or a statistical package like R can be used to calculate the logarithms of each of the variables and estimate the regression model shown in equation (5.2). There are 3 separate demand equations that can be estimated: one equation for the demand for product Q, a second equation for the demand of product R and a third equation for the demand for product S.

For each these equations, the dependent variable is the logarithm of the quantity for the relevant product. For all three equations, the independent variables on the right-hand side of the equation are the logarithms of the 3 price variables.

[28] Instead of obtaining prices across time, one can instead obtain prices across different regions or across different stores. The data in **Error! Reference source not found.** would then have a row for each region or store rather for each time period.

Table 9: Quantity and price data for first 10 weeks of hypothetical example data

Week ending	Quantity product Q	Quantity product R	Quantity product S	Price product Q ($)	Price product R ($)	Price product S (S)
7/01/2019	4,074	3,350	1,800	1.87	1.41	1.65
14/01/2019	5,601	2,961	1,867	1.35	1.50	1.70
21/01/2019	3,674	2,959	2,255	2.07	1.42	1.53
28/01/2019	7,724	2,353	1,630	0.98	1.85	2.04
4/02/2019	5,722	2,619	1,985	1.32	1.61	1.73
11/02/2019	6,041	2,103	2,198	1.25	1.79	1.78
18/02/2019	4,289	2,203	2,642	1.77	1.62	1.54
25/02/2019	4,123	3,113	1,919	1.83	1.45	1.65
4/03/2019	5,628	2,152	1,656	1.35	1.96	2.10
11/03/2019	4,393	2,795	2,215	1.73	1.49	1.59

Source: Frontier Economics

Note: We have included more weeks of data in this example, because the regression model estimates a number of parameters, and this requires more data than when calculating correlations.

Interpreting the results

In **Table 10** we present the estimated equation for the demand for product Q estimated using the 10 weeks of data shown in **Table 9**. To interpret the results, please refer to the discussion in Section 5.2.1 on pages 32 to 34. The F-statistic for this equation is 14,616 with a p-value of 0.000. This indicates that the estimated model is statistically highly significant and that the probability that these results could have been produced by chance is practically zero.

The R^2 statistic for this equation is 0.9999 and the adjusted R^2 is 0.9998. The measures of how well the model fits the data indicate an almost perfect fit. This is largely due to the fact that there are very few observations compared to the number of coefficients that are estimated. For larger samples, it is unlikely that the fit will be so good.

The results in **Table 10** show that the estimate of the own-price elasticity (the coefficient for log(price Q)) is equal to -0.985. This indicates that a 1% increase in the price of product Q would lead to a 0.985% decrease in the quantity of product Q sold. The standard error of the estimate is very small compared to the estimate, and the t-value is very large and has a p-value of 0. This means that the estimated own-price elasticity is estimated quite precisely and is statistically highly significant.

The coefficients for log(price R) and log(Price S) are the estimates for the cross-price elasticities, i.e. the impact on the sales of product Q if there is change in either the price of product R or the price of product S. Both of these coefficients are positive, which indicates that an increase in the price of either of these products leads to an increase in sales for product Q, indicating that these products are substitutes for product Q. However, the coefficients are very small and have large p-values – much higher than any commonly used level of significance. Hence, we cannot reject the hypothesis the prices of products R and S have no impact on the sales of product Q.

Table 10: Demand equation for product Q

	Beta	Standard error	t-value	p-value
log(price Q)	-0.985	0.008	-117.43	0
log(price R)	0.006	0.019	0.29	0.778
log(price S)	0.015	0.022	0.7	0.512
constant	8.916	0.013	707.63	0

Source: Frontier Economics

When interpreting cross-price elasticities it is important also to take into account the own-price elasticity of the product raising its price and the relative sales of the products. Assume the product raising its price by 10% had sales of 100,000 units before the price change and an own-price elasticity of -0.4. Raising the price by 10% would then lead to a reduction in sales of 0.4 x 10% = 4% or 4,000 units. Assume that 1,000 of these units are not replaced by a substitute product, but that the other 3,000 lost sales are replaced by a substitute product. Assume that prior to the price change the substitute product had sales of 10,000 units; then the increase of 3,000 in sales would amount to a 30% increase in sales, giving a cross-price elasticity of 30%/10% = 3. However, if the substitute product had sales of 150,000 units prior to the price change, then the increase in sales of 3,000 units would only amount to a 2% increase giving a cross-price elasticity of 2%/10% = 0.2.

This example illustrates that while the signs of cross-price elasticities are informative about the substitutability of products, the sizes of cross-price elasticities need to be interpreted with considerable care. The key information when interpreting cross-price elasticities is the sign of the estimated elasticities and the statistical significance. A cross-price elasticities that is positive and statistically significantly different from 0 provides evidence that the product with the positive cross-price elasticity and product on the left-hand side of the equation are substitutes and are part of the same market.

Estimating demand systems

One disadvantage of the double-log model is that restrictions on the elasticities that derive from economic theory, the Slutsky restrictions, cannot be imposed or tested in this model. The inability to impose theoretical restrictions sometimes also leads to implausible estimates for the elasticities when using small datasets. The most commonly used model for the estimating price elasticities that overcomes this limitation is the Almost Ideal Demand System (AIDS). The econometric specification and estimation of this model are considerably more complex than the double-log demand model. We present further details of this model in Appendix A.

Limitations of the econometric estimation of price elasticities

Robust econometric estimation of elasticities using either the double-log model or the AIDS model involves checking the models for potential violations of a range of assumptions that underpin any econometric estimation task. This includes the possibility that the estimates are spurious due to trends in the variables. To avoid such potential pitfalls, it is important that these estimations are carried out or validated by experienced econometricians.

5.3.4 Use of GIS mapping software

A key consideration for competition authorities is often the geographical extent of the market. While some markets may be trivially economy-wide in scope, others, such as restaurants in remote villages, may have strong geographical elements with narrow geographic markets. For others, the extent of geographic markets may be best described as an empirical question. This section will provide a guide on the use of geographical information system (GIS) software in the context of market definition, in particular the geographical extent of a market participant's influence on other market participants.

Approach

The use of mapping software (GIS) involves several facets:

- The representation of locations and areas in a spatial environment
- Performing calculations using the spatial objects
- Presenting results spatially

As a first step, it is necessary to obtain a geographic representation of the data. This may be locations of market participants (latitude and longitude, or perhaps postcode/suburb information), or the geographical description of areas such as postcodes or suburbs. It is then possible to construct various measures, for example:

- Distance, drive time or drive distance between market participants
- Spatial markets for each market participant, which may be defined as the area within some specified radius, or drive time, or suburb in which the participant is located, or alternatively if data permits, the area that contains some proportion of a retailer's customers
- Characteristics of derived spatial markets, by using available data at the postcode level
- Identities of all retailers in a derived spatial market, along with distance to the focal participant

The tools available in GIS are powerful, compatible with commonly used software platforms and can be supplemented by plugins and APIs for third party providers.

What are the key strengths?

