Lecture 1
Introduction to Semantic Mapping, Both Metric and Ultrametric.
(Using the following Analysis of Questionnaires: Geometric Data Analysis for Interpretation and Communication of Results)

Web address:
http://www.multiresolutions.com/Courses/
ESSCaSS2015Course
User name: Courses
Password: Wel1Come [Note: letters: Wel, number: one, letters: Come]

Fionn Murtagh

Themes and Objectives (Lecture 1)

- A broad and general introduction to semantic mapping through Correspondence Analysis.
- The focus here is on Multiple Correspondence Analysis, to handle both the data and contextual information. A link will be drawn to noted social scientist, Pierre Bourdieu.
- Also included will be clustering.
- So metric and ultrametric mapping of the data is at issue.
- The case study used will be from a survey questionnaire.
- Final part: short discussion of the role of Big Data.


* In The Craft of Sociology, Bourdieu wrote: "I use Correspondence Analysis very much, because I think that it is essentially a relational procedure whose philosophy fully encompasses what my view constitutes social reality. It is a procedure that "thinks" in relations, so I try to do it with the concept of fields."

Later here, and in the accompanying case study, I will follow closely this little book:
Brigitte Le Roux and Henry Rouanet, Multiple Correspondence Analysis, SAGE. Quantitative Applications in the Social Sciences Series, 2010 (115 pp.)

Left: Brigitte Le Roux; right: Henry Rouanet (1931-2008)
Le Roux and Rouanet’s “Taste Example” data (the start of the data follows).

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>TV</th>
<th>Film</th>
<th>Art</th>
<th>Eat</th>
<th>Gender</th>
<th>Age</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Action</td>
<td>Landscape</td>
<td>StakeHouse</td>
<td>Women</td>
<td>55-64</td>
<td>£20-29</td>
</tr>
<tr>
<td>2</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Horror</td>
<td>Fall</td>
<td>Indian</td>
<td>Real</td>
<td>Women</td>
<td>45-54</td>
</tr>
<tr>
<td>3</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Horror</td>
<td>Steak</td>
<td>Indian</td>
<td>Real</td>
<td>Women</td>
<td>55-64</td>
</tr>
<tr>
<td>4</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Action</td>
<td>Landscape</td>
<td>Steak</td>
<td>House</td>
<td>Women</td>
<td>65+</td>
</tr>
<tr>
<td>5</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Horror</td>
<td>United</td>
<td>States</td>
<td>Women</td>
<td>65+</td>
<td>£10-19</td>
</tr>
<tr>
<td>6</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Horror</td>
<td>Fall</td>
<td>Indian</td>
<td>Real</td>
<td>Women</td>
<td>55-64</td>
</tr>
<tr>
<td>7</td>
<td>Active</td>
<td>TV=Soap</td>
<td>Action</td>
<td>Landscape</td>
<td>StakeHouse</td>
<td>Women</td>
<td>25-34</td>
<td>£10-19</td>
</tr>
</tbody>
</table>

- The data, inspired by Bourdieu’s work on taste, are from an ESRC project, “Cultural Capital and Social Exclusion”, collected in the UK in 2003-2004.
- Likes of respondents in regard to: (1) television programmes; (2) films; (3) paintings and art; (4) restaurants and eating out.
- Number of response modalities of these: TV - 8; Film - 8; Art - 7; Eat - 6. In addition there were the identification variables: Gender, Age (6 categories), and Income (7 categories, including unknown).
- 1215 individuals responded to all questions. Also there were an additional 37 respondents.

