Big Data skills in the Social Sciences
Transactional and consumer data

Dr Andy Newing
Lecturer in Retail Geography
School of Geography
University of Leeds
• Lecturer in Retail Geography
• Consumer Data Research Centre (CDRC)
Demand side

Supply Side
– Electronic Point of Sale (EPOS) transactions/sales
– Loyalty cards
– GPS location data
– Mobile phone data – including apps (e.g. exercise)
– Transport (e.g. Oyster Cards)
– Footfall
– Mobilities
– Social media
– Utilities
Welcome to the Consumer Data Research Centre (CDRC)

Vast amounts of UK consumer data are generated each day, providing valuable insight to help organisations operate more efficiently.

It's not just businesses that benefit. Researchers can utilise data to make the UK a better place.

Our aim is to work with organisations to open up their data to trusted researchers so we can provide solutions that drive economic growth and improve our society.
• Increasing recognition among businesses of the value of these data for wider reuse – e.g. ‘social good’.
• Awareness that their data covers only part of the market or a subset of consumers – skills required to link these data to other sources or to model missing data.
• Big data collected by these organisations is often underutilised.
• Innovation and collaboration.
Table A51
Average weekly household expenditure by Output Area Classification (OAC) supergroup
UK, 2014

<table>
<thead>
<tr>
<th>Category (Opening)</th>
<th>Average weekly household expenditure (£)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAC 1</td>
<td></td>
</tr>
<tr>
<td>Super group 1</td>
<td></td>
</tr>
<tr>
<td>Super group 2</td>
<td></td>
</tr>
<tr>
<td>Super group 3</td>
<td></td>
</tr>
<tr>
<td>Super group 4</td>
<td></td>
</tr>
<tr>
<td>Super group 5</td>
<td></td>
</tr>
<tr>
<td>Super group 6</td>
<td></td>
</tr>
<tr>
<td>Super group 7</td>
<td></td>
</tr>
<tr>
<td>Super group 8</td>
<td></td>
</tr>
<tr>
<td>All households</td>
<td></td>
</tr>
</tbody>
</table>

1. Food and non-alcoholic drinks
2. Alcoholic drinks, tobacco and narcotics
3. Clothing and footwear
4. Housing: rent and mortgage
5. Housing: goods and services
6. Health
7. Transport
8. Communication
9. Recreation and culture
10. Education
11. Restaurants and hospitality
12. Miscellaneous goods and services
13. Total expenditure

Total expenditure: £220.80
Average weekly expenditure per person: £58.90

---

*Other expenditure items: 20.00, 220.10, 100.30, 10.90, 90.30, 35.40, 30.00, 21.00, 16.00

---

*Additional notes:
1. This table uses 2011 Output Area Classifications, replacing the 2001 classifications used in previous publications. OAC and OAC classifications were comparable.
2. Excludes mortgage interest payments, council tax and Northern Ireland rates.
General details

Title of output: Living Costs and Food Survey
Abbreviated Title: LCF
Designation: National Statistics
Geographic Coverage: UK
Date of last SQR or QMI*: February 2008
Contact details: SocialSurveys@ons.gov.uk

Executive summary

The Living Costs and Food Survey (LCF) is an annual survey, designed primarily to measure household expenditure on goods and services. It also gathers information about the income of household members. Respondents, including children, keep a detailed diary of expenditure for two weeks. Respondents also record the weights and volumes of food and drink items bought.

A household expenditure survey has been conducted each year in the UK since 1957. From 1957 to March 2001, the Family Expenditure Survey (FES) and National Food Survey (NFS) provided information on household expenditure patterns and food consumption. Both surveys were well-established important sources of information for government and the wider community, charting changes and patterns in the UK’s spending and food consumption since the 1950s. In April 2001 these surveys were combined to form the Expenditure and Food Survey (EFS). The EFS was renamed as the Living Costs and Food (LCF) survey in January
## Research Opinion Poll

Your views count. Manufacturers and service providers depend on your input to help them offer the right products in the right way at the right price. Please take a few minutes to complete this survey. Just ignore any questions you would prefer not to answer. Thank you for helping us.

Your answers will be protected by the Data Protection Act and controlled by Ascot Ltd who will use this information for managing products & services, marketing research, updating and developing databases. Ascot Ltd may pass the information given to other reputable organisations, who will use it for their own market research and analysis purposes. They may send you, by mail or other media, details of their products and services. When answering questions for your partner, please check first that they are happy for you to do so. If you would rather that your details and/or your partner were not passed to these organisations than please tick the appropriate box(es) you / your partner.

Questions have been included in the survey on behalf of The Telegraph Group PLC; Newspapers Q1-5, Yorkshire Forward (Regional Development Agency) various questions about Yorkshire, and You and Your Family Q3, HUGS Matters Q3-6, Fenda Matters Q7, Derby Family Health Q2.

Thank you very much for your help.

PS: Don't forget the extra Digital Radio draw, to enter, simply return your survey within 7 days!