GIS is a powerful tool that can help competition authorities understand local markets in situations where markets are likely to possess strong local elements, for example supermarkets or petrol retailers.

There are many calculations that may be performed using GIS that are otherwise impractical, for example calculating the share of a suburb within 5 kilometres of a retail store location. GIS can allow the analyst to merge datasets with data on geographic areas, for example census data at the suburb level, to characterize local markets.

GIS can allow the competition authority as a first step to determine where there may be competition concerns, based on local market concentration or other measures, and then further examine the local markets by overlaying maps sourced from providers such as Google with key features such as the location of competitors. In serving as a filter, allowing the competition authority to focus only on potentially problematic areas, GIS may be able to reduce the time taken in competition analysis

What are the key limitations?

A key limitation is the availability of geographic data. The competition authority would need to source data on competitor locations. While this can be obtained from participants, some data cleaning would typically be required due to incorrect data. Further, data on the characteristics of geographic areas may be highly aggregated, available at the state level rather than the suburb level for example. The data describing, geographically, these areas may also not be available.

Another key limitation is the applicability of geography to the market considered. If it is established that a market is economy-wide, there would be little benefit from carrying out spatial analysis using GIS as location would have a minimal impact on competition. Market feedback may however inform the authority of this in the initial stages.

5.3.5 Exercise 3: Defining geographic bounds of markets

Box 5: Defining geographic bounds of markets

Please use your cleaned customer location dataset to derive the appropriate geographic bounds of the markets. In this particular case, the radius of the geographic market for each store and product is the same across all stores and products. For this reason, it will not be necessary to estimate a separate radius each store or product. We suggest you estimate the radius on the assumption that the market should include 80 per cent of the relevant customers. Hint: The Haversine formula may be used to find the distance between a pair of geographic coordinates, where a coordinate is described by a latitude and a longitude.

5.3.6 Exercise 4: Calculating price correlations to determine product markets

Box 6: Calculating price correlations to determine product markets

Please use your cleaned retail sales dataset to calculate price correlations. Aggregate the data over all stores to produce a series for the prices for each product and each week.

You can use Excel or a statistical software package to calculate the pairwise correlations between the prices of the three products, Q, R and S. What conclusions can you draw about which products may be substitutes?

5.3.7 Exercise 5: Estimating demand elasticities to determine product markets

Box 7: Estimating demand elasticities to determine product markets

Please use your cleaned retail sales dataset to estimate demand elasticities. Aggregate the data over all stores to produce a series for the prices and quantities for each product and each week, then convert to a wide format.

(1) You can use Excel or a statistical software package to estimate the single-equation double-log relationships described by equation (5.2) between the log of the quantities sold for each product and the logs of prices.

(2) To allow for changes in common cost drivers for the products, repeat the exercise in (1), but add the logarithm of the automotive fuel price index as an additional explanatory variable.

What conclusions can you draw about which products may be substitutes? Be sure to check that your results are consistent with your knowledge from other sources that products R and S are substitutes.

(3) If you are comfortable using R, you can also use the micEconAids package to estimate Marshallian elasticities by estimating the AIDS demand system described in Appendix A.

5.3.8 Exercise 6: Estimating a diversion ratio between stores

Box 8: Estimating a diversion ratio between stores

Please your cleaned retail sales dataset to estimate a diversion ratio from Store A to Store B for product Q. During the period June 2020 onwards, Store A did not offer product Q. What proportion of the sales of Q lost by Store A went in additional sales to Store B and to Store C?

5.3.9 Exercise 7: Estimating a diversion ratio between products within a market

Box 9: Estimating a diversion ratio between products within a market

This exercise is based on the assumption that Store A and Store B are in the same market. During the period June 2020 onwards, Store A did not offer product Q. What was the impact on sales of R and S as a result of this event? What can we conclude from this result?

5.4 Assessing market power

Market power is the ability of a company to act independently of competitive constraints. These competitive constraints might come from other incumbents in the market or from potential entrants to the market.

When thinking of prices, market power might be defined as the ability of a company to maintain prices above the competitive level for a sustained period. When thinking of profits, market power might be defined as the ability of a company to earn a return on funds invested above the cost of those funds for a sustained period.

5.4.1 Natural experiments

When assessing market power, no single measure will do: one usually has to use a range of measures. In extreme cases, one can sometimes observe a natural experiment – the corporation may have behaved in way that can only be explained by its substantial market power. For example, a firm may have been the only firm in a market for 20 years and be highly inefficient in its operations. In such a case, one can readily infer substantial market power, because if the firm had been subject to significant competitive pressure, such a highly inefficient firm could not have survived for so long.

5.4.2 Analysis of profits

In the absence of such a clear natural experiment, one way of assessing the market power of a corporation is to compare its rate of return on funds invested with the cost of those funds. However, care needs to be exercise because the profits recorded in a firm's financial accounts are not recorded to reflect economic profits.[29] Rather, they are recorded according to rules that are designed to ensure that shareholders are given a fair picture of the financial state of the firm. A key difference between economic and financial rates of return arises from rules concerning depreciation and the valuation of assets.

One way to avoid these problems is to estimate the company's internal rate of return (IRR) from its cash flow records. This requires many years of data; and it still requires opening and closing values of

[29] See Franklin M Fisher and John J McGowan, "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, Vol 73 (1983) pp 82-97.

the assets of the company. However, if many years of data are available, a truncated IRR can be calculated as an alternative to the company's return on capital employed.

5.4.3 The influence of concentration

All competition agencies use measures of market concentration as a rough indication of market power. This is particularly appropriate as an indicator of unilateral effects of competition in the case of homogeneous products because the Cournot model indicates a relationship between market concentration (as measured by the Hirschman-Herfindahl index) and the margin of price on marginal cost – where the latter is often used as an indicator of market power. The Hirschman-Herfindahl Index (HHI) is written as:

$$\text{HHI} = \sum_i s_i^2$$

where s_i is the market share of the ith firm in the market and the HHI is the sum of these market shares after they have been squared.

If the issue of market power concerns a differentiated product market, measures of market concentration, such as the HHI, may tell little about the market power of the firm. The more-relevant question is whether the firm in question is constrained by close competitors.

5.4.4 Stability of market shares

Although static measures of market shares tell us little about the market power of a firm in a differentiated product market, the stability of market shares over years (for example, a decade) may tell us much more. As John Sutton pointed out in his classic study, *Sunk Cost and Market Structure*,[30] differentiated-product markets may display market shares that change little over time because of endogenous sunk costs that create barriers to entry and to expansion. If a differentiated product market displays market shares that are reasonably constant over time, one should check the Sutton hypothesis by seeking independent evidence concerning the sunk costs associated with developing, launching and advertising products in the relevant market.

Box 10: Case study: The NZCC's market study into the grocery sector

As part of its market study in the grocery sector, the New Zealand Commerce Commission commissioned Frontier Economics to undertake some econometric analysis to examine how the structure of local grocery markets affects prices and margins in New Zealand. The analysis consisted of two studies, each of which focused on a particular issue.