Oranges data from F. Husson, S. Lê and J. Pagès,
Exploratory Multivariate Analysis by Example Using R,
Chapman & Hall, 2011.
Taking 6 oranges crossed by 15 properties.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
<th>Value 5</th>
<th>Value 6</th>
<th>Value 7</th>
<th>Value 8</th>
<th>Value 9</th>
<th>Value 10</th>
<th>Value 11</th>
<th>Value 12</th>
<th>Value 13</th>
<th>Value 14</th>
<th>Value 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taste</td>
<td>7.82</td>
<td>2.62</td>
<td>1.86</td>
<td>3.42</td>
<td>3.78</td>
<td>2.67</td>
<td>3.20</td>
<td>2.70</td>
<td>3.32</td>
<td>2.52</td>
<td>3.32</td>
<td>3.12</td>
<td>2.70</td>
<td>2.70</td>
<td>3.12</td>
</tr>
<tr>
<td>Texture</td>
<td>2.78</td>
<td>2.62</td>
<td>1.86</td>
<td>3.42</td>
<td>3.78</td>
<td>2.67</td>
<td>3.20</td>
<td>2.70</td>
<td>3.32</td>
<td>2.52</td>
<td>3.32</td>
<td>3.12</td>
<td>2.70</td>
<td>2.70</td>
<td>3.12</td>
</tr>
<tr>
<td>Flavor</td>
<td>2.63</td>
<td>1.80</td>
<td>4.01</td>
<td>3.45</td>
<td>2.67</td>
<td>3.20</td>
<td>2.70</td>
<td>3.32</td>
<td>2.52</td>
<td>3.32</td>
<td>3.12</td>
<td>2.70</td>
<td>2.70</td>
<td>3.12</td>
<td>2.70</td>
</tr>
<tr>
<td>Aroma</td>
<td>3.76</td>
<td>3.17</td>
<td>1.86</td>
<td>3.42</td>
<td>3.78</td>
<td>2.67</td>
<td>3.20</td>
<td>2.70</td>
<td>3.32</td>
<td>2.52</td>
<td>3.32</td>
<td>3.12</td>
<td>2.70</td>
<td>2.70</td>
<td>3.12</td>
</tr>
<tr>
<td>Appearance</td>
<td>3.20</td>
<td>3.05</td>
<td>2.67</td>
<td>3.42</td>
<td>3.78</td>
<td>2.67</td>
<td>3.20</td>
<td>2.70</td>
<td>3.32</td>
<td>2.52</td>
<td>3.32</td>
<td>3.12</td>
<td>2.70</td>
<td>2.70</td>
<td>3.12</td>
</tr>
<tr>
<td>Juiciness</td>
<td>3.07</td>
<td>2.73</td>
<td>3.64</td>
<td>3.24</td>
<td>3.05</td>
<td>3.31</td>
<td>3.05</td>
<td>2.80</td>
<td>3.31</td>
<td>3.05</td>
<td>2.80</td>
<td>3.31</td>
<td>3.05</td>
<td>2.80</td>
<td>3.31</td>
</tr>
</tbody>
</table>

**Triangular Inequality Holds for Metrics**

**Example: Euclidean or “as the crow flies” distance**

\[
d(x, z) \leq d(x, y) + d(y, z)
\]

For \( x = (2, 5) \) and \( z = (5, 3) \) Euclidean distance is
\[
d(x, z) = \sqrt{(2-5)^2 + (5-3)^2} = \sqrt{9 + 4} = 3.6
\]

**Strong Triangular Inequality, or Ultrametric Inequality, Holds for Tree Distances**

\[
d(x, z) \leq \max\{d(x, y), d(y, z)\}
\]

\[
d(x, z) = 3.5
\]

\[
d(x, y) = 3.5
\]

\[
d(y, z) = 1.0
\]

**Closest Common Ancestor Distance is an Ultrametric**

**Line Drawing**

J.P. Bensaïd & Collaborateurs,
1. Analyse des Données, Tome 2,
L’Analyse des Correspondances,

See depiction of axial moments of rotational inertia in lower right.

Christian Huygens, 1629-1695
who developed moments of inertia in the context of work on oscillations of pendulum systems.
Geometric Data Analysis

- For two-way frequency tables, CA is applicable.
- For quantitative data, PCA is applicable.
- For categorical variables, MCA (multiple correspondence analysis) is useful.
- Categories: modalities of (nominal) variables.
- Individual observations, observables, cases, items, time periods, and so on.
- Variables, attributes, qualities, properties, and so on.