### Groceries

<table>
<thead>
<tr>
<th>1. Where do you shop for groceries?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ASDA/Netto</td>
</tr>
<tr>
<td>2. Morris</td>
</tr>
<tr>
<td>3. Co-op</td>
</tr>
<tr>
<td>4. Iceland</td>
</tr>
<tr>
<td>5. Tesco</td>
</tr>
<tr>
<td>6. Waitrose</td>
</tr>
<tr>
<td>7. Matalan/Spencer</td>
</tr>
<tr>
<td>8. Garages Forecourt</td>
</tr>
</tbody>
</table>

2. From the list above please write the number of your MAIN supermarket: [ ]

### Newspapers

<table>
<thead>
<tr>
<th>1. What newspapers do your family read?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Daily Paper</td>
</tr>
<tr>
<td>2. Sunday Paper</td>
</tr>
</tbody>
</table>

2. Write in the number of your MAIN: [ ]

### Home

1. In which month do you renew the following insurance policies? [ ]

2. How many adults live in your home? [ ]

3. Is your home: [ ]
   - Flat
   - Bungalow

4. How many bedrooms do you have? [ ]

5. Do you own your home: [ ]
   - Yes
   - No

6. When did you move to this address? [ ]

7. Please tell us the number and postcode of your previous address: [ ]

8. Are you planning to move to the next: [ ]
   - 0-3 months
   - 3-6 months
   - 6-12 months
   - No
• “looks like the man drawer of stuff” (David Kemp, HP, Big Data Week Belfast 2015).
  – unstructured and undocumented
  – held by various teams
  – in various formats
  – within various systems
  – in numerous locations
  – some may be ‘digital exhaust’
  – considerable manual input to organise, extract and explain
• Data scarce to data rich – many social science methods traditionally designed for small datasets with well known properties.
• Spreadsheets to large databases
• How will major surveys (e.g. ONS LCF) keep pace?
• Social science has started developing capabilities to store and analyse transactional data.
• Much of these data are geographical in nature so spatial analytics increasingly important (Graham and Shelton, 2013).
• Segmentation – how do we reduce unwieldly data sets to manageable specific segments to analyse?
• Data mining
  – find unexpected relationships
  – Summarise data in meaningful ways
• Time series analysis and near real time insight
• Visualisation – communicating insight
Sharwood's Green Label Mango Chutney

A deliciously tangy chutney, made with juicy tropical mangoes

Pair up with Sharwood's Poppadoms

MILD
“it is a customer problem, not a technological one”

Mick Yates, formerly Dunnhumby, International Big Data Symposium 2015

• Is our analysis data driven or problem driven?
• How do we exploit the longitudinal nature of these data?
• Do we want to understand at the aggregate or individual level?
• How do we link those observed behaviours to known individuals and/or known spatial origins?
• Collaborative research with commercial sector
  – Data access
  – Timeliness vs. quality
  – Publication potential
• Skills for dealing with the commercial sector – the academic way of doing things doesn’t work!
• Commercial awareness – understand the context
• Huge overlap with marketing – not always the traditional the domain of social sciences
• Often requires new interdisciplinary collaborations – marketing, computer science
• CDRC Training and capacity building

Autumn Training Programme Announced

We have just announced the Autumn Training Programme, which includes courses in R, QGIS and Spatial Analytics.

Find out More

• Or simply have a go ....
Real-world data to put your theory into practice

**Breakfast at the Frat**
Using sales and promotion info on pretzels, frozen pizza, boxed cereal, and mouthwash gathered from a sample of stores over 156 weeks, this dataset facilitates time series analyses in areas including promotional effectiveness and price sensitivity.

**Let's Get Sort-of-Real**
The data's not real, but there sure is a lot of it. With 300M+ at-lill transactions over 117 weeks, we've replicated the typical patterns found in real in-store sales data to help curious data scientists test their techniques and algorithms in a very real way.  

*Multiple download options available. Click to view.*

**Carbo-Loading**
Carbo-Loading contains household level transactions over a period of two years from four categories: Pasta, Pasta Sauce, Syrup, and Pancake Mix. These categories were chosen so that interactions between the categories can be detected and studied.

**The Complete Journey**
This dataset contains household level transactions over two years from a group of 2,500 households who are frequent shoppers at a retailer. It contains all of each household’s purchases, not just those from a limited number of categories.

https://www.dunnhumby.com/sourcefiles
Abstracts

Alexander Bland (2015) Identifying fuel and poverty characteristics through e.on consumer records and geo-demographic segmentation data (Sponsor: e.on)

Thomas Berry (2015) Exploring the utility of the 2011 Workplace Statistics to help The Co-op better understand transient new store locations, worker flows and worker demographics (Sponsor: The co-operative food!) – ESRC Prize Winner


Yiqao Huang (2015) Shopping centre’s turnover estimation using microsimulation: an exploratory research in Inverness (Sponsor: CACI)

Radoslaw Kowalski (2015) Topic modelling online customer reviews (Sponsor: Argos)

Karol Lugomer (2015) Relations between structure and performance of retail centres in England and Wales and demographics of their catchment areas (Sponsor: LDC) – ESRC Prize Winner

Anastasia Ushakova (2015) Can we identify vulnerable energy customers from smart meter data? (Sponsor: British Gas) – ESRC Prize Winner

Clemens Zaucher (2015) Identifying the main drivers of customer satisfaction and dissatisfaction by mining customer verbatim feedback (Sponsor: easyJet) – ESRC Prize Winner

https://www.cdrac.ac.uk/retail-masters/

http://geog.leeds.ac.uk/people/a.newing
Big Data skills in the Social Sciences
Transactional and consumer data

Dr Andy Newing
Lecturer in Retail Geography
School of Geography
University of Leeds