Web scraping for food price research

Judith Hillen
Agroscope Tanikon, Ettenhausen, Switzerland

Received 4 February 2019
Revised 10 June 2019
4 September 2019
Accepted 6 September 2019

Abstract

Purpose – The purpose of this paper is to discuss web scraping as a method for extracting large amounts of data from online sources. The author wants to raise awareness of the method's potential in the field of food price research, hoping to enable fellow researchers to apply this method.

Design/methodology/approach – The author explains the technical procedure of web scraping, reviews the existing literature, and identifies areas of application and limitations for food price research.

Findings – The author finds that web scraping is a promising method to collect customised, high-frequency data in real time, overcoming several limitations of currently used food price data sources. With today's applications mostly focussing on (online) consumer prices, the scope of applications for web scraping broadens as more and more price data are published online.

Research limitations/implications – To better deal with the technical and legal challenges of web scraping and to exploit its scalability, joint data collection projects in the field of agricultural and food economics should be considered.

Originality/value – In agricultural and food economics, web scraping as a data collection technique has received little attention. This is one of the first articles to address this topic with particular focus on food price analysis.

Keywords E-commerce, Big Data, Data collection, Food price, Digitalization

Paper type Research paper

1. Introduction

Web scraping is a relatively new method for collecting online data. The term describes the automated process of accessing websites and downloading specific information, such as prices, from each (Kienle *et al.*, 2004). Allowing the creation of large, customised data sets at low costs, web scraping is already applied for scientific and commercial purposes in many areas, such as marketing, industrial organisations, or inflation measurement (for an overview, see Cavallo and Rigobon, 2016; Edelman, 2012).

In food price research, however, this data collection technique has received little attention. In agricultural economics and food system analysis, we mostly rely on more traditional data sources, such as official price indices or retail scanner data, of consumer prices. Yet, several issues are associated with these data sources.

For example, official prices and price indices for products, segments such as food, or even the whole economy are mostly published on a monthly or quarterly basis, with some publication delay. The public provision by official agencies and the availability of long time series are attractive for research purposes. Yet, one must rely on correct data collection, weighting and aggregation by official sources. Because, normally, no access to the raw data is given, it is not possible to detect errors or even manipulations (Cavallo, 2013).

In comparison, scanner data obtained at the point of sale at retailers are available at a higher frequency (generally weekly) and provide more details at the product level. A main advantage is that they include transaction data, i.e. the quantities purchased of a good at a given price (Campbell and Eden, 2014; Cotterill, 1994; Silver and Heravi, 2001). However, these data need to be purchased from market research institutes such as Nielsen N.V., and can be very costly, especially if longer time series or multiple retailers and locations are required.

As an increasing number of prices is published online and as online grocery retail is slowly gaining market shares in many parts of the world (Nielsen, 2015; Rigby, 2018), web scraping may be a promising alternative to get data for food price research.

British Food Journal
Vol. 121 No. 12, 2019
pp. 3350-3361
© Emerald Publishing Limited
0007-070X
DOI 10.1108/BFJ-02-2019-0081

In the following, we will not give detailed instructions on how to build a web scraper, and we will omit technical details and coding issues[1]. Rather, the aim is to discuss the method's potential for agricultural and food economics research. Section 2 gives an overview of what exactly web scraping is and how it works, and weighs the pros and cons regarding food price analysis. Section 3 reviews existing applications and considers further applications for studying online and offline food prices. The paper finishes with an outlook and suggestions on how this new data collection method could best be used in the discipline of agricultural and food economics.

2. About web scraping

2.1 Definition

Throughout this paper, we use the term web scraping. However, we found several related terms and concepts, which are not always distinctively defined (for definitions, see e.g. Kienle *et al.*, 2004; Massimino, 2016; Nakash *et al.*, 2015).

As a minimal definition, web scraping (or screen scraping, information scraping) describes the automated process of accessing web documents and downloading specific, pre-defined information, such as prices, from each, to then transform and save them into a structured format.

Web crawling, on the other hand, means accessing web content and indexing it via hyperlinks; thus, only the URL but no specific information is extracted. Instead, the full content is made available through the hyperlink but is generally not archived. Search engines, including Google, crawl the web, analyse the online content and compile all the links they find to match the search request. Crawlers (or spiders) are also used for price comparison tools. Shopbots are programmes that crawl websites to obtain price information from several sellers in order to find the lowest price (Hemenway and Calishain, 2004).

For a conceptual distinction between scraping and crawling, see Table I. The remainder of this paper deals only with web scraping because we are interested in collecting food price data for research from public websites.

