

Text as Data: Basic Approaches to Computer-Assisted Content Analysis

Dr Ekaterina Kolpinskaya
Associate Lecturer & Researcher, Q-Step Centre

E.Kolpinskaya@exeter.ac.uk

@DrKolpinskaya

Outline

- What is content analysis?
- Overview of content analysis software
- Analysing texts (and more) using NVivo
- Analysing texts using Yoshikoder
- Triangulating texts with quantitative data

Quantitative Content Analysis (QCA)

‘a research technique for the objective, systematic, and quantitative description of the manifest of communication’ (Berelson, 1952)

It uses a text rather than reads it.

It differs from more qualitative approaches, because

- Involves large-scale analysis of many texts, rather than close readings of few texts
- Requires interpretation to be operationalized
- Does not explicitly concern itself with the social or cultural predispositions of analysts

Three major approaches

- **Supervised-learning** (compares word frequencies to manually set categories)
- **Scaling** (treats words as data in statistical model)
- **Pattern-matching** (automated classic content analysis with post-hoc statistical analysis)

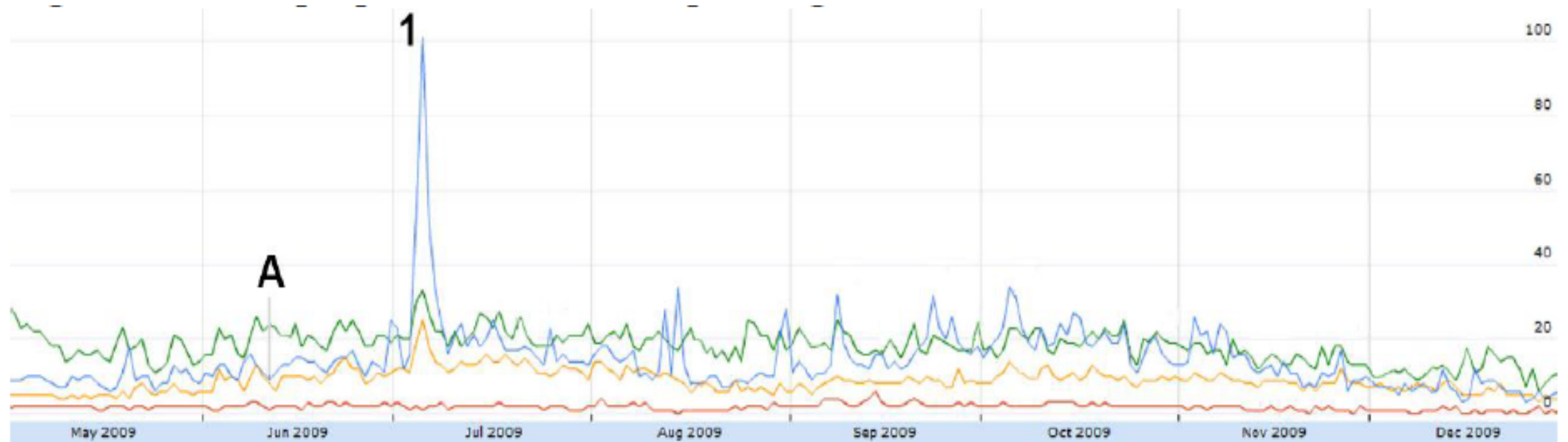
What can we use QCA for?

“best for linguistically constituted facts, attributions, social relationships (or at least their public face), public behaviours, institutional realities... because language is “repetitive, routine, public, [and] institutionalized” (Krippendorf, 2004)

- **Studies of political behaviour**
- **Studies of elites (e.g., parties and politicians)**

This includes ideal point estimation, salience, framing, valence/sentiment analysis, conflict scores, etc. E.g., <http://www.sentiment140.com/>