The first study tested whether there is a relationship between concentration and prices in local grocery markets. The study compared the way in which variations in concentration across local grocery markets were related to variations in prices. It controlled for various other factors that may influence variations in prices across markets.

A separate market was defined around each of the 694 stores in the sample and two different measures of concentration were used. The first measure was the share of revenue of the group to which the focal store belongs, when these shares were weighted to account for the distance of competitor stores to the focal store. The second measure was the proximity of the closest store of each of the six banner groups in the study.

The results of the tests varied somewhat depending on the measure of market concentration. In models with weighted revenue share as the concentration variable, increased concentration in local markets tended to be associated with higher prices for certain banner groups. In models with proximity to major banners as the concentration measure, we found that some banner groups tended to have lower prices if certain other banners were nearby.

[30] MIT Press, 1991.

The second study analysed the impact of the entry, exit or rebranded entry on the prices charged by close competitor stores in the 12 months following the event. The study used a sample of 32 events of entry, exit or rebranding These events had a good geographic spread with different demographic characteristics, such as population density. The study found that most of these events had no economically-significant impact – where an economically-significant impact was defined as at least 0.5% change (increase or decrease). Of the 9 cases with an economically significant impact, two were positive (such as entry associated with a price increase) and seven were negative. The absolute magnitude of the impact of every event on the prices of the close competitor stores was less than 2.25%. The study concluded that there is no systematic relationship between the entry or exit of a store and the prices charged by close competitor stores in the 12 months after entry.

Source: NZCC

What are the key strengths?

Many methods are available to assess a corporation's degree of market power. The methods employed will depend on what data are available. However, it is not advisable to rely on only one method.

What are the key limitations?

Assessing the closeness of competitors in a differentiated product market can be undertaken using econometric methods to estimate demand elasticities. However, there are many reasons why such estimation might produce invalid results. For this reason, it should only be undertaken by staff with specialist training and experience in econometrics.

5.4.5 Exercise 8: Determining market shares

Box 11: Determining market shares

Assume that BP and Caltex are proposing to merge. What will be the effect of this merger on the shares in the market centred on Store F? Hint: this may be done one of two ways:

- you can use distance-weighted revenue to calculate market shares; or
- you can use distance weights assuming all firms have the same revenue

5.5 Predicting outcomes of horizontal mergers

Most competition agencies assess mergers according to whether they are likely to have the effect of substantially lessening competition. The application of this test involves a comparison of the future state of competition in the relevant market with and without the merger.

Because we do not know for certain what the future will bring, the assessment of the likely effects of a merger involves a comparison of two unknowns:

- the future state of competition in the market with the merger
- the future state of competition in the market without the merger.

The structure that guides the comparison of these two states necessarily involves thinking in terms of models. This is true whether the model is implicit or explicit; and it is a good discipline on the staff of the agency to be explicit about the model they have in mind. This does not mean that staff have to use formal modelling to estimate the likely outcome of the merger in terms of units of currency. Rather, it means that staff should try to be explicit about the kind of model they have in mind when thinking about the particular merger before them.

As we suggested when discussing the assessment of market power, the key decision an agency must make when considering a horizontal merger is whether the merger involves firms producing homogeneous products (such as oil) or whether the merger is between companies producing differentiated products (such as carbonated soft drinks).

5.5.1 Modelling mergers in electricity markets

In consideration of modelling mergers in electricity markets, it is important to recognise that regulators and competition authorities can utilise market modelling to project different market outcomes in different settings or scenarios e.g., to assess the long-term impacts of changes in the electricity market. For example, market modelling can be used to assess the impact of adding new generators and interconnectors to the energy system each year, or the introduction of electricity policies (e.g., renewable energy targets and carbon emissions constraints) which would impact market participants. The model would be able to determine the efficient (least-cost) operation and investment in an electricity market over a long-term investment horizon and regulators would be able to determine the optimal size, type, location, and timing of new generation capacity and/or network changes to meet demand over the analysis horizon.

The standard model economists use when considering a merger between firms producing homogeneous products is the Cournot model. As we indicated in **Section 5.4.3** above, the HHI is critical to the outcome of this model. Another factor that is critical is the shape of the marginal cost functions of the producers.

In the Cournot model, firms compete by simultaneously choosing a quantity that maximises their profit, having regard to quantities offered by competitors. In a market with two identical firms, if both produce high quantities, market prices will be low, and each firm would be better off by producing lower quantities and achieving higher prices. If both firms produce low quantities, market prices will be high, and each firm would be better off capturing additional volume at lower prices by producing higher quantities. If one firm produces low, and the other high, the high producing firm will benefit from the higher prices caused by the low producing firm's reduced output. In response, the low producing firm should increase its level of output. A Cournot equilibrium exists where each firm respond optimally to the production incentives of the other firm.

Cournot equilibria can be used to predict outcomes in states of the world with and without a merger on a forward-looking basis. A case study relating to the Australian Electricity Market is provided in Box 12.

Box 12: Case study: Electricity generation merger in Australia

AGL, a major vertically-integrated electricity generator-retailer in Australia, announced plans to acquire two large black-coal generators from the New South Wales government in 2014. The Australian Competition and Consumer Commission opposed the acquisition because it asserted the transaction would likely result in a substantial lessening of competition in the market for retail electricity and for wholesale electricity contracts.

AGL challenged the ACCC's decision by applying for authorisation from the Australian Competition Tribunal. Under the application, AGL had to show that the transaction would lead to public benefits despite any competition concerns, i.e., the Tribunal would need to be satisfied there were net public benefits from the transaction proceeding.

One aspect of AGL's case provided by its experts was simulation modelling based around Cournot competition. Market outcomes were modelled under a business-as-usual case and a merger case, with Cournot equilibria the tool used to predict outcomes in each scenario. The Tribunal found this modelling approach to be compelling and provided authorisation in late 2014.

Source: https://www.judgments.fedcourt.gov.au/judgments/Judgments/tribunals/acompt/2014/2014acompt0001

5.5.2 Estimating upward pricing pressure

As we discussed in 5.4.4, the key influence on market power of a firm in a differentiated product market is the closeness of its competitors. When considering a merger, one of the key factors that needs to be considered is how close the two firms are in product space – that is, whether the two firms are producing goods that are very close substitutes or less-close substitutes.

Consider two firms (1,2) that are proposing to merge. Prior to the merger, each is setting its prices to maximise its profits. Setting profit-maximising prices involves considering the trade-off between price and quantity: the higher the price, the less that will be sold. A profit-maximising firm will set the price at the profit-maximising level, taking into account that, if it pushed the price any higher, it would lose some of its sales.

A key issue when considering a proposed merger between firms 1 and 2 is, if 1 pushed its price a bit higher, what proportion of its lost sales would be lost to 2; and if 2 pushed its price a bit higher, what proportion of its lost sales would be lost to 1? Proportions of this kind are known as diversion ratios. Suppose when firm 1 last increased its prices, it estimated that 50% of its lost sales were lost to firm 2. If the merger proceeds, then the sales that would previously have been lost to firm 2, would now not be counted as lost – because firms 1 and 2 would have the same owner. Therefore, a merger of parties with high diversion ratios (a high degree of substitution between products produced by 1 and 2) is likely to cause prices to increase.