2.2 Technical procedure

There are several ways to build a web scraper, and probably no one-size-fits-all approach exists. Although this paper does not aim to give detailed instructions on how to code a web scraper, we will briefly give an intuitive description of what a web scraper does technically and how this tool can be implemented for creating food price data sets.

Generally speaking, one needs to write a script that accesses the websites hosting the data, finds the relevant, previously defined elements and then downloads, and stores them in structured data sets. In any case, the price, product name and a timestamp recording when the content was accessed need to be stored to ensure a consistent output over time. If available and desired, additional information, such as package size, customer rating,

	Web scraping	Web crawling
Process	Automatically requesting web documents and collecting information from them	Repetitively finding and fetching hyperlinks starting from a list of initial URLs
Target information	Pre-defined data on specific websites	URLs to access all kinds of information, depending on search request
Output	Downloaded data in structured format	Indexed hyperlinks, stored in database
Use	Data collection (e.g. price series)	*Ad hoc* requests (e.g. search engines, price comparison tools)

Source: Own representation

Table I. Distinction between web scraping and web crawling

category, country of origin, labels, etc., can be included. Recording a unique product ID and the URL can help to trace any irregularities.

In principle, the script imitates a web user, navigates through the sites and extracts the pre-defined information. To keep the websites' traffic at a moderate rate, this download should happen with some delay between the requests (Hemenway and Calishain, 2004).

Figure 1 gives a schematic overview of the several steps of building a web scraper. Technically, the script can be coded in most of the common programming languages (Hemenway and Calishain, 2004; Massimino, 2016). Some languages, such as Python or R, even provide pre-programmed libraries for this purpose. Although such code elements are certainly helpful as building blocks, there is always a need to adapt the script to the targeted website, depending on how the website is set up, the page structure, authentication requirements, etc. because emany websites use reactive elements, simple HTTP requests cannot be used. Especially large online food retail websites mostly have online catalogues but no application programming interface to access their food prices directly (Papin *et al.*, 2018). Therefore, the script needs to navigate a "real" web browser and "click" through the website. Also, sufficient time should be allocated for debugging and testing. Once the script is successfully written and tested, its execution can be fully automated: a scheduler starts the download at defined time intervals. However, the scraper can be failure-prone to changes in a website's layout, product group structure or other even minor changes. To be aware of such issues, one can build in alerts, e.g. sending an e-mail if a script did not run through completely or if the download size is unusually small. The accessed data can be saved in any format, e.g. as a text or CSV file, with defined elements (name, ID, price, timestamp, etc.). For each observation over time, the new data are simply added to this file, resulting in one consistent data set. If several websites (e.g. different retailers or geographic locations) are scraped, separate files can be created, but formatting should be the same, simplifying later analyses.

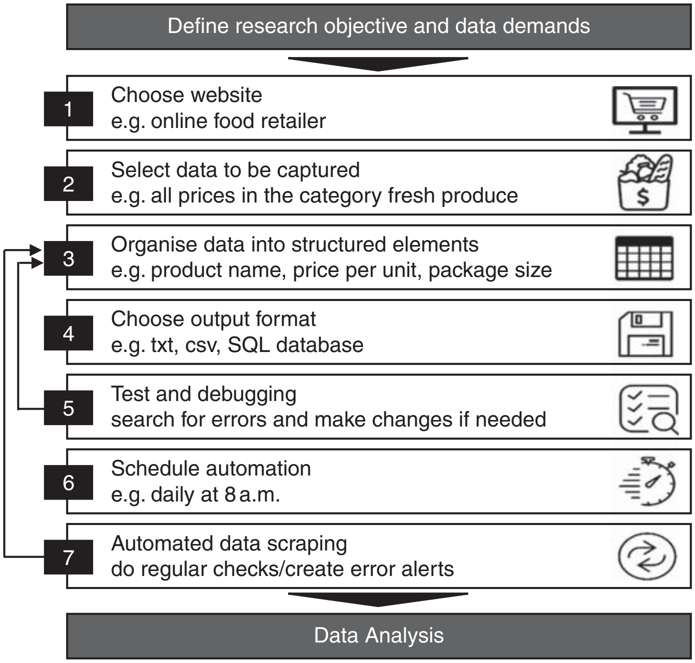
Define research objective and data demands
1
Choose website
e.g. online food retailer
2
Select data to be captured
e.g. all prices in the category fresh produce
3
Organise data into structured elements
e.g. product name, price per unit, package size
4
Choose output format
e.g. txt, csv, SQL database
5
Test and debugging
search for errors and make changes if needed
6
Schedule automation
e.g. daily at 8 a.m.
7
Automated data scraping
do regular checks/create error alerts
Data Analysis

Figure 1.
Schematic web scraping procedure

Source: Own representation, icons made by Freepik from www.flaticon.com

Because many applications require daily scraping, one may not want to occupy the personal computer with this task. Alternatives are a private server or dedicated single-board computers such as the Raspberry Pi.

