Collecting Clothing Data from the Internet

Robert Griffioen, Jan de Haan and Leon Willenborg

Statistics Netherlands
Aim of the paper

Our aim is to describe Statistics Netherlands’ experiences with the collection of prices on clothing items from the Internet for CPI purposes.

We do not compare different approaches to estimating price indexes for clothing nor do we propose a particular method (though we do present tentative index numbers).
Outline

Background
Data collection via web scraping (‘Internet robots’)
Dynamics of items observed
Classification
Tentative price index numbers
Issues and risks
Conclusions
Background

Potential advantage of online data
Efficiency - collecting prices by visiting stores is costly

Other potential advantages
Inclusion of online purchases – currently not well covered
Extending sample sizes (‘big data’)
Higher frequency of price observation

Biggest disadvantage
Data on quantities/expenditure unobservable
Background

Pilot project

Clothing only

Single retailer “S”- has also many physical stores across the Netherlands

Both prices and some characteristics are observed through web scraping

Note: more Internet robots running – largest web store, various housing websites
Web scraping

Information on website of “S”
Item prices
Short item descriptions
Long item descriptions and photos

Web scraping strategy: visit as few web pages and use as little information as possible
will make software robust against website changes
makes software administration simpler
reduces ‘respondent burden’ (web server requests)
Web scraping

Data extracted

**ID**: unique web address of item

**Type**: information on ‘department’ (women, men, children)

**Name**: item name; not necessarily a unique identifier

**Short description**: item description, used for classifying

**Price**: ‘offer’ price of item

IDs are unique identifiers and stable across time

Some data cleaning needed
Web scraping

Daily observation (early in the morning)
Start: 31 January 2012
Data is still being collected

Robot needed fixing 10 times because of website changes;
3 major changes (repair took us almost two days)

Experiences led us to build in standard checks during data collection (rather than checking for errors afterwards)
Dynamics of items observed

Selection of items observed daily (February – July 2012)

<table>
<thead>
<tr>
<th>web ID</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The table and chart show the dynamics of items observed daily from February to July 2012. Each cell indicates the presence or absence of an item on a specific date.
Dynamics

Most items observed during a relatively small number of days
items may be temporarily out of stock, or
they can still be purchased but have been replaced by
similar items, perhaps with a different color

In general:
Set of items observed can be affected by the way in which the
robot navigates through the website, so …. 
…. it is not necessarily true that via web scraping we observe the
entire population of items available
Dynamics

Daily number of items observed
Background

Daily number of ‘births’
Background

Daily number of ‘regular items’ that become ‘sales items’
Dynamics

Daily number of ‘regular items’
Dynamics

Daily number of ‘sales items’

Number of items

Ladies sale  Men sale  Kids sale
Dynamics

Summary of findings

The robot currently does not observe the entire collection on a daily basis.
Number of items observed fluctuates substantially from day to day.
As expected, the number of regular items and that of sales items exhibit a seasonal pattern.
Sales periods as identified on the website correspond to sales periods observed in physical shops.
Classification

Classification system for clothing was developed using automatic coding; 98% correct matches using short item descriptions should apply to any (web) store. Similar breakdown for each ‘department’; matrix structure.

Further breakdown possible by using long item descriptions; lot of work; cannot be read and processed directly. Might be store-specific.
Classification

More detailed classification
Will increase homogeneity, but ....
.... can become unstable over time: many new and disappearing product categories

Collecting additional information (using long item descriptions) still useful
to control for compositional change, or more generally
to control for quality change using hedonics or otherwise
Tentative monthly price indexes

Elementary index numbers calculated at lowest level of existing classification, e.g. for women’s tops, men’s jeans, girls’ dresses, men’s jackets

Ratios of unweighted average prices – daily observation, so items that are observed more frequently within a month have a bigger weight

[simulation with using only data from three Mondays each month were very similar]

Upper level aggregation: fixed annual weights from external source

Indexes at department level exhibit a seasonal pattern
Tentative monthly price indexes

Price indexes for three departments
Tentative monthly price indexes

Elementary aggregates not homogeneous!

Frequency distribution of prices for men’s jackets
Issues and risks

Methodological issues

Potential representativity issue as not all items are observed on a daily basis

- Less important on a monthly basis

Collection in physical stores could be a subset of the collection shown on website (though not for “S”)

Does an average of daily price observations approximate a unit value including both regular and sales prices?

- Impossible to check because scanner data for “S” is unavailable
Issues and risks

Additional information on characteristics required to refine classification or to adjust for compositional/quality change

However, information needed possibly depends on method chosen:

‘Big data’ and hedonic quality adjustment versus small samples and manual item selection / quality adjustment

Prices information on website tends to be ‘correct’ (though some data checking is always useful)
Issues and risks

Potential risks of data collection via web scraping

Changes in structure of the website that affect information needed for navigation or data extraction
Retailer may close its website for the robot, e.g. when web scraping adversely affects the website’s performance
   Good working relationship with the retailer should prevent this
Conclusions

Main advantages of web scraping
Low collection costs
Use of ‘big data’ to circumvent small sample problems
Data quality tends to be good
Some characteristics can be easily observed

Main disadvantages
No weighting information
Characteristics information may be insufficient
Website changes can lead to data problems
Choice of web scraping strategy affects data observed