5. Big data for consumer policy

5.1 Focus of the use case

Consumer price indices are a key indicator for economic and monetary policy as it is one of the gauges for the cost of living for citizens. **Monitoring sales conditions** is relevant for consumer protection policy and for the establishment of the European Digital Single Market and the European Internal Market.

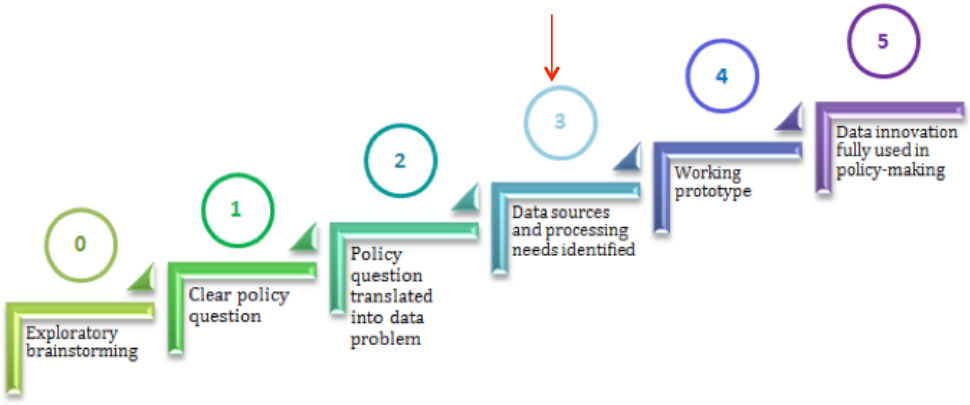
This use case explores the possibilities and potential benefits of **online retail monitoring for consumer policy**. The focus is on potential areas for retail monitoring:

1. The monitoring of e-commerce websites compliance to European consumer protection policy (sales conditions), among which:
   - Illegal sales practices (pre-ticked boxes, coupled sales);[34]
   - Geographical price discrimination: "versioning", territoriality clauses with wholesalers;
   - Availability of contracts; and
   - Availability of contact details of the merchant.
2. The approximation of Consumer Prices Indices (CPI), used for key economic indicators such as inflation.

In these two areas, **web scraping** is a promising approach. Web scraping, in short, involves automatically browsing relevant websites, mining relevant data, then cleaning the data with human input or eventually machine learning, and then running data analytics that report on consumer prices and, for example, price differences per country for the same product at the same retailer, or coupled sales of insurances with flights using pre-ticked boxes (both are disallowed practices).

From a design perspective, this use case has reached stage 3 (Figure 11): the policy problem is clearly identified and so are the data sources and the processing needs.

Figure 11 Use case readiness level



[34] http://europa.eu/rapid/press-release_MEMO-11-675_nl.htm

5.2 The rationale

An emerging challenge in consumer policy relates to the effective monitoring, implementation and enforcement of consumer protection regulations in the context of **online commerce**. Real-time monitoring of disallowed practices may feed into the all stages of the policy cycle and allow for determining where to focus efforts for enforcement (or how to adapt policies). Consumer price monitoring based on web scraping data would yield more up-to-date CPI based on a wider array of products. This would enable policymakers – responsible for monetary, economic, competition policy etc. - to observe the effects of their policy interventions earlier and at a more detailed level. It also provides them with early signals about inflation, deflation and (indirectly) competition intensity and economic growth, which could trigger timely debates about policy interventions

Using web scraping on websites that sell products or services can return vast sets of information that can be used for multiple purposes. In this use case, two aspects are highlighted: **consumer price monitoring** (i.e. the price of the items themselves) and **sales conditions monitoring** (i.e. the conditions under which the items are sold to the customer).

The possibilities of gathering data for consumer price indices have been demonstrated in other publications. Several parties such as the *Billion Prices Project*[35] and *Eurostat* together with *Statistics Netherlands*[36] have demonstrated a working method for consumer price monitoring. There are also commercial suppliers of CPI data based on web scraping such as *Pricestat*.[37] The literature concludes that CPI composition through web scraping is possible, accurate and reliable. However, changes in website layouts, product assortments and product quality (and thus price), and the difficulty of tracking the same or similar products between different retailers (because of the different names or descriptions for the same product), imply that **the exercise remains a challenge and requires human attention to the job**.

Because of this level of maturity of web scraping for consumer prices monitoring, we will further disregard the consumer price component of the case and focus on **sales conditions**.