If technological infrastructure or coding know-how is insufficient to build one's own solution, one can use commercial providers of web scraping services. However, such commercial options may lack the transparency and open code documentation required for most scientific research and peer-reviewed publications. For such purposes, it may be more appropriate to use and adapt open source libraries.

2.3 Advantages

For food price research, web scraping is a promising new method of data collection. It helps to overcome some issues of traditional data sources, such as official statistics and scanner data, as Table II shows.

Low costs. Good data sets can be expensive. Especially at the retail level, access to highly frequent and disaggregated data, such as scanner data, is costly. Collecting prices via web scraping is basically free if done with open source software. Costs associated with web scraping include the time required to write and test the code. At the defined times when the script is executed, electricity and online access are necessary. With all this in place, scalability across countries and products is high, decreasing the marginal cost per observation to almost zero (Cavallo, 2018). Alternatively, if the budget allows, tasks can be outsourced to an increasing number of commercial providers offering Data as a Service (Massimino, 2016).

Frequent, real-time sampling. Once the script is written, it is up to the user whether it should run and extract prices and other data monthly, weekly, daily or even at a higher frequency (e.g. hourly, as done by Ellison and Ellison, 2009). For food price research, daily data are probably sufficient for most applications. Such a high sampling frequency allows for a more detailed analysis of price dynamics and for the application of other statistical methods, compared with the analysis of time-aggregated price data (Edelman, 2012). Furthermore, as data are collected in real time, availability is given without any publication delay. This is an advantage for analysing recent events or policy changes, as well as for forecasting.

Product range and details. In official statistics, consumer prices are mostly reported for widely defined product or product category levels. Further details such as package size, brand, quality differentiation, etc., which may be interesting for food price analysis, are generally not available. In contrast, web scraping can be used to extract prices for precise products, including all available product attributes. This may be relevant information for some research questions. If not, taking this information as a starting point, the researchers themselves can aggregate the data to the level they need, using the methods they consider appropriate.

	Scraped data	Scanner data	National statistics data[a]
Cost per observation	Low	High	Free[b]
Data frequency	Daily	Weekly	Monthly
Real-time data	Yes[c]	No	No
Full product range	Yes[c]	No	No
Product details	Yes	Yes	Limited
International comparability	Yes	Limited	Limited
Transaction data	No	Yes	Yes (weighted)

Notes: [a]Price indices; [b]If publicly available; [c]nearly, accounting for download delay and potential errors
Source: Own representation based on Cavallo and Rigobon (2016, p. 156)

Table II. Alternative data source comparison

Store type. Most secondary data simply disclose that they were collected at “retail level”, or sometimes it is distinguished between supermarkets, discounters and small single retailer stores. With web scraping, the researchers decide which stores, retailers, wholesalers or online delivery service they are interested in and collect those prices, if available, without any hidden aggregation. To access different websites in regions and countries, self-identification, e.g. through entering a postal code or choosing a country, may be necessary and can be included in the web scraper code.

Transparency and customising. Although scanner data and many official statistics provide very reliable and well-structured data, they may not always contain all the information researchers ideally need for their projects. It could be that these data are aggregated to some level or are not available in the same structure or by the same provider for all the geographic locations of interest, resulting in limited international comparability. Web scraping allows the creation of customised data sets, targeted to the respective needs and transparent in how the data were obtained, without omitted variables or black boxes. If open source programmes are used and code and data are shared, this transparency and reproducibility is also given to the scientific community. Enabling researchers to integrate data collection into their empirical work instead of relying on secondary data may therefore improve the quality and precision of empirical research (Cavallo and Rigobon, 2016).

2.4 Limitations

Despite all the above-named advantages, the method has some limitations.

No historic data. Web scraping means collecting real-time data. Hence, to come up with a sufficiently long time series, one needs to start data collection from day zero of the respective time period. For *ad hoc* analyses, this may not always be possible. Not being able to access historic data is certainly a drawback of data collection via web scraping, requiring careful planning ahead.

Too Big Data. We saw that the marginal cost and effort of scraped data are minimal. Hence, it may be tempting to collect literally Big Data beyond what is needed to answer a given research question, or to start collecting data even before having a well-defined research question. Then, we may end up doing purely explorative data mining instead of theoretically motivated research (Massimino, 2016). Although computing power and storage possibilities are constantly improving, adequately analysing the data may be challenging, especially for professionals and researchers without much previous experience with Big Data handling. When considering web scraping, it is also worth considering what is sufficient for the planned research, regarding product scope, time frame, frequency and level of detail. For example, if aiming to measure overall food price development in a country or region, the selection of representative retailers and product categories will be a core part of the research design.