Estimates of diversion ratios can be obtained in various ways:

- they may be known to the parties as a result of asking leaving customers where they are going – or asking new customers who they used to buy from
- they can sometimes be estimated from a natural experiment – if a cinema is temporarily closed for renovation, by how much did attendance at the other cinema increase?
- they can be estimated from consumer surveys – by asking consumers who they buy from and from whom they used to buy
- they can be calculated direct from estimates of cross-price elasticities – which show the percentage increase in the quantity sold of one product in response to a one per cent increase in price of the other product.

Some competition agencies make use of diversion ratios to estimate the likely effect on prices of a merger in a differentiated product market. The Upward Pricing Pressure (UPP) on prices of firm 1 is estimated in its simplest form as the diversion ratio from 1 to 2 (the proportion of the sales lost by 1 that are captured by 2) multiplied by the margin of price on marginal cost of firm 2:

$UPP1 = D12\ (P2 — MC2)$[31]

For example, consider a margin of 25% and a diversion ratio of 20%, the UPP would be 5%. The UPP formula would predict that the prices of firm 1 would increase by 5% as a result of the merger. The margin of firm 2 is relevant because it shows the increase in firm 2's profit that would be lost for each unit lost to firm 2 if firm 1 increased its price and the merger were not to proceed.

Sometimes UPP is referred to as GUPP, where the G stands for gross. This acknowledges that the tendency of horizontal mergers to cause an increase in prices may be offset by decreases in marginal costs.

[31] See Joseph Farrell and Carl Shapiro, "Antitrust Evaluation of Horizontal Mergers: An Economic Alternative to Market Definition", *The B. E. Journal of Theoretical Economics*, Vol 10 (2010) pp 1-39.

5.5.3 Differentiated Bertrand simulation

While the standard Bertrand and Cournot Models are useful for analysing homogeneous goods, they also work well in considering competition between firms producing differentiated product. It is similar to the logic of UPP, in that it is based on elasticities of demand and margins of prices on marginal costs.[32]

Although it is common for competition agencies to look for evidence of diversion ratios, it is relatively rare for agencies to undertake full simulations of mergers using the differentiated Bertrand model. The technique of such simulations is relatively simple to describe. A profit function is created for each firm, when profits are defined as the price-cost margin for the firm multiplied by quantity sold by the firm - where quantity sold by a firm is written as a function of its own price and the prices of its competitors. The coefficients on the prices in the quantity function have been previously estimated with a demand system as discussed in **Section 5.3.3** above. Each firm is assumed to choose its own price to maximize its profits. The expression for the profit-maximizing price for each firm will be its reaction function – because it will be a function of the prices of its competitors. The equilibrium in the market exists when each of the reaction functions holds simultaneously. This can be solved in Excel.

An early instance of full simulation using a differentiated Bertrand model was the challenge by the United States Federal Trade Commission (FTC) of the proposed merger of two wholesalers of bread in Chicago and Los Angeles. The FTC used a logit demand system to estimate the price elasticities of demand which were inputs into a Bertrand simulation model.[33] The merging parties criticized the inputs that were used in the modelling – both the demand elasticities and the price-cost margins. The Commission decided not to rely on the results of the modelling because of the novelty of the approach (using differentiated Bertrand simulation) and because of the level of disagreement.

What are the key strengths?

Formal models are valuable for checking the reasoning of staff. For this reason, staff should be encouraged to state the economic model on which they are relying. This will also promote transparency for the businesses that are subject to regulation. They have a right to know the reasoning of the agency.

What are the key limitations?

All theory is based on assumptions. These assumptions can be simple (such as the importance of diversion ratios) or more complex (such as when simulation modelling based on estimates of elasticities and margins of price on marginal costs). The more complex the assumptions, the more controversial will be the results of the modelling. Agencies that wish to promote sound reasoning and transparency will try to make explicit the economic models on which they rely; but the use of full merger simulation models will create grounds for large-scale disputes.

5.6 Determining appropriate penalties

There is a substantial economics literature on the economics of penalties.[34] The literature is concerned principally with the function of penalties in deterring similar conduct in the future. The principal lesson from the literature is that the penalty should be a multiple of the costs the conduct has imposed on others. The penalty has to be a multiple of the costs because the probability of detection is less than one. For example, if the probability of detection is 25%, the penalty needs to be four times the cost imposed on others so that a company contemplating an infringement will consider an expected cost of the conduct equal to the cost the conduct will impose on others.

---

32 A good non-technical introduction and a numerical example is to be found in David Besanko,David Dranove and Mark Shanley, Economics of Strategy,2nd edition pp. 254-258.

33 The procedure is explained in G Werden, “Expert Report in United states v Interstate Bakeries Corp and Continental Baking Co”, *International Journal of the Economics of Business*, Vol 7 (2000) pp 139-148.

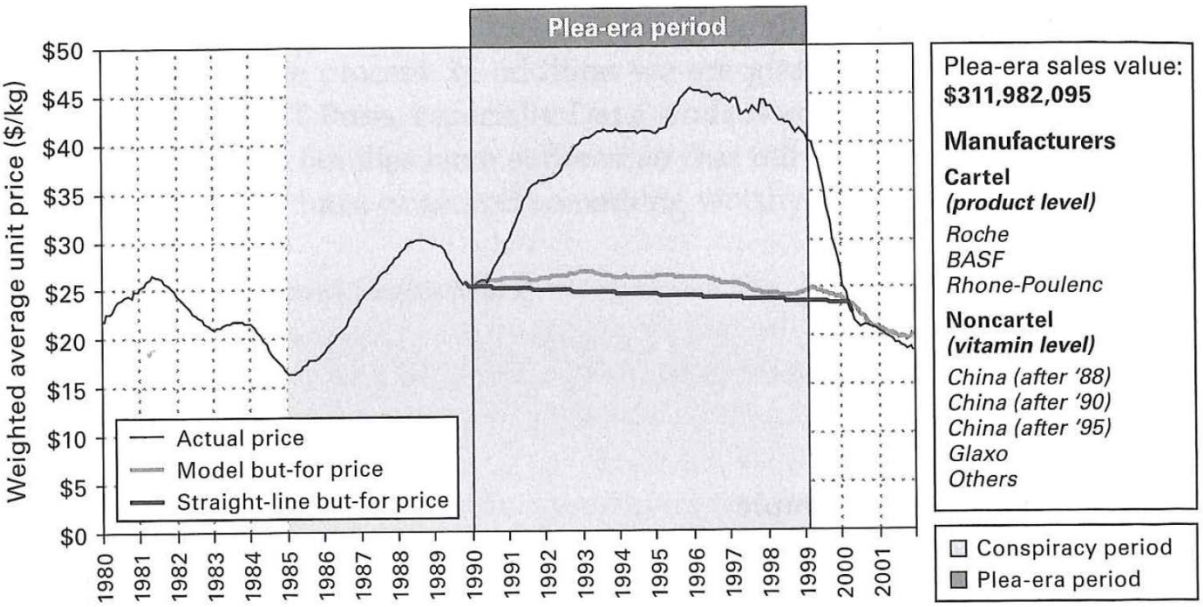
34 A seminal paper is Gary S Becker, “Crime and Punishment: An Economic Approach”. Journal of Political Economy, Vol 76 (1968).