People expressing flu-relating concerns on Twitter

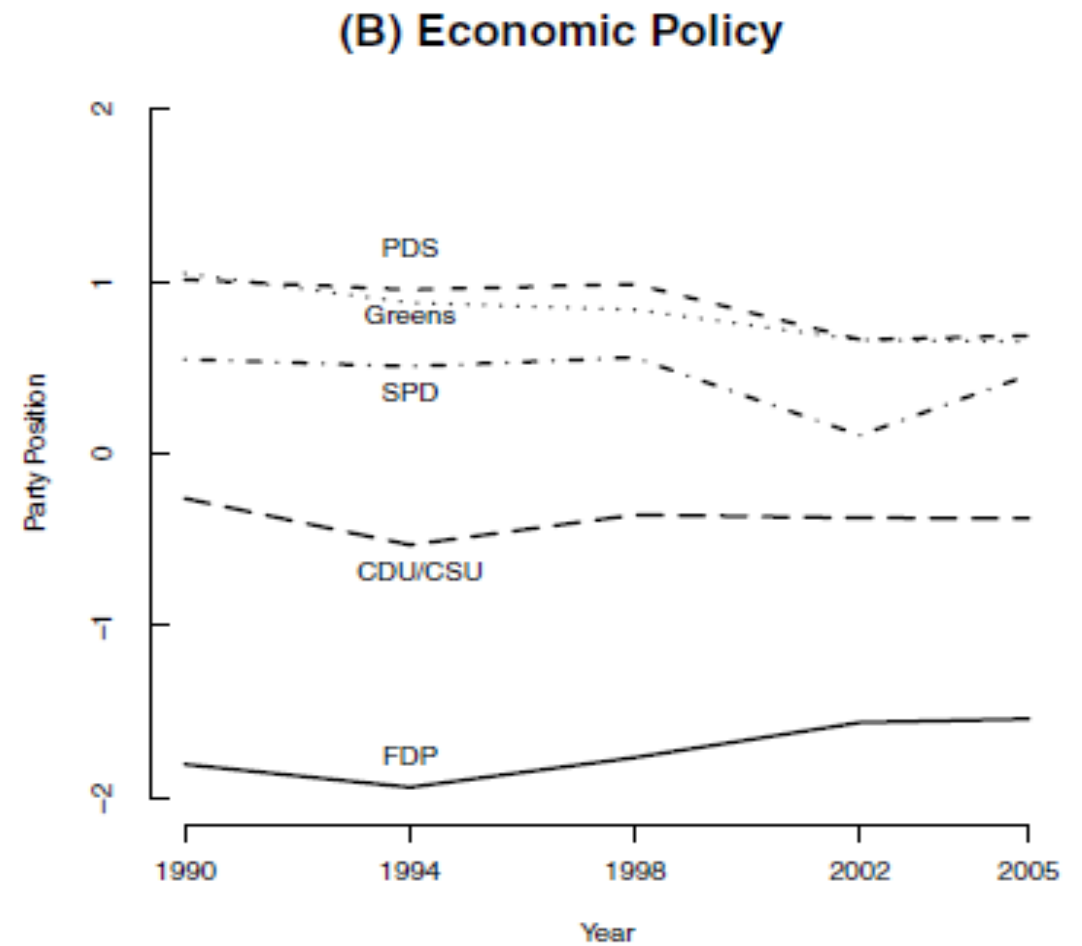
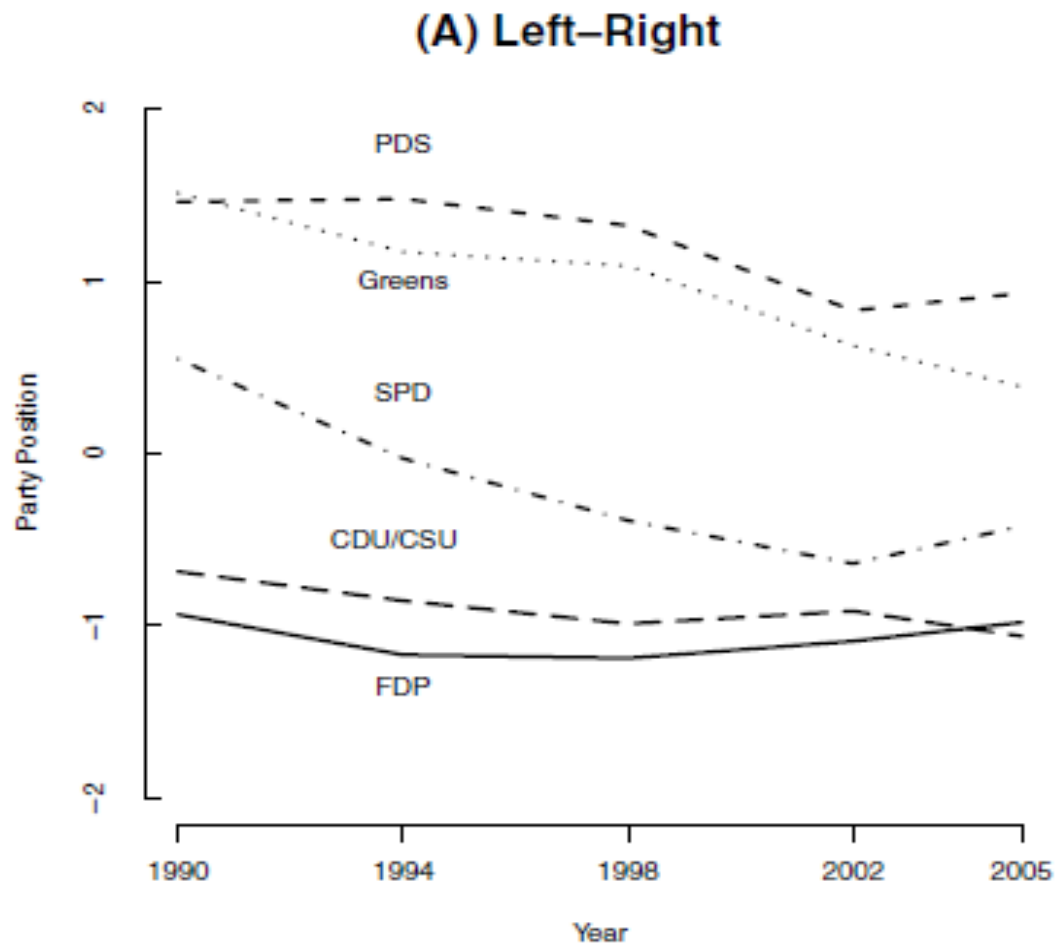