No transaction data. Generally, web scraping extracts only prices but no data on how often products are clicked on, or eventually purchased, because these pieces of information are not publicly available. This lack of transaction data certainly is a drawback, as it may be that, especially for high prices, no purchase is done at all, making the price completely irrelevant to consumers (Chevalier and Goolsbee, 2003). Like in any market, also online there are most likely some top-selling products and a lot of rarely purchased products. Products or brands with fewer buyers are also bought less frequently by those few buyers. This relationship is known as the double jeopardy (Ehrenberg *et al.*, 1990). Without transaction data, the above-mentioned product groups cannot be identified and analysed separately. However, Gorodnichenko *et al.* (2018) compared unweighted prices with price quotes weighted by clicks and found that they were quite similar. Categories like “bestsellers” or sorting by “most popular” can help to give some indication on frequently purchased products, if possible on a given website.

Online availability. Especially in developing countries in which food security and hence food prices at consumer level are an issue and of interest for research, online availability of prices on standardised websites may still not be given for the majority of transactions, particularly if informal markets are taken into account. Yet, even in many developing countries, those prices available online may be a more reliable source than official statistics (Cavallo, 2013). Moreover, the availability of prices published online is increasing globally. Regarding online grocery shopping, Asia-Pacific is leading in terms of market share, and especially Chinese consumers are quickly adopting online food shopping (Nielsen, 2017; Wang and Somogyi, 2018). Developing markets in all parts of the world are quickly catching up as internet and smartphone penetration rates increase, at least in urban areas (Nielsen, 2017).

Legal and ethical limitations. Web scraping itself is a technology and is not *per se* illegal or legal. Rather, one must assess the legal situation carefully for each individual application. Obviously, only public content should be accessed and copyright policies must be complied with. When downloading and using someone else's data, the respective Terms of Use or Terms of Service apply. Reading the Terms of Use and the Robots Exclusion Protocol (robots.txt file) are good starting points to see whether scraping data from a site is allowed or not (Kienle *et al.*, 2004). The robots.txt file is a clearly codified access policy in a standardised format and can be found at the URL http://[www.domain.com]/robots.txt (Hemenway and Calishain, 2004). The Terms of Use are often available somewhere on a website, but one does not explicitly need to agree with these ("browse-wrap agreement"). Whether they are enforceable contracts and legally binding has been examined by courts with different outcomes on a case-by-case basis (Toto and Buffington, 2016).

Also other aspects of how to deal with web scraping are still subject to discussion in the legal literature (see, for discussion, Hirschey, 2014; Zhu and Madnick, 2010), and the law is still evolving. In the past, particularly relevant for the legal evaluation seemed to be why a site was scraped and how the scraped data were used (Hirschey, 2014). Given this uncertain environment in most jurisdictions, one may want to seek professional legal advice before starting a research project.

Besides legal constraints, ethical concerns can and need to be debated. Especially large websites generally require user registration and are able to detect scrapers. Papin *et al.* (2018) propose a web scraping system that actively bypasses the security measures of websites by using a headless browser and the TOR network (Dingledine *et al.*, 2004) to stay completely anonymous. Although this approach is technically possible, the question remains whether it is ethically acceptable to adopt such practice for scientific, often publicly funded research. The best practice is certainly to obtain the explicit permission of the website operator.

3. Areas of research

So far, many web scraping applications have been used to analyse online pricing of non-food consumer goods. Because large-scale online retail started with product segments such as books and electronics, also most early research about pricing on the internet focussed on these products (e.g. Bakos, 1997, 1998; Chevalier and Goolsbee, 2003). Gorodnichenko and Talavera (2017) conducted a large study on online prices for more than 100,000 goods over five years, but, even here, fresh food products are not represented.

3.1 MIT Billion Prices Project

The largest web scraping effort for scientific purposes is MIT's Billion Prices Project, launched in 2008. Initially used as an alternative to relying on manipulated official statistics about Argentina's inflation rates (Cavallo, 2013), the price collection through web scraping was soon scaled up to other countries and an even broader product range. Cavallo and Rigobon (2016) describe how billions of online prices collected in this project help to triangulate inflation measurement. This project does pioneer work not just regarding the scope of the data collection but also with its approach to make much of the data publicly accessible to other researchers.