Three key ideas of MCA especially from the Geometric Data Analysis viewpoint (Le Roux and Rouanet, 2010):

1. Geometric modelling: one cloud of points from the individual to the categorical variables.
2. Formal approach: “As in all data analysis, in good mathematics, it is simply searching for relationships, all the science for the art of it is just finding the right metric to diagonalize.”
3. Description first: geometric modelling before probabilistic modelling, in the spirit of inductive philosophy. “The model should follow the data, not the reverse.”

Interpretation

- Projections onto factors by decreasing order of importance.
- Also contributions and correlations to factors, to judge importance and relevance.
- The eigenvalues - moments of inertia about the axes - in relative percentage terms inform us about the extent of approximation to the data of a limited set of factors (associated with the 1st, 2nd, 3rd, ... eigenvalues).
- It may be feasible to characterize or label the factors in terms of basic variables.
- Joint display of rows, columns - individuals and their attributes. Thus, mental and/or complementary understanding.
- Subgroups and clusters. We may well want to carry out a complementary clustering of the individuals and/or their properties. That can refine selection of, and further exploration of, clusters or it is often a hierarchical clustering, furnishing a set of partitions of the individuals and/or attributes. An aim with such a hierarchical clustering may be the generating of a typology.
- Subsequent issues in the domain concerned: reduct action of the data to study, maybe select the principal or supplementary data in a different way, or perhaps collect new data.

Pierre Bourdieu’s use of CA, MCA

- Place of the German edition of Le Mots de la Société, 1981.
- “From Correspondence Analysis very much, because I think that it is essentially a relational procedure whose philosophy fully expresses what in my view constitutes social reality. It is a procedure that is indiscussed, as I try to do it with the concept of field.”


- “Those who know the principles of multiple correspondence analysis will grasp the affinities between this method of mathematical analysis and the thinking in terms of fields.”

- Used in Bourdieu’s other work, including: Nova Accedemia, Héberc d’École, Structures Sociales d’Économie, etc.

Hierarchical clustering

- For n observations, we have - by our agglomerative hierarchical clustering algorithm - n-1 agglomerations (i.e., merging) steps.
- Each level of the hierarchy (i.e., make a horizontal cut) furnishes a partition into clusters of the observation set.

- The factor space (output from CA) is Euclidean and equidistanted, for the individuals and attributes. This is suitable input to agglomerative hierarchical clustering.
- Given the inertia criterion used in CA, an appropriate hierarchical agglomerative criterion is the minimum variance (in merging two clusters at each successive agglomerations).

Dual spaces, transition formulae. Importance includes (1) interpretation in simultaneous display; (2) computational - we will not dwell on this here; (3) supplementary elements.
Complete disjunctive form - one response category is answered; for a question, the total of response category values = 1.

I × Q table. Disjunctive I × K table.
Q = 3 questions and K = 2 + 2 + 3 = 7 categories.

<table>
<thead>
<tr>
<th>I</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>a1</td>
<td>b1</td>
<td>c1</td>
</tr>
<tr>
<td>i'</td>
<td>a2</td>
<td>b2</td>
<td>c2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>K</th>
<th>a1</th>
<th>a2</th>
<th>b1</th>
<th>b2</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>i'</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\( n = \text{Q}_2 \)

---

Case Study: Outputs and R Code
Based on Bourdieu's Taste

The accompanying R code reproduces Everything in this book.

Le Roux and Rouanet's "Taste Example" data (the start of the data follows).

- The data, inspired by Bourdieu's work on taste, are from an ESRC project, "Cultural Capital and Social Relations", collected in the UK in 2003-2004.
- The likes of respondents in regard to (1) television programmes; (2) films; (3) paintings and art; (4) restaurants and eating out.
- Number of response modalities of these: TV - 8; Film - 8; Art - 7; Eat - 6. In addition there were the identification variables: Gender; Age (6 categories); and Income (7 categories, including unknown).
- 1215 individuals responded to 44 questions. Also there were an additional 37 respondents.

- This most comprehensive survey (118 citations) sets out new contemporary issues of sampling and population distribution estimation. An enormously important take-home message is the following.
- "There is the potential for big data to evaluate or calibrate survey findings ... to help to validate cohort studies". Examples are discussed of "how data ... tracks well with the official", far larger, repository or holdings.