A = June 11: WHO pandemic level 6 announcement

1 = July 5: Harry Potter actor Rupert Grint has H1N1

Subconcepts: Blue = concern for others; Red = concern for self; Yellow = concerned emoticons; Green = general concern.

Estimated party positions in Germany, 1990-2005



Content analysis: Strengths & limitations



Transparency
Replicability
Reliability
Deals with large amount of text
Well suited for political texts

Cannot replicate an in-depth
discourse analysis
Looks at manifest content only
Complexity of linguistic
structures usage
Comparing across time and
languages
Construct validity

Steps of Quantitative Content Analysis:

1. Select and collect the data
2. Define the unit of coding and analysis
3. Choose the analytic technique and develop a coding scheme (dictionary)
4. Pilot test the coding scheme on samples of text & assess its consistency
5. Code data
6. Run the analysis
7. Draw conclusions from the coded and analysed data
8. Report methods and findings

General sampling advice

1. Control what you can
2. Randomize what you cannot

Regularly ask yourself:

- What makes me analyse these texts, not other ones? Can I justify this?
- How many texts is 'enough'?

Textual data AKA corpus

Corpus is a collection of texts, complete, thematically unified to serve research as a result of the systematic approach to collecting data.
Practically, it is the data/dataset.

3 principles:

- Relevance to RQs – should have one thematic focus
- Homogeneity – different types of materials should not be mixed with images. They have to be organised separately
- Synchronicity – respect the cycle of stability and change

Define coding units and units of analysis

A code is 'one that captures the qualitative richness of the phenomenon'
(Boyartis, 1998)

What can be coded?