The term 'sales conditions' is used for all aspects related to the context of the online sale of products and services. For example, retailers may couple the sale of an item with a service (e.g. extended guarantee) using pre-ticked selection boxes, or be unclear on the total costs of purchase. Moreover, retailers could discriminate, without clear justification, between customers according to the geographical origin of the customer or make specific "versions" of the same product per country, on which a price differential is based. These practices are illegal based upon the European Commission Directive 2011/83/EC (Consumer Rights Directive, CRD). However, manual monitoring requires significant resources. For example, policymakers and consumer protection agencies could scan and check a sample of relevant websites. One could also rely on reporting by consumers, but this may not give a complete and actual representation, and it would imply that actions are taken only when consumers complain.

5.3 The policy context

To accelerate progress towards the **Digital Single Market** and the **European Internal Market** are two priorities of the European Commission. As such, enhanced monitoring of retail prices and sales condition can contribute directly to pursuing Commission objectives. Unrestricted market access for consumers favours effective pricing, while a digital single market for retailers ensures the widest possible audience.

---

[35] http://bpp.mit.edu
[36] http://ec.europa.eu/eurostat/web/ess/-/web-scraping-price-collection-and-detailed-average-prices
[37] http://www.pricestats.com

More specifically, the digital single market would benefit from easier e-commerce and would require addressing the barriers mentioned above as well as outright geo-blocking.

The European Market transcends national borders and is therefore a European responsibility. The most relevant regulation for consumer policy is the **Directive on Consumer Rights** (2011/83/EC). The directive must be transposed into national regulations by Member States and falls under the responsibility of Directorate-General for Justice and Consumers. One of the consequences of European policy through a Directive rather than a Regulation is that national implementations differ, which may be a barrier to e-commerce.

5.4 The data process: from data collection to analysis and visualisation

Data sources

The data needs for monitoring sales conditions are already available; all the data is out there on the internet. The challenge is in designing **webscraping algorithms** that can extract the right kind of information from as many websites as possible and that can handle changes in the websites. Furthermore, changes in the data would ideally be monitored too, so as to monitor for any changes in prices and practices of online retailers.

The **data sources for this case are websites** offering goods for sale over the internet. This is a vast collection of websites that cannot be analysed completely; web scrapers require specific tuning per website so as to be able to find and harvest the desired information. The Billion Prices Project uses information from "hundreds of online retailers", updated on a daily basis. For this use case, the selection would be hundreds of European retailers. The number of websites to be covered can increase, as the automated procedures are improved.

The selection of retail websites should be broad enough to get a decent spread over sectors and products, but can include some duplication of products or sectors so as to compare sales conditions for similar items by similar retailers in various countries.

Data collection process

The data would be collected using **web scraping** and **simulated customer visits**. As mentioned before, these methods require tailoring to the websites visited: the more detail one requires, the more tailoring work is required for the web scraper. Price information may be the easiest to recognise (hence this kind of information is already made available on a commercial basis).

One of the possibilities to monitor illegal sales conditions is to scan websites for the presence of pre-ticked selection boxes; e.g. a pre-ticked box for additional insurance with airplane tickets. Another possibility is to monitor for barriers in international trade using mystery shoppers[38]: actual persons or (in a couple of years) web scrapers that operate from different locations and mimic a shopping process up to the point of ordering, then breaking the operation and repeating it with slightly changed parameters (country of origin, shipping method, etc.) and thus filling a dataset.

Web scrapers would further download documents such as standard conditions and contracts, and scan for the presence of the merchant's contact details and a privacy statement. A deeper scraping would yield information on the presence of pre-ticked boxes and possibly price differentiation varying with the customer's locale.

The technological maturity of web scrapers is high, but the algorithms are not yet autonomous because semantic algorithms are not advanced enough to recognize

[38] https://ec.europa.eu/jrc/sites/default/files/JRC92294_Supply%20side%20barriers%20to%20ecommerce.pdf

the unstructured information that guides human visitors through the page. Recognising sales items in a grid such as on a product listing website has been done by various teams, but buying more complex services (energy, air fares, insurances) and configurable items such as computers or cars would be too farfetched for web scrapers because the process of acquiring these items is more complex to navigate through.

The technical infrastructure one would need for web scraping is rather limited; any consumer grade computer would be able to run at least a dozen actions simultaneously. Scraping for compliance monitoring on a European scale requires some more computational power. However, most effort (and thus expenses) would be put in keeping the scrapers in tune with the websites they're scraping.