This reasoning is applicable to anti-competitive conduct such as the formation of a cartel. Cartels which increase prices create benefits to the members of the cartel and impose costs on the rest of the community. The costs they impose on the rest of the community are larger in magnitude than the benefits to the members. The excess of benefits to the members of the cartel over the costs imposed on others is referred to by economists as the 'deadweight loss' caused by the cartel. If businesses contemplating the formation of a cartel expect a penalty equal to the cost they will be imposing on others, they will be unlikely to form the cartel.

5.6.1 Estimating damages

The economic literature on penalties suggests one needs to have some idea of the cost the conduct has imposed on society before one can fix the penalty. In some cases, this is relatively simple to do. The increase in the price of vitamin A caused by the international vitamin's cartel is depicted in **Figure 5** below. As discussed in **Section 5.2.1**, the but-for price is an estimate of the prices that would have prevailed but for the cartel.

Figure 5: Price of vitamin A acetate 650 feed



Source: Expert Report of B Douglas Bernheim, reproduced in Robert C Marshall and Leslie M Marx, The Economics of Collusion (MIT Press) p 2.

In cases where the effect of the illegal conduct on prices is so dramatic, econometric models may not contribute much to the estimate of the costs the conduct imposes on society. The cost is the increase in prices multiplied by the volume of sales. To this, needs to be added an amount for the deadweight loss that the conduct causes by reducing demand for the product.

In other cases, an estimation of the cost imposed by the conduct is more difficult. Not all cartels are as successful for the participants as the international vitamin's cartel. Some cartels, such as the OPEC oil cartel, form and then collapse and re-form over time. When simple plots indicate little about the effect of the cartel on prices, more advanced techniques may be required to estimate the costs of illegal conduct such as the formation of a cartel.

A commonly used approach is to construct an econometric model to estimate what the price would have been during the cartel period if there had been no cartel, the "but-for" price. The dependent variable in this approach is the price of the product, and the explanatory variables are exogenous demand and supply factors that may influence the price. A dummy variable is used to separate the sample period into the period of the cartel and the periods before and/or after the cartel. before, during and after the cartel. The coefficient on the dummy provides an estimate of the average price increase that occurred during the cartel period. After estimation, the "but-for" price can be obtained by

predicting the prices for each year of the cartel period after setting the cartel dummy variable equal to 0.

Estimating a robust model for “but-for” price prediction involves many challenges. Often, the cartel applied to a range of products, not just a single product. The model needs to account for differences in the characteristics of these products. The product specifications may also change over time, confounding the impact of the cartel on prices with that impact of the changing specifications. Prices are often also subject to various discounts. Detailed knowledge of the products and industry practices is required to properly account for all the factors that affect the prices of the products. It also requires a large database that contains detailed information on all these factors. Such databases and usually messy and have many data problems, e.g., negative quantities (if some of the products are returned), implausibly large prices, missing values, etc. Extensive data handling skills and experience are required to turn such raw data into a dataset suitable for econometric analysis.

There are also some common problems that can arise in any econometric modelling, such as misspecification, endogeneity, multicollinearity, non-stationarity and measurement errors, the presence of which could invalidate the results. To overcome such problems, suitable diagnostic tests should be carried out on the model to provide evidence of the robustness of the results derived from the model.

6 Organisation structure

Competition authorities organise themselves in different ways. This will be influenced by a range of factors including, for example, any administrative rules in the legal instrument establishing the authority, the objectives of the authority and the functions that have been allocated to it, the size and competency of the authority, and the traditions and institutional structure of member economies.

There is no unique institutional design which would fit all economies, and any particular design involves trade-offs that must be assessed against the local conditions of an economy. Notwithstanding these differences, competition authorities need to ensure that they are effective in discharging their duties. In an era of increasing digitisation and reliance on big data, many competition authorities have recognised a need to engage people with experience in digital markets, data science techniques and empirical analysis.

6.1 Meeting the demands of big data

Competition authorities have adopted different approaches to adapt to the increasing demands of digitized markets and the growing prevalence of big data.

One approach has been to engage staff with explicit experience in data science, machine learning and empirical techniques. In Australia, the ACCC established the Strategic Data Analysis Unit to provide expert quantitative analytical support across the Commission. Members of the unit work on key inquiries and matters where the use of data and analysis can assist. The unit assists with framing analysis, determining data sources, performing analysis and communicating analytical results. Typically, the team supplements, rather than replaces, economists and econometricians on the case team.

In Singapore, as part of a whole-of-government effort to improve data driven decision making, CCCS staff members attend training programs on data analytics so that they may better utilise these techniques when assessing competition matters. The CCCS has also established a Data Monitoring and Analytics Unit. They support the case team by collecting and managing data but are not involved in the economic analysis of the data (which is principally undertaken by the economist on the case team).

Another approach has been to establish a team of people with a specific mandate to consider digital markets. This approach has been adopted by the UK Government through its proposed establishment of the Digital Markets Unit, a new regulatory arm designed to oversee plans to promote greater competition in digital and online markets. The UK Government has indicated that the statutory duty of the Digital Markets Unit should be to promote competition (which includes promoting competitive outcomes) in digital markets for the benefit of consumers. The unit currently exists in ‘shadow’ form under the CMA, with its final objectives and functions subject to statutory approval.

6.2 Establishing a data analytics team

While both approaches discussed above have their advantages, competition authorities seeking to enhance their data gathering and analytical techniques across digital and non-digital markets should adopt the former – that is, by establishing a separate data analytics team. The process involved in establishing this team will vary from one competition authority to another. The box below provides a general guide that authorities can follow to ensure the team is fit for purpose.

Box 13: General guide to establishing a separate data analytics team

- *Determine objective*: The objective of the data analytics team may be to provide expert quantitative analytical support across the authority.
- *Establish functions*: The internal offerings of the data analytics team may include:

- building software tools to tackle repetitive or slow tasks
- data gathering and manipulation at scale and pace
- providing analysis and insight using data science
- understanding and explaining technology and digital markets
- analysing algorithms, and how they are used

- *Confirm size:* The size of the data analytics team will depend on the authority's size, remit, sources of data and IT estate. The competition authority should start with a small team covering the capabilities above and expand organically and iteratively as utilisation of the team increases.
- *Develop governance framework*: It is important to ensure that robust processes and procedures are in place to guide collaboration and communication between the data analytics team and other teams within the competition authority. The governance framework should clearly set out how the resources of the data analytics unit will be utilised across the authority and establish lines of responsibility. There is no one way this can be done, and the framework for one authority will be heavily influenced by existing rules and practices.
- *Engage staff:* The authority should engage people for the data analytics team with suitable expertise and experience, including with respect to:
  - designing, building and maintaining datasets
  - developing sophisticated models using different programming languages
  - undertaking econometric analysis
  - algorithms, machine learning and artificial intelligence
- *Monitor over time*: It will be important to review and monitor the performance of the data analytics team over time. This can be achieved with regulatory post-mortems or feedback sessions. These reviews will allow the authority to identify any gaps in experience which need to be filled, or any changes to the governance framework that may be needed to clarify how resources are utilised.