3.2 Online retail pricing strategies

It is often assumed that online posted prices are more flexible compared with offline prices, which display rather rigid pricing patterns, except for temporary promotions (Herrmann *et al.*, 2005). In online markets, menu costs to change prices are negligible, allowing suppliers to adjust prices at a high frequency, reacting to demand and supply changes, and ultimately leading to more efficient markets (Gorodnichenko and Talavera, 2017; Smith *et al.*, 2001; Tang and Xing, 2001). On the consumer side, search costs for prices have decreased, and price comparison for a defined good is quick and easy, especially thanks to price comparison websites (Bakos, 1997; Gorodnichenko and Talavera, 2017).

This flexibility on the seller and buyer side is assumed to increase price conversion (i.e. less price variation across different sellers of the same good) and price transmission (i.e. faster and more complete passing-on of price signals from other markets). No longer having to deal with menu and search cost, which are commonly used to explain why the law of one price[2] does not hold, there is hope that online prices allow for new insight regarding price transmission and conversion (Gorodnichenko *et al.*, 2018). Some empirical studies have shown that online prices change more frequently, and in smaller magnitude, than offline prices (Brynjolfsson and Smith, 2000; Ellison and Ellison, 2009). However, these data were collected on marketplaces such as eBay or price comparison tools (Google Shopping), which may not be representative of overall online retail because online retailers are heterogeneous in their characteristics and price setting (Einav *et al.*, 2018; Pan *et al.*, 2002).

An alternative strand of literature concludes that exactly the opposite is true, that prices do not converge more, but that online markets give even more room to differentiate between customer groups with different price sensitivities and to apply targeted price discrimination (Ancarani, 2002; Baylis and Perloff, 2002). On top of that, the convenience of online shopping and home delivery may attract less price-sensitive consumer segments (Degeratu *et al.*, 2000).

Online retail could even allow for dynamic pricing based on the analysis of current and past customer demand, competitor price setting, and other factors such as holidays or weather conditions (Grewal *et al.*, 2011; Shpanya, 2013). However, detecting such practices would require a very advanced web scraping code, pretending to log on from different IPs, with different user profiles. Here, again, the question is whether it is ethical for researchers to adopt the ruse of a legitimate customer inquiry for their data collection.

3.3 Online grocery retail

Even in the quite well-studied non-food sector, we found no consistent results showing how online price setting may differ from offline price setting. For online grocery retail, and fresh products in particular, even less is known about online pricing, and empirical studies are scarce, especially ones with large sample sizes.

There is no clear evidence regarding price-level differences between online and offline grocery sellers. *Ad hoc*, non-representative sample observations in the UK and USA suggest that online grocery retailers were more expensive for a long time, but have recently lowered prices, even below established offline supermarket prices (Oliver Wyman, 2014, 2018).

Cavallo (2017) included food items in his comparison of online and offline prices, conducted in 10 countries between 2015 and 2016 for more than 24,000 products. For the food subsample (5,953 observations), there was a small mark-up of online prices (1 per cent), whereas non-food items were on average even slightly cheaper online than offline (drugstore −3 per cent, household −2 per cent). The results differed among countries and only measured within-retailer price dispersion for multi-channel retailers. Online-only retailers or brick-and-mortar stores, such as traditional small shops or discount stores, were not considered. Yet, the results suggest differences between food and non-food online pricing.

Also, the argument of reduced search costs through online price comparison websites does not seem to hold for online food retail. Currently, such tools are mostly available for homogeneous, durable, and easy-to-ship products, but not for (fresh) foods with differences in appearance, freshness, and taste and presumably limited arbitrage opportunities (Gorodnichenko and Talavera, 2017). Furthermore, groceries are generally not bought as individual items but bundled (e.g. per week or at least per meal), making an individual price comparison futile.

Fedoseeva *et al.* (2017) analysed the price setting in the German online chocolate market, comparing daily prices of twelve products across eight sellers over three months. They found no evidence for more homogeneous prices among different sellers (lower search cost hypothesis) or more frequent price adjustment (lower menu cost hypothesis) than in offline markets. Yet, data were collected manually, sample size was small and time span was rather short.

Applying web scraping methods could help to understand pricing in the growing online food retail business. Some recent studies made use of website content analysis, e.g. to better understand online food delivery companies' performance, but they also did not use an automated web scraping procedure (Pigatto *et al.*, 2017).

We found one very recent effort to scrape online food prices by Papin *et al.* (2018). The authors proposed to continuously track online food prices in multiple locations through web scraping. As a proof of concept, they collected food prices from Monoprix.fr, the website of a major French retail chain. However, so far they have not published any analyses of their collected data.