Keiding and Louis (contd.)

- It is well pointed out how one case study discussed "shows the value of using 'big data' to conduct research on surveys (as distinct from survey research)".
- Limitations though are clear: "Although randomization in some form is very beneficial, it is by no means a panacea. Trial participants are commonly very different from the external ... pool, in part because of self-selection, ...."

Keiding and Louis (contd.)

- This is due to, "One type of selection bias is self-selection (which is our focus)".
- Important points towards addressing these contemporary issues include the following.
- "When informing policy, inference to identified reference populations is key"
- This is part of the bridge which is needed, between data analytics technology and deployment of outcomes.
"In all situations, modelling is needed to accommodate non-response, dropouts and other forms of missing data."

While "Representativity should be avoided", here is an essential way to address in a fundamental way, what we need to address:

"Assessment of external validity, i.e. generalization to the population from which the study subjects originated or to other populations, will in principle proceed via formulation of abstract laws of nature similar to physical laws."

The following 50 page report makes interesting reading.

"The classic statistical paradigm was one in which researchers formulated a hypothesis, identified a population frame, designed a survey and a sampling technique and then analyzed the results .... The new paradigm means it is now possible to digitally capture, semantically reconcile, aggregate, and correlate data."

From AAPOR Report: What is Big Data?

... large data sources have been mined to enable insights about economic and social systems, which previously relied on methods such as surveys, experiments, and ethnographies to drive conclusions and predictions ...

Example 1: Online prices. ... prices collected daily from hundreds of online retailers around the world to conduct economic research. One statistical product is the estimation of inflation in the US. Changes in inflation trends can be observed sooner in PriceStats than in the monthly Consumer Price Index (CPI).

Example 2: Traffic and infrastructure. Big Data can be used to monitor traffic or to identify infrastructural problems .... real-time information .... to fix problems and plan long term investments...

Example 3: Social media messages. ... Twitter ... generating early predictions of initial claims for unemployment insurance. The predictions are based on a factor analysis of social media messages mentioning job loss and related outcomes
"Because the GBCS is not a random-sample or representative survey" other ways can and are being found to draw great benefit.

Another different study on open, free text questionnaires (Züll and Scholz, 2011) notes selection bias, but also: "However, the reasonable use of data always depends on the focus of analyses. So, if the bias is taken into account, then group-specific analyses of open-ended questions data seem appropriate".

The bridge between the data that is analyzed, and the calibrating Big Data, is well addressed by the geometry and topology of data. Those form the link between sampled data and the greater cosmos.

- Bourdieu's concept of field is a prime exemplar.

- Consider, as noted by Lebaron (2009), how Bourdieu's work, involves "putting his thinking in mathematical terms", and that it "led him to a conscious and systematic move toward a geometric frame-model".

- This is a multidimensional, "structural vision".

"... the GBCS [Great British Class Survey] data have three important limitations. First, the GBCS was a self-selecting web-based survey, ... This means it is not possible to make formal inferences. ... the nationally representative nature of the Labour Force Survey (LFS) along with its detailed and accurate measures ... facilitates a much more in-depth investigation ..."


- Bourdieu's analytics "amounted to the global [hence Big Data] effects of a complex structure of interrelationships, which is not reducible to the combination of the multiple [...] effects of independent variables".

- The concept of field, here, uses Geometric Data Analysis that is core to the integrated data and methodology approach used in the Correspondence Analysis platform.
• An approach to drawing benefit from Big Data is precisely as described in Keiding and Louis. Their noting of the need for the "formulation of abstract laws" that bridge sampled data and calibrating Big Data is addressed,

• for the data analyst and for the application specialist,

• as geometric and topological.

Finally, content of a slide from a presentation by Frédéric Lebaron:

**Investigating Fields**

• Geometric modelling of data as instrument of synthesis and representation of the fields.
• In the middle of the 1960s, formulation of the concept of « field » (first article 1966).
• The « geometric data modelling » as a way to combine statistical analysis and the notion of field : « Those who know the principles of MCA will grasp the affinities between this method of mathematical analysis and the thinking in terms of field » (Bourdieu, 2001, p. 70).