In this context the importance of the **human component** in monitoring web practices must be mentioned. An online sales process usually requires multiple steps; selecting the product, possible additions to the product, shipping and payment information; and these steps often vary per retailer that have websites with varying layouts, data structures and sales processes. To fully automate this process one would need to run semantic analyses on the website's content or apply machine learning. In essence, gathering more detailed information requires more effort and computers are not always up to that task.

A remark has to be made that **retailers do not always appreciate automated visits**.[39] Documented cases exist where retailers sue exploiters of automated visitors, either because the scraping generates additional load to the website's servers that do not generate revenue for the operator (trespass to chattels[40]), because the information on the website being scraped falls under copyright law, or because automated visits violate the website's terms of use. The first argument is most relevant in the case of automated visits by government agencies: if every government agency in the world started scraping websites for various policy enforcement purposes, the load incurred on the service's servers may become unduly high. Although the argument of trespass to chattels is no longer upheld in US jurisprudence[41], the argument of unfair server load generation remains an interesting one. Once detected, automated visiting can be sabotaged by changing the website's layout or (invisible to human visitors) some technical aspects of the website. Terms of use stated by a website may also prohibit automated visits.[42]

Based on the assumption that web scraping done by or for the EC for (compliance) monitoring purposes would not be hindered by legal issues, a data point from a single virtual visit would, for example, contain the following parameters (next page).

[39] https://en.wikipedia.org/wiki/Web_scraping#Legal_issues

[40] Trespass to chattel is a particular type of trespass whereby a person has intentionally interfered with another person's lawful possession of a chattel. A chattel refers to the movable or immovable personal property of an individual except real estate. Generally, the basic elements of a claim of trespass to chattels are lack of an owner's consent to trespass, interference with possessory interest, and intention of the trespasser (US Legal, 2014).

[41] http://blog.icreon.us/web-scraping-and-you-a-legal-primer-for-one-of-its-most-useful-tools/

[42] http://www.out-law.com/en/articles/2015/january/website-operators-can-prohibit-screen-scraping-of-unprotected-data-via-terms-and-conditions-says-eu-court-in-ryanair-case/

Table 3 Example of survey structure of web scraping for consumer policy monitoring

Parameter	*Format*
Date	dd/mm/yyyy
Origin of retailer	ISO country code
Origin of (automated) shopper	ISO country code
URL of website	text
Unique ID of item	number
Specifications of item	Standardised list: product type, purpose, (technical) specifications, etc.
Price of item in country of retailer	number in EUR
Prices of item in other countries*	number in EUR per ISO country code
Appearance of pre-ticked boxes during sales process	yes/no
If pre-checked boxes: what items were checked	List of all items checked in pre-ticked boxes
Available languages for contract	List of contract languages (ISO country code)
Content of contracts per language	Freeform text
Availability of contact details	Yes/no
Availability of privacy statement	Yes/no
Content of privacy statement	Freeform text

*Elaborate on method: spoof IP for geolocation or go through sales process on website to select country of delivery or country of customer.

If non-compliance is encountered, an automated complaint could be filed taking into account the European Consumer Complaint Registration system.[43]

Collecting data can be done per country, by national statistical offices. Of course, these agencies can collaborate and exchange tools and lessons. How compliance monitoring would fit in the portfolio of statistical offices is not entirely clear. If the focus is on legal and/or trade issues, there might also be a role for national justice departments or the department of trade/economics/consumers. Another data collection stakeholder would be national consumer agencies. However, A CPI web scraper would be the first step towards building an automated consumer policy monitor, and it might be inefficient to duplicate the effort elsewhere for another, extended purpose: compliance monitoring. The data collection would ideally be run by a single agency with assistance from relevant government departments and possibly consumer organisations.

Another option to collect data is to use an **Application Programming Interface** (API). APIs enable computer systems to communicate directly through standardised commands and replies. The website in between, which is in fact no more than an interface between the human user and the information on the server, could be circumvented using an API. For CPI monitoring this could be a very effective way of standardising data gathering while reducing server load. For what concerns compliance monitoring, an API would be less useful as the API cannot shed light on how the information present on the website is presented to the user as the actual content of the page is omitted.