3.4 Overall food price dynamics

Web scraping can only access prices published online. Nonetheless, it gives insight into offline and overall price development. As discussed in Section 3.3, the online–offline price-level difference for food seems to be small in most countries (Cavallo, 2018, 2017; Oliver Wyman, 2014, 2018). Also, in many countries, offline retailers have started to publish their prices online (Nielsen, 2017). Hence, web scraping can provide data on overall price developments. To date, this method has been used to calculate general consumer price indices, both in research and by official statistical agencies and national banks (Cavallo and Rigobon, 2016).

3.5 Beyond consumer price research

So far we have focussed on consumer price research because these data are currently most widely available online. However, as online publishing becomes more common, web scraping applications could expand to a wholesale and production level, as well as farm input prices.

Besides the use for scientific research, there are also more practical field applications. Scraping market prices and forwarding relevant price developments to farmers' mobile phones could help them to improve agricultural production and marketing decisions. There is evidence that access to information technology can improve farmers' and overall welfare (e.g. Jensen, 2007; Aker, 2011). Camacho and Conover (2011) found that Colombian farmers who received targeted digital price and weather information on their phones considered this useful and managed to reduce their crop losses.

Such projects could be scaled up at relatively low cost, exploiting the high scalability of web scraping. Also, the technology is not limited to prices but can be used for other reasonably well-structured information available online, such as weather data (Yang *et al.*, 2010).

4. Conclusion

We have shown that web scraping is a promising, low-cost data collection method for food price research. Using web scraping to build customised data sets can help to overcome common problems such as incomplete data, omitted variables and sample selection. So far, the main application consists of online retail prices because of their good availability. Data obtained through web scraping can help to analyse how prices are set in the growing online food retail, and how this price setting may impact the value chain and potentially change global food systems. Also, online prices give insight into general (offline) price development and can help to fill in gaps where public data sources are not available or reliable. As more and more price data are published online, areas of application for web scraping will also expand. Especially for global data collection and comparisons, this method will open up new research opportunities. Exploiting this new technology gives us the chance to find new answers to old questions, or even to ask new questions. Once aware of the technical possibilities, researchers and practitioners may come up with innovative applications.

However, we found some limitations that may discourage or inhibit individual researchers to engage in web scraping, such as ethical and legal uncertainties, unavailability of historic data, and technical difficulties. Working together and centralising data collection can help to overcome these barriers (Massimino, 2016; Papin *et al.*, 2018). Larger data collection projects could bundle technical and legal expertise and exploit the scalability of web scraping technologies. A good example of such a project is the herein discussed MIT Billion Prices Project, which makes the scraped data publicly available for scientific use. A similar web scraping project could be started for food price research, collecting and publishing consistent and cleaned data. After some time of ongoing real-time data collection, the data could benefit a wide range of researchers in agricultural economics, development economics and related disciplines.

Yet, some universities, institutes or associations would need to take the lead and make an initial investment. In a first step, the initiators could organise a public repository, where researchers can upload and share their scraped data sets that may be relevant to colleagues who also study food prices. Although competition for the best scripts and biggest data may be, in part, fruitful, cooperating and sharing data sources and methods to advance in research is likely to lead to less biased and broader data sets as well as to better results.

Notes

1. For readers interested in a technical guide to web scraping, we recommend the book *Practical Web Scraping for Data Science* by vanden Broucke and Baesens (2018).
2. The law of one price is an economic theory stating that a good should have the same price anywhere, if traded in a free market. In the long run, any price differences should be eliminated due to arbitrage opportunities.

References

Aker, J.C. (2011), "Dial 'A' for agriculture: a review of information and communication technologies for agricultural extension in developing countries", *Agricultural Economics*, Vol. 42 No. 6, pp. 631-647.

Ancarani, F. (2002), "Pricing and the internet: frictionless commerce or pricer's paradise?", *European Management Journal*, Vol. 20 No. 6, pp. 680-687.

Bakos, J.Y. (1997), "Reducing buyer search costs: implications for electronic marketplaces", *Management Science*, Vol. 43 No. 12, pp. 1676-1692.

Bakos, Y. (1998), "The emerging role of electronic marketplaces on the internet", *Communications of the ACM*, Vol. 41 No. 8, pp. 35-42.

Baylis, K. and Perloff, J.M. (2002), "Price dispersion on the internet: good firms and bad firms", *Review of Industrial Organization*, Vol. 21 No. 3, pp. 305-324.

Brynjolfsson, E. and Smith, M.D. (2000), "Frictionless commerce? A comparison of internet and conventional retailers", *Management Science*, Vol. 46 No. 4, pp. 563-585.