- Themes/topics
- Types of arguments (parliamentary documents)
- Values
- Attitudes
- Emotions
- Group dynamic (focus groups, TV debates)
- Presentation/Representation
- Colour, posture, character (paintings, images)
- Mixed: topics + values + rhetorical approach (presidential debates)

Choose the analytic technique

- Word scoring - uses the frequencies of single words to estimate the position of a text on a single dimension (e.g., left-right scale, reference text) (e.g., http://www.tcd.ie/Political_Science/wordscores/index.html)
- Pattern matching - pre-defined dictionary approach useful for estimating salience, whereby a content dictionary is designed using words and phrases as indicators of concepts (e.g., Saalfeld and Bischof 2013)
- Sentiment analysis – how people feel about products/politicians, etc. (e.g., Hopkins and King 2010, see <http://gking.harvard.edu/readme>)

Overview of content analysis software

- NVivo
- QDA Miner/Simstat
- Yoshikoder

And many more on <http://www.kdnuggets.com/software/text.html> & <http://www.content-analysis.de/software/quantitative-analysis>

Practice (1): KWIC using NVivo

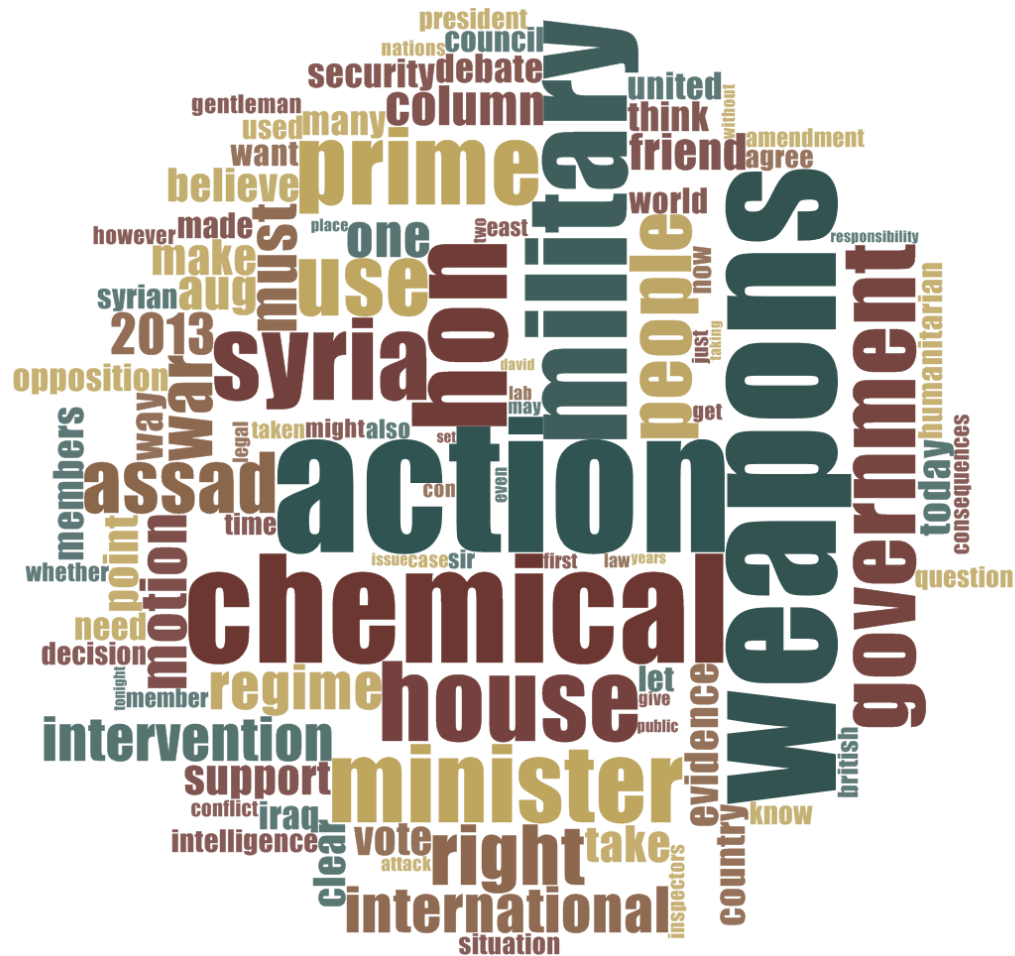
- House of Commons debate on intervention in Syria on August 29, 2013
- House of Commons debate on intervention in Syria on December 2, 2015

What are the themes?