6.3 Key strengths and limitations

The key advantages of establishing a separate data analytics team are that:

- it puts focus on hiring and training data scientists
- it facilitates the sharing of data scientists and resources across the agency
- it facilitates inter-agency sharing of data science learnings, tools and techniques
- it helps the agency keep up-to-date with developments in data tools and techniques

However, there are also a number of risks associated with establishing a separate data analytics team which, if not appropriately managed, may undermine its effectiveness.

A key challenge is ensuring that there is sufficient communication between the case team and the data analytics team. The case team's decision on a particular competition issue may be compromised if it does not have sufficient knowledge of how the analysis was undertaken, the key assumptions that were made, and/or the limitations of the results. Conversely, if the data analytics team does not have a proper understanding of the facts of the case and the proposed theories of harm, the empirical analysis undertaken by the team may become detached from the economic underpinnings of the case. In addition, the absence of sufficient communication may result in the possible duplication of work between the data analytics team and case team, leading to suboptimal usage of resources.

These risks can be addressed by ensuring that there are robust processes and procedures in place to guide collaboration and communication between the data analytics team and other teams within the competition authority. This includes, for example, establishing appropriate lines of responsibility, ensuring participation and representation from both the case team and the data analytics team at meetings, and conducting regular post-mortem and feedback sessions.

A The Almost Ideal Demand System

One disadvantage of the double-log model is that restrictions on the elasticities that derive from economic theory, the Slutsky restrictions, cannot be imposed or tested in this model. This inability to impose theoretical restrictions sometimes also leads to implausible estimates for the elasticities when using small datasets.

The most commonly used model for the estimating price elasticities that overcomes this limitation is the Almost Ideal Demand System (AIDS) model.[35] The AIDS model considers the demand for products in the group of interest as a system, with one equation for each product. The specification of the equation for each product j in the system of equations is:

(B.1) $$w_{jt} = \alpha_j + \sum_i \beta_{ji} log(P_{it}) + \gamma_j log\left(\frac{E_t}{S_t}\right) + \varepsilon_{jt}$$

where:

- w_{jt} = the share of expenditure at time t for product j within the group of products under consideration. This is the dependent variable for equation j in the model
- P_{it} = the price of product i at time t, which includes both the price of the product itself, and the prices of potential substitute products
- E_t = total expenditure at time t
- S_t = a specific price index at time t that combines the prices of all the products in the group. This is known as the Stone price index
- ε_{jt} = the residual term in the equation for product j.

It is also possible to add other variables to the model to take account of economic and other factors that may have an impact on the expenditure shares for different products.

The parameter β_{ji} is an estimate of the impact of a change in the price of product i, P_{it}, on the revenue share of product j. But it can't be interpreted directly as an elasticity. Additional calculations need to be done to obtain the price elasticities.[36] The estimation of the AIDS model, and the calculation of the elasticities, is considerably more complex than the double-log model and cannot be done with standard regression routines. However, both Stata and R have special routines for undertaking these calculations.

Whether one uses the double-log model or the AIDS model to estimate price elasticities may depend on ease of estimation and interpretation versus a desire to impose restrictions consistent with economic theory.

Estimation of elasticities using either of both the double-log model and the AIDS model may face potential problems often faced in other econometric estimation tasks, including the possibility that the estimates are spurious due to trends in the variables. To avoid such potential pitfalls, it is important that these estimations are carried out by experienced econometricians.

[35] The AIDS model was introduced in the literature before the spread of the AIDS disease. To avoid confusion between the econometric model and the disease, some economists refer to the model as the Nearly Ideal Demand System (NIDS).

[36] The elasticities for the AIDS model are conditional elasticities, since they are estimated on the assumption that total expenditure on the group of products under consideration is not affected by a change in the price of one of the products in the group. In practice, the impact of this assumption is usually quite small.

B Solutions to exercises

Exercise 1: Data-cleaning exercise

It should be feasible to compare the cleaned datasets to the clean versions provided; any discrepancies represent data errors that remain uncorrected.[37]

Retail data

In combining the four retailer data sets, the column headings should be edited to be consistent. For example, with BP "Name" should be changed to "Product", and with ColesExpress "AveragePrice" should be changed to "Price". It should also be noted that United uses the "WeekStarting" convention whereas the others use the "WeekEnding" convention for describing weekly retail data. For now, the date variable could be changed to something consistent, noting that the date for United will need to be edited to reflect the "WeekStarting" date later. Once names align, the data can be appended to form a combined dataset in the long format.

The summary command in R is quite useful at this point. Using this command will reveal that the date variable is in a character format, and that there are some observations with zero units sold, though the minimum revenue is positive.

Using the table command, we can observe that there are nine different stores, though there are fewer observations for store A (this is due to the withdrawal of product Q at store A).

Similarly, examining the different products using the table command reveals some errors, with both upper-case and lower-case versions of the product names. The product names need to be corrected to a consistent format.

For the three products, by examining the subset of data with zero units sold, we can observe that the various quartiles of price are similar to the other observations. We can derive Units as Revenue divided by Price for the subset of data with zero recorded for Units; the value reflected missing data rather than no units sold.

Re-examining the data, by using the summary command on the subset of data where the product is Q, we observe some very small values of Units. Examining these, it appears that the units are expressed in terms of thousands of units for some observations. This can be corrected by multiplying the Units variable by 1000 for such observations (for example those where the absolute value of Price*Units/Revenue minus one exceeds 0.1).

Examining the date variable, it appears that the date is in three formats: year-month-day, day-month-year, and day-month-year where the month is in a three-character format. These dates need to be converted into an appropriate consistent format, one that the software can recognise as a date format for the purpose of analysis (for example, allowing a time trend requires some numerical representation). The approach depends on the software used. Once this is achieved, the date convention used by United (Store H) can be fixed by adding 6 days to the date for store H's observations. Finally examining the unique dates in the data should reveal 104 dates from the start of 2019 to the end of 2020.

There are several duplicates in the data, these can be dropped easily in R or Stata using the unique() command or "duplicates drop". This should yield 2777 observations.

Stores

Using the summary command should reveal that there are issues with the Latitude variable, with a positive value (i.e. Northern Hemisphere) which is inconsistent with Australian data. Examination suggests that the negative sign is missing. The values of the store name and brand appear fine, and align with name used in the retail data.

[37] With the exception of the location of Store D; the address must be retrieved from Google Maps or similar provider, the coordinates will be subject to some negligible error.

After correcting the positive latitude by taking the negative value, we can examine the geographic distribution using mapping software. One of the stores appears to be east of where it should be, by looking up the address in Google Maps we can correct the location. We may also determine the locations of the other stores as a check, given there are only nine stores.