Camacho, A. and Conover, E. (2011), "The impact of receiving price and climate information in the agricultural sector", IDB Working Paper Series No. IDB-WP-220, Washington, DC.

Campbell, J.R. and Eden, B. (2014), "Rigid prices: evidence from U.S. scanner data", *International Economic Review*, Vol. 55 No. 2, pp. 423-442.

Cavallo, A. (2013), "Online and official price indexes: measuring Argentina's inflation", *Journal of Monetary Economics*, Vol. 60 No. 2, pp. 152-165.

Cavallo, A. (2017), "Are online and offline prices similar? Evidence from large multi-channel retailers", *American Economic Review*, Vol. 107 No. 1, pp. 283-303.

Cavallo, A. (2018), "Scraped data and sticky prices", *Review of Economics and Statistics*, Vol. 100 No. 1, pp. 105-119.

Cavallo, A. and Rigobon, R. (2016), "The billion prices project: using online prices for measurement and research", *Journal of Economic Perspectives*, Vol. 30 No. 2, pp. 151-178.

Chevalier, J. and Goolsbee, A. (2003), "Measuring prices and price competition online: Amazon.com and BarnesandNoble.com", *Quantitative Marketing and Economics*, Vol. 1 No. 2, pp. 203-222.

Cotterill, R.W. (1994), "Scanner data: new opportunities for demand and competitive strategy analysis", *Agricultural and Resource Economics Review*, Vol. 23 No. 2, pp. 125-139.

Degeratu, A.M., Rangaswamy, A. and Wu, J. (2000), "Consumer choice behavior in online and traditional supermarkets: the effects of brand name, price, and other search attributes", *International Journal of Research in Marketing*, Vol. 17 No. 1, pp. 55-78.

Dingledine, R., Mathewson, N. and Syverson, P. (2004), *Tor: The Second-Generation Onion Router*, US Naval Research Laboratory, Washington DC.

Edelman, B. (2012), "Using internet data for economic research", *Journal of Economic Perspectives*, Vol. 26 No. 2, pp. 189-206.

Ehrenberg, A.S.C., Goodhardt, G.J. and Barwise, T.P. (1990), "Double jeopardy revisited", *Journal of Marketing*, Vol. 54 No. 3, pp. 82-91.

Einav, L., Farronato, C., Levin, J. and Sundaresan, N. (2018), "Auctions versus posted prices in online markets", *Journal of Political Economy*, Vol. 126 No. 1, pp. 178-215.

Ellison, G. and Ellison, S.F. (2009), "Search, obfuscation, and price elasticities on the internet", *Econometrica*, Vol. 77 No. 2, pp. 427-452.

Fedoseeva, S., Grein, T. and Herrmann, R. (2017), "How German online retailers price foods: an empirical analysis for chocolate products", *International Journal on Food System Dynamics*, Vol. 8 No. 1, pp. 32-44.

Gorodnichenko, Y. and Talavera, O. (2017), "Price setting in online markets: basic facts, international comparisons, and cross-border integration", *American Economic Review*, Vol. 107 No. 1, pp. 249-282.

Gorodnichenko, Y., Sheremirov, V. and Talavera, O. (2018), "Price setting in online markets: does IT click?", *Journal of the European Economic Association*, Vol. 16 No. 6, pp. 1764-1811.

Grewal, D., Ailawadi, K.L., Gauri, D., Hall, K., Kopalle, P. and Robertson, J.R. (2011), "Innovations in retail pricing and promotions", *Journal of Retailing*, Vol. 87 S1, pp. S43-S52.

Hemenway, K. and Calishain, T. (2004), *Spidering Hacks*, O'Reilly Publishing, Sebastopol, CA, ISBN 0596005776.

Herrmann, R., Moeser, A. and Weber, S.A. (2005), "Price rigidity in the German grocery-retailing sector: scanner-data evidence on magnitude and causes", *Journal of Agricultural & Food Industrial Organization*, Vol. 3 No. 1, pp. 1-37.

Hirschey, J.K. (2014), "Symbiotic relationships: pragmatic acceptance of data scraping", *Berkeley Technology Law Journal*, Vol. 29 No. 4, pp. 897-927.

Jensen, R. (2007), "The digital provide: information (technology), market performance, and welfare in the South Indian fisheries sector", *The Quarterly Journal of Economics*, Vol. 122 No. 3, pp. 879-924.

Kienle, H.M., German, D. and Müller, H.A. (2004), "Legal concerns of web site reverse engineering", Web Site Evolution, *Conference Proceedings of the 6th International Workshop on Web Site Evolution (WSE 2004) – Testing, Chicago, IL, 11 September*, pp. 41-50, doi: 10.1109/WSE.2004.10000.