[43] http://ec.europa.eu/consumers/consumer_evidence/data_consumer_complaints/index_en.htm

Data analytics and visualisation

The data gathered from multiple websites on a vast amount of products needs to be **cleaned**, so identical products can be identified on multiple websites and the product's price and sales conditions can be monitored over time. To make a comparison one has to find similar products to compare with and verify whether or not they are the same. For some items this is easily done. Especially consumer electronics often have a version or model identifier. However, versioning[44] also occurs and this obfuscates the comparison process. For clothing, this may be more difficult. Researchers from Statistics Netherlands encountered many challenges in matching clothing items across websites. To overcome product-matching challenges, a categorization was introduced.[45] The pilot study attained a 98% correct matching of categories (within the same website), but had to admit that this was also due to a well structured website. The researchers explained that their categories are broad but that adding more categories would cause the data set to become inconsistent over time; some categories would (dis)appear following fashion trends.

As for compliance monitoring, a web scraper could identify the presence of a contact address, a privacy statement and contracts. The actual content and meaning of the privacy statement or the contracts would be harder to analyse. However, commercial providers of contract analysis software for due diligence research already exist.[46] The software would have to be adapted to the case of privacy statements and user contracts, but since these are usually similar across merchants this would be an acceptable investment.

To summarise, some data items that result from the web scraping are sufficient for direct analytical purposes. These are:

- The price of items for sale (while mapping any differences between countries);
- Presence of contact details;
- Presence of location based price discrimination;
- Presence of contracts; and
- Presence of privacy statement.

For the following items, a more in-depth analysis is needed:

- The content and its legality of contracts; and
- The content and legality of pre-checked boxes.

This could be analysed through machine learning, but requires significant human "training" for each language that is offered to the machine, which would therefore be a costly exercise.

Use of the data in policymaking

A dataset would contain (at least) the information as proposed in

[44] The practice of labelling (almost) identical goods with different version numbers to allow for geographic discrimination of prices and customer bases.
[45] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/UNECE-ILO_2014_Griffioen_deHaan_Willenborg.pdf
[46] Kira Systems

Table 3 above. Compliance monitoring data would first of all be relevant for **enforcement** purposes. Further analyses of data could answer the following questions, per sector, market segment and per country:

- Where does violation or compliance occur?
- What kind of violation is most often encountered?
- Do these parameters differ for cross border transactions?
- How far do contracts differ over countries and in which clauses?

Answers to these questions would be most relevant for consumer rights policy. Knowing how different market segments comply in different scenarios would enable policymakers to start a well-directed dialogue with these stakeholders to improve compliance or (threaten to) **adapt or introduce legal measures**, which will in turn benefit consumer protection and the single digital market.

5.5 Reflections on challenges and next steps

This use case explored the possibilities of web scraping for monitoring Consumer Price Indexes and sales conditions. Information on these topics is relevant for monetary, economic, and competition policy, consumer rights enforcement and development of the European digital single market.

Web scraping in itself is not innovative; commercial enterprises already offer information on, among other things, consumer prices based on web scraping methods. The main challenges are in **interpretation of information that goes further than prices of products**:

- How to compare identical products from different vendors?
- How to monitor for sales conditions?
- How to simulate buying processes for more complex (configurable) products?

Issues 1 and 2 have been overcome, although significant human effort may be required to aid the machines in analysing the content they harvest. For issue number 3, more **advanced semantics** would be required for algorithms that autonomously navigate pages and know how to **interpret unstructured information** (i.e. text that guides consumers through pages). These efforts would need to be duplicated per country or language that the vendor operates in.

An option to reduce this effort is to take into account (for example) English websites only; many vendors offer their website in a native language and English. This would introduce inclusion issues for those countries where English is not as common.

Note that the proposed method would yield very little privacy issues as the data used is publicly available and the analysis can be discussed with companies before it is published.

The less complex tasks of this case (price information, presence but not content of contracts, pre-ticked boxes) could be implemented on some scale in each Member State by their statistical agencies or consumer protection agencies. Issues with comparability of products within countries over varying vendors would remain but can be overcome by categorisation. Scalability is limited for analysing the content of the scraped data; this requires significant machine learning that is of the same size no matter the amount of content offered. For smaller languages this may form a huge barrier as compared to German, English or French.

A feasible approach would be to start rolling out web-scraping methods over statistical offices for CPI composition and perhaps monitoring the presence of required data (contracts, contact information, etc.). This would yield valuable experience and data sets in each country. In the meantime, semantic analysis would become more powerful and could later on be used to analyse more types of data that has been scraped.

Further reading

Related projects: Billion Prices Project and Clothing price data