Using Web Scraped Data to Construct Consumer Price Indices

Nigel Swier
NTTS Conference, 10-12 March 2015, Brussels
Background

• One of 4 “big data” pilots in ONS
• Prices collection manually based
• Difficulties accessing retail scanner data
• Web scraping as a possible alternative (although lacks quantity information)
• More detailed, more frequent and cheaper
• Price scraping for supermarket groceries relatively unexplored
Prototype web scrapers

- 3 supermarkets
- 35 CPI/RPI item categories
- Written in Python (scrapy)
- Daily collection (around 6500 price quotes)
- Item counts monitored daily
Warburtons Toaste Sliced White Bread 800G

Delivering the freshest food to your door. Find out more >

£1.45 (£0.18 / 100g)

Add to basket
## Mapping categories

### Annex A: Consumer Prices Index: Representative Items in 2012

#### 01.1 Food

<table>
<thead>
<tr>
<th>Bread and Cereals</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large white loaves - sliced and unsliced</td>
<td>Bread rolls, Flour, Pasta</td>
</tr>
<tr>
<td>Large wholemeal loaf</td>
<td>Various selected biscuits, Doughnuts</td>
</tr>
<tr>
<td>Garlic bread</td>
<td>Crackers, Frozen pizzas</td>
</tr>
<tr>
<td>Rice</td>
<td>Dehydrated noodles/pasta, Cereal bars</td>
</tr>
<tr>
<td>Breakfast cereals</td>
<td></td>
</tr>
<tr>
<td>Corn based snacks</td>
<td></td>
</tr>
<tr>
<td>Sponge cakes</td>
<td></td>
</tr>
<tr>
<td>Pack of individual cakes</td>
<td></td>
</tr>
<tr>
<td>Fruit pies</td>
<td></td>
</tr>
<tr>
<td>Chocolate wafers</td>
<td></td>
</tr>
<tr>
<td>Hot oat cereal</td>
<td></td>
</tr>
</tbody>
</table>

---

**Tesco.com | My Orders | My Account**

---

**Enter the code XXMNT4 at the online checkout**
## Data Manipulation (Wrangling)

<table>
<thead>
<tr>
<th>ONS Item Category</th>
<th>Item Description</th>
<th>Search Term</th>
<th>Correct Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apples, dessert, per kg</td>
<td>WAITROSE PINK LADY APPLES 4S</td>
<td>'APPLE*'</td>
<td>Yes</td>
</tr>
<tr>
<td>Apples, dessert, per kg</td>
<td>SAINSBURY'S APPLE, KIWI &amp; STRAWBERRY 160G</td>
<td>'APPLE*'</td>
<td>No</td>
</tr>
</tbody>
</table>
Price quote distributions

Whiskey:

Onions:
Experimental Monthly Indices

Random item from each item category with an index day (bootstrapping)
Daily Price Index (Whiskey)
Next Steps

- Experimental high frequency index
- Analysis of mySupermarket data
- Targeted use of web scraped data for temporal sampling project (HICP compliance)
- Machine learning for product categorisation
Acknowledgements

• Rob Breton (Office for National Statistics)

• Rob O’Neill (University of Huddersfield)