Sample

Using the summary command should reveal that there are issues with the Latitude variable, with a positive value (i.e., Northern Hemisphere) which is inconsistent with Australian data. Examination suggests that the negative sign is missing. The values of the store name appear to have errors, with some observations repeating the store name. By only taking the first character, we obtain a sample of 100 customers for each of the nine stores. We can examine the geographic distribution using mapping software, no obvious errors can be observed (aside from some in rivers/parks), and so we rely on the data as presented.

Fuel index

The A2328636K series of the ABS data should be extracted, and the date converted to a format consistent with that used in the cleaned retail data. This provides values of the automotive fuel price index at the end of each quarter. However, we require values for each date as used in the retail data so that we may merge the fuel price index to the retail data. There are several alternative approaches: one is to apply the quarter-end index value to all dates that fall within the quarter, alternatively one may perform interpolation so that the value depends on the value at the start of the quarter, the value at the end of the quarter, and the fraction of the quarter that has elapsed. In this context the former approach may suffice.[38] As a check, one could perform checks to see if residuals close to the start of a quarter are different to those close to the end of the quarter, or by examining model fit.

To use the fuel index, we will adjust revenues/prices in the retail data to reflect changes in the fuel index. To do this, we choose a base period (for example the value at the start of the retail data) and deflate prices/revenues using the value of the index relative to the base value.

Data format

Finally, we wish to use data in an aggregated and wide format for some regressions. To do this, we first sum units and revenue for each week/product, and then derive the price as revenue divided by units. This can be accomplished by using the collapse (sum) command in Stata. As we will be using prices/revenues adjusted for the fuel price index, we perform the adjustments to create AdjPrice and AdjRevenue variables. We also obtain log values of Units and AdjPrice

For regressions, we will use data in a wide format, this can be accomplished using dcast (reshape package) in R or “wide” command in Stata.

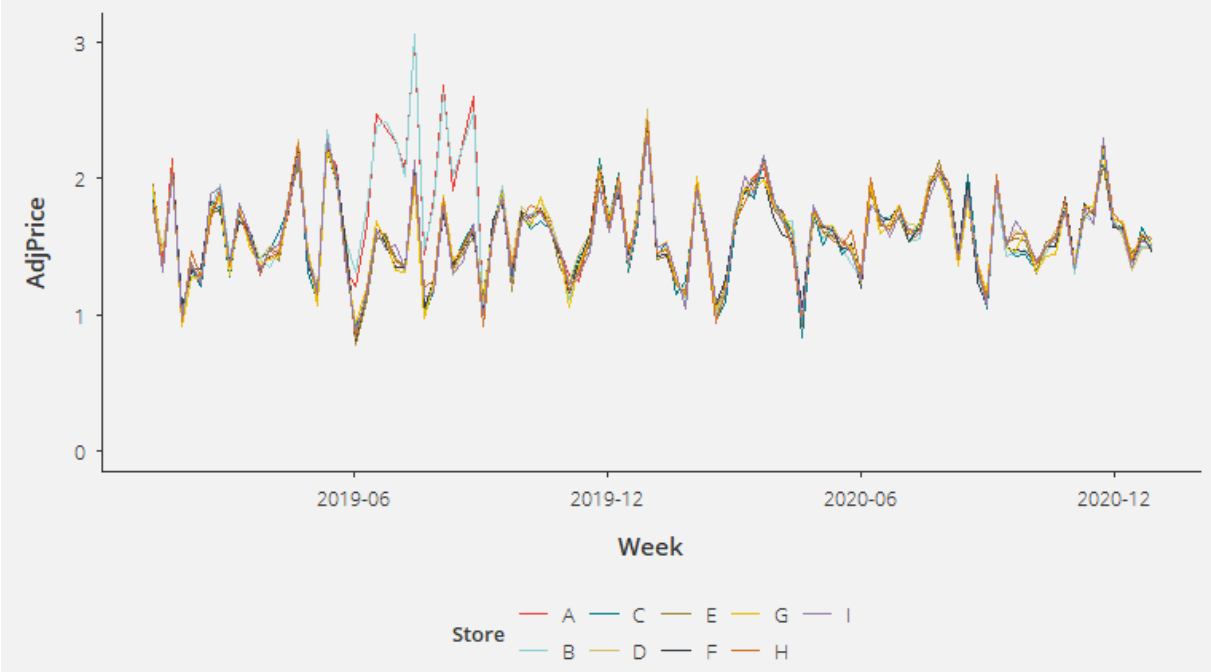
Exercise 2: Detecting a cartel

To perform this analysis, we first take the data in the long format, and select the observations corresponding to product Q.

We can then plot the adjusted price over time for the nine stores (data may need to be converted to the wide format for this). **Figure 6** shows that the prices of Q for the nine tend to be very close to each other in any given week, except for a period in June 2019 – August 2019 during which stores A and B priced well above the others.

[38] The former approach was used to generate the underlying data.

Figure 6: Collusion



Source: Frontier Economics analysis

To determine the impact of the alleged collusion, we can create a variable $Conduct_t$ that is equal to one if the week is during the alleged conduct period (June 2019 – August 2019), and two further variables equal to one if the store is A or B (A_i and B_i respectively).

One regression we might start by performing is $\log AdjPrice_{it} = \beta_1 + \beta_2 * Conduct_t + \beta_3 * (Conduct_t * A_i) + \beta_4 * (Conduct_t * B_i) + \epsilon_{it}$. Estimating this regression equation yields estimates of 0.403 and 0.411 on Conduct A and Conduct B, implying that pricing at A and B during the conduct period was 50% higher than expected.[39] The t-stats were 6.7 and 6.8 respectively (highly significant). One could go further and add dummy variables for the individual stores (this would not accomplish much as stores didn't tend to price high or low relative to others aside from the conduct). Dummies for individual weeks could also be used to reflect that while prices at individual stores vary over time, they are highly correlated with each other.[40,41]

Exercise 3: Defining geographic bounds of markets

First, take the store location dataset, keep the store name and latitude/longitude, and rename the coordinates to StoreLat/StoreLon or similar. Merge the customer location dataset with the store location dataset, and then find the distance between the customers lat/lon and the store lat/lon. In R this can be performed using the distHaversine function with a radius of around 6,371,100m. Alternatively the Haversine equation can be used manually in Excel.

We then take the 80th percentile of the 900 customer-store distances as the appropriate radius, obtaining 2011.7m, slightly over 2km.

[39] log(1.5)=0.405

[40] Adding both sets of dummies increased the t-stats on the two variables of interest to 35 and 37, while the coefficient estimates for Conduct*A and Conduct*B are 0.405 and 0.412. Note that due to multicollinearity the Conduct variable should be omitted.

[41] Normally it would not be beneficial to include so many explanatory variables. However, they substantially improve the fit in this circumstance.

We might be interested in whether the radius differs between stores, in particular if stores differ in some meaningful way (perhaps some brands have large convenience stores attached to the petrol station). Examining the 80th percentiles of the individual stores yields **Table 11** below which shows minimal differences between the stores.