Massimino, B. (2016), "Accessing online data: web-crawling and information-scraping techniques to automate the assembly of research data", *Journal of Business Logistics*, Vol. 37 No. 1, pp. 34-42.

Nakash, J., Anas, S., Ahmad, S.M., Azam, A.M. and Khan, T. (2015), "Real time product analysis using data mining", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, Vol. 4 No. 3, pp. 815-820.

Nielsen (2015), "The future of grocery: E-commerce, digital technology and changing shopping preferences around the world", available at: www.nielsen.com/wp-content/uploads/sites/3/2019/04/nielsen-global-e-commerce-new-retail-report-april-2015.pdf (accessed 24 September 2019).

Nielsen (2017), "What's in-store for online grocery", available at: www.nielsen.com/wp-content/uploads/sites/3/2019/04/nielsen-global-connected-commerce-report-january-2017.pdf (accessed 24 September 2019).

Oliver Wyman (2014), "AmazonFresh in the U.S.", available at: www.oliverwyman.de/content/dam/oliver-wyman/global/en/2014/aug/OW_AmazonFresh_ENG.pdf (accessed 24 September 2019).

Oliver Wyman (2018), "AmazonFresh undercuts supermarkets by 10–20 percent", 4 January, available at: www.oliverwyman.com/our-expertise/insights/2018/jan/amazon-fresh-undercuts-supermarkets-by-1020-percent.html (accessed 24 September 2019).

Pan, X., Ratchford, B.T. and Shankar, V. (2002), "Can price dispersion in online markets be explained by differences in e-tailer service quality?", *Journal of the Academy of Marketing Science*, Vol. 30 No. 4, pp. 433-445.

Papin, J., Andrès, F. and d'Orazio, L. (2018), "A method to build a geolocalized food price time series knowledge base analyzable by everyone", First Latin America Data Science Workshop, 27 August, Rio de Janeiro, available at: http://ceur-ws.org/Vol-2170/paper13.pdf (accessed 3 September 2019).

Pigatto, G., Machado, J.G.D.C.F., Negreti, A.D.S. and Machado, L.M. (2017), "Have you chosen your request? Analysis of online food delivery companies in Brazil", *British Food Journal*, Vol. 119 No. 3, pp. 639-657.

Rigby, N. (2018), "Amazon grocery 2017 review: one click retail study", 16 January, available at: http://oneclickretail.com/amazon-grocery-2017-review/ (accessed 24 January 2019).

Shpanya, A. (2013), "Five trends to anticipate in dynamic pricing [Blog post]", Retail Touchpoints, 14 May, available at: www.retailtouchpoints.com/features/executive-viewpoints/5-trends-to-anticipate-in-dynamic-pricing (accessed 24 January 2019).

Silver, M. and Heravi, S. (2001), "Scanner data and the measurement of inflation", *The Economic Journal*, Vol. 111 No. 472, pp. 383-404.

Smith, M.D., Bailey, J. and Brynjolfsson, E. (2001), "Understanding digital markets: review and assessment", MIT Sloan School of Management Working Paper No. 4211-01, Cambridge, MA.

Tang, F.-F. and Xing, X. (2001), "Will the growth of multi-channel retailing diminish the pricing efficiency of the web?", *Journal of Retailing*, Vol. 77 No. 3, pp. 319-333.

Toto, C.S. and Buffington, K. (2016), "How binding is your browsewrap agreement? [Blog post]", Pillsbury Winthrop Shaw Pittman LLP, 6 June, available at: www.lexology.com/library/detail.aspx?g=4b4b93da-c40a-4724-916c-3bdd9011698d (accessed 24 September 2019).

vanden Broucke, S. and Baesens, B. (2018), *Practical Web Scraping for Data Science*, Apress, New York, NY, ISBN-13: 978-1-4842-3582-9.

Wang, O. and Somogyi, S. (2018), "Consumer adoption of online food shopping in China", *British Food Journal*, Vol. 120 No. 12, pp. 2868-2884.

Yang, Y., Wilson, L.T. and Wang, J. (2010), "Development of an automated climatic data scraping, filtering and display system", *Computers and Electronics in Agriculture*, Vol. 71 No. 1, pp. 77-87.

Zhu, H. and Madnick, S.E. (2010), "Legal challenges and strategies for comparison shopping and data reuse", *Journal of Electronic Commerce Research*, Vol. 11 No. 3, pp. 231-239.

Corresponding author
Judith Hillen can be contacted at: judith.hillen@agroscope.admin.ch