Table 11: Individual store radii

Store	Radius (m)
A	1,918
B	1,965
C	1,891
D	2,012
E	2,023
F	2,134
G	1,982
H	1,814
I	2,142

Source: Frontier Economics

Exercise 4: Calculating price correlations to determine product markets

After aggregating retail sales across stores, adding prices adjusted for the fuel price index (while omitting revenues) and converting to a wide format, we obtain a dataset of 104 weekly observations with seven variables: the week, and prices both adjusted and unadjusted for each of the three products.

We then find the correlations between the prices by using the correlate commands in R/Stata or manually using correl() in Excel. This yields **Table 12** and **Table 13**.

Table 12: Correlations between prices

	Price_Q	Price_R	Price_S
Price_Q	1.000	0.221	0.142
Price_R	0.221	1.000	0.857
Price_S	0.142	0.857	1.000

Source: Frontier Economics

Table 13: Correlations between prices adjusted for the fuel price index

	AdjPrice_Q	AdjPrice_R	AdjPrice_S
AdjPrice_Q	1.000	-0.054	-0.202
AdjPrice_R	-0.054	1.000	0.758
AdjPrice_S	-0.202	0.758	1.000

Source: Frontier Economics

The results suggest that the prices of R and S are highly correlated, though there is little relationship between the price of Q and R/S once the common impact of the fuel price index is excluded as shown in **Table 13**, highlighting the need to control for common price drivers as much as possible

Exercise 5: Estimating demand elasticities to determine product markets

After aggregating retail sales across stores, adjusting prices for the fuel price index (while omitting revenues), taking log prices and log units converting to a wide format, we obtain a dataset of 104 weekly observations with seven variables: the week, and the logs of units and prices for each of the three products. We then perform three regressions: for each of product Q, R and S we have log of units sold as the dependent variable, with logs of adjusted prices of the three products as explanatory variables. **Table 14** provides the regression results.

Table 14: Regression results – adjusted prices

Dependent variable		(Intercept)	logAdjPrice_Q	logAdjPrice_R	logAdjPrice_S
LOGUNITS_Q	Coef	8.92	-0.99	-0.01	0.01
	t	(1905.27)	(-258.87)	(-1.03)	(1.28)
LOGUNITS_R	Coef	8.24	0.00	-2.30	1.32
	t	(1205.08)	(0.19)	(-168.59)	(83.86)
LOGUNITS_S	Coef	8.25	0.00	1.54	-2.53
	t	(1106.53)	(-0.16)	(103.20)	(-147.92)

Source: Frontier Economics

The results show that the own price elasticities are negative and highly significant as we would expect. The cross-price elasticities show strong substitution between R and S (positive, reasonably large and statistically significant). However, the impact of prices of R and S on sales of Q is small and insignificant, similarly for the impact of the price of Q on sales of R and S. We conclude that Q is not a substitute for R and S, and vice versa. It appears that there is a product market for Q, and a separate one for R and S combined.

Alternative, if we do not adjust for the fuel price index, we would obtain the results in **Table 15**. Note the significant impact of the price of Q on sales of R and S, and the significant impact of the price of S on sales on Q. These results, driven by the common price driver the fuel index, may incorrectly lead one to conclude that Q is in the same product market as R and S.

Table 15: Regression results

Dependent variable		(Intercept)	logPrice_Q	logPrice_R	logPrice_S
LOGUNITS_Q	Coef	8.64	-0.86	0.03	0.37
	t	(328.03)	(-30.4)	(0.38)	(4.56)
LOGUNITS_R	Coef	7.97	0.13	-2.26	1.66
	t	(309.98)	(4.64)	(-30.48)	(21.19)
LOGUNITS_S	Coef	7.97	0.13	1.57	-2.18
	t	(298.51)	(4.56)	(20.42)	(26.80)

Source: Frontier Economics

In addition, one could estimate elasticities using the AIDS framework. The micEconAids package in R may be used to estimate the demand system, the Marshallian elasticities that are then obtained are very similar to those in **Table 14**.

Exercise 6: Estimating a diversion ratio between stores

Using the cleaned retail dataset, take the subset of data for which the store is A or B and the product is Q. Create a new variable indicating if the week is after 1 June 2020 (call this Withdrawal).

Find the average sales of Q at A prior to the withdrawal, perhaps by looking at the period November 2019 through May 2020 (a period of seven months to match the seven-month withdrawal period would be sufficient). This gives 589 units, which we might take as what A would have expected to have sold if it did not withdraw Q. Next, find the increase in sales of Q at B during the withdrawal period by taking the subset of data for which the store is B, estimating the regression equation $Units_t = \beta_1 + \beta_2 Withdrawal_t + \epsilon_t$. This should give a coefficient of 497 on withdrawal, indicating that sales were 497 higher than expected. The diversion ratio can therefore be calculated as 497/589=84%. That is, 84% of the sales of Q lost by A were diverted to B. This implies that A and B are close competitors in the market for Q.

Exercise 7: Estimating a diversion ratio between products within a market

In this exercise we assume that A and B are in the same geographic market, with no other participants (we can check this by finding distances between stores). Using the cleaned retail dataset, take the subset of data for which the store is A or B and the product is R or S, and then aggregate over stores to find sales of R and S. Convert to wide format, create the withdrawal variable, and examine the data from November 2019 onwards.

Taking a similar approach to before, we use regression analysis to determine the impact of the withdrawal of Q at A on sales of R and S. The coefficient on withdrawal in the R regression is -35 with a t statistic of -1.2, implying that sales of R went down by 35 (relative to the intercept of 604 sales per week). Similarly, sales of S went up by 51 with a t statistic of 2.3 (relative to the intercept of 432 sales per week). This implies that diversion from Q to R/S is small if anything.

Merger

Examining the Stores dataset, we create two new variables equal to the latitude and longitude of store F, then calculate the distance from each store to F by comparing the two sets of lat/lon coordinates using the Haversine equation. We then take all stores that fall within the radius of 2011.7m determined earlier. This yields **Table 16** below.

Table 16: Constituents of the geographic market centred at F

Store	Brand	Distance
E	ColesExpress	1,825
F	BP	-
G	Caltex	450
H	United	848
I	BP	1,342

Source: Frontier Economics

We can then calculate distance weighted market shares as (Radius-Distance)/Radius for each store, then normalise so that shares sum to one. We then sum market shares by brand (assuming that all stores within a brand are controlled by that brand) to obtain market shares in **Table 17**.[42]

Table 17: Market shares of the geographic market centred at F

Brand	Share
ColesExpress	3.3%
BP	47.9%
Caltex	27.9%
United	20.8%

Source: Frontier Economics

The results that BP and Caltex have the highest market shares of the geographic market centred at F, with a 76% combined share post-merger, an increment of 28%. This would raise preliminary concerns; further analysis of the area should be undertaken. For example, the presence of supermarkets, whether stores are on the same major roadway, travel times between locations.

[42] Alternatively, one could distance weight the revenues based on the retail data.