Compare and contrast the most/least featured themes

Plot the dynamic of the debate

Lets compare – 2013 vs. 2015



NViVo 101

1. Create a new project (tick 'Write user actions as an action log')
2. Import the documents of your interest
3. Create the coding framework = nodes + references
4. Start coding manually, then you may switch to automatic coding
5. Use Explore and Query tabs to examine the content of the sources and the relationships between the sources and/or nodes
6. Export the Query results into Excel and/or save them as a part of project

Practice (2): Relational QCA using Yoshikoder

EDMs tabled during the parliamentary debate on the intervention in Iraq between 25 November 2002 and 18 March 2003 – **728 EDMs**.

Each EDM is coded by number and date and saved as separate .txt file

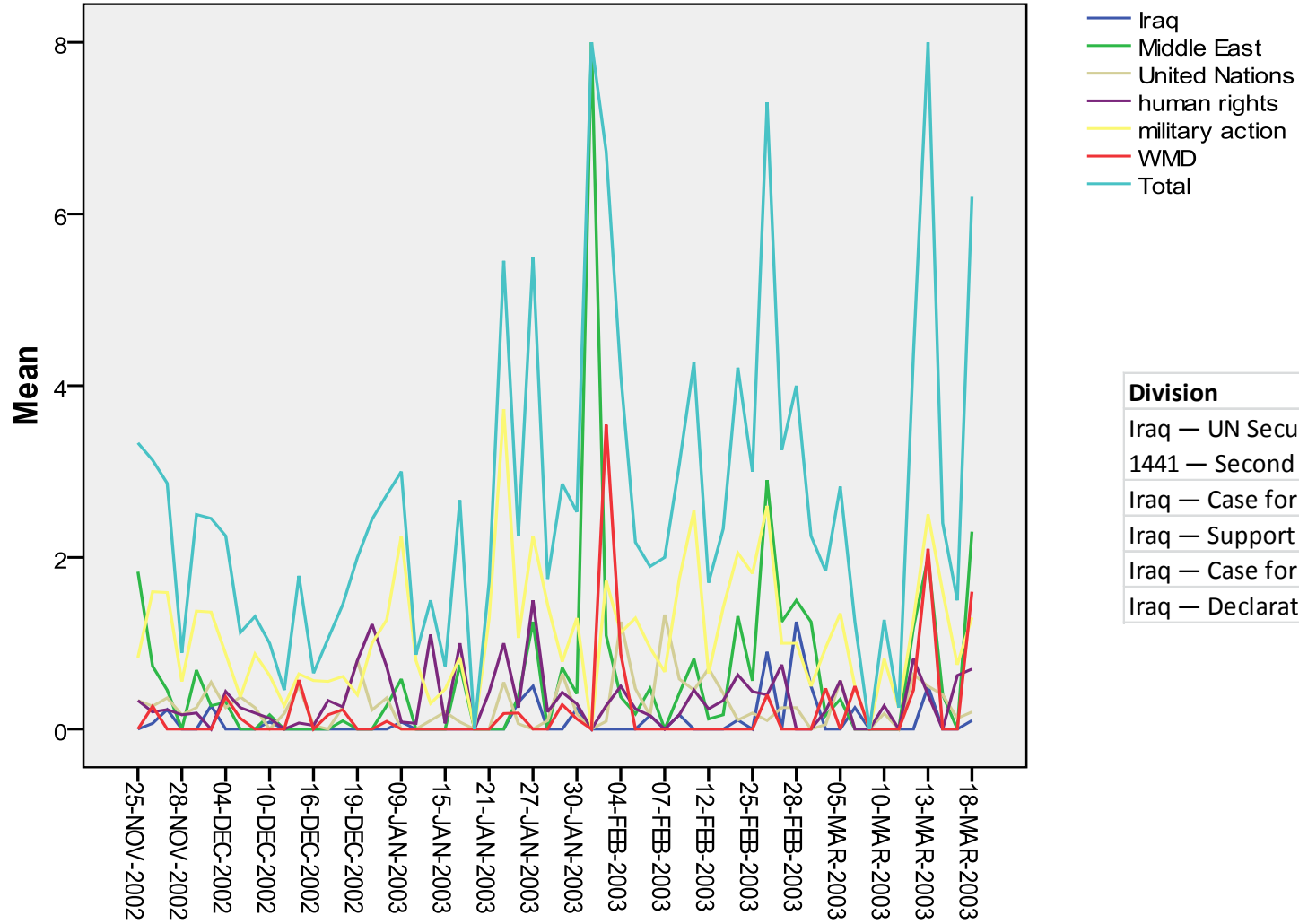
Six concepts relating to Iraq are included in a dictionary:

- 'Iraq'
- 'Middle East'
- 'United Nations'
- 'Human Rights'
- 'Military Action'
- 'Weapons of Mass Destruction (WMD)'

Yoshikoder 101

- Collect and code the data units as txt files
- Design a dictionary
- Pilot test and refine the dictionary on samples of data
- Run the analysis (Report → Apply Dictionary)
- Export result as an Excel file

References to Iraq-related issues by date



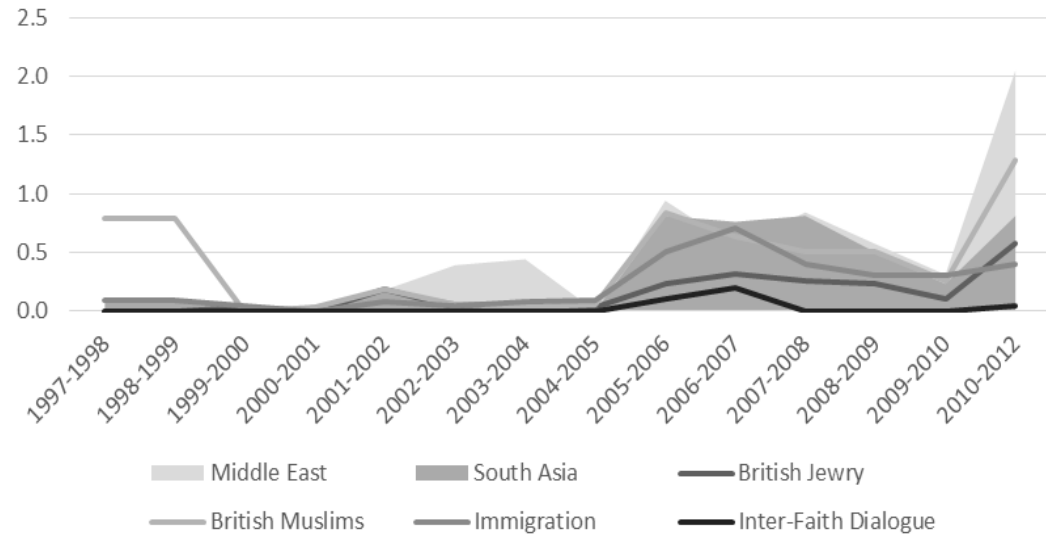
Division	Date	Result
Iraq — UN Security Council Resolution 1441 — Second resolution necessary	25 Nov 2002	rejected
Iraq — Case for war is unproven	26 Feb 2003	rejected
Iraq — Support for the Government	26 Feb 2003	accepted
Iraq — Case for war not established	18 Mar 2003	rejected
Iraq — Declaration of War	18 Mar 2003	accepted

Next step: Add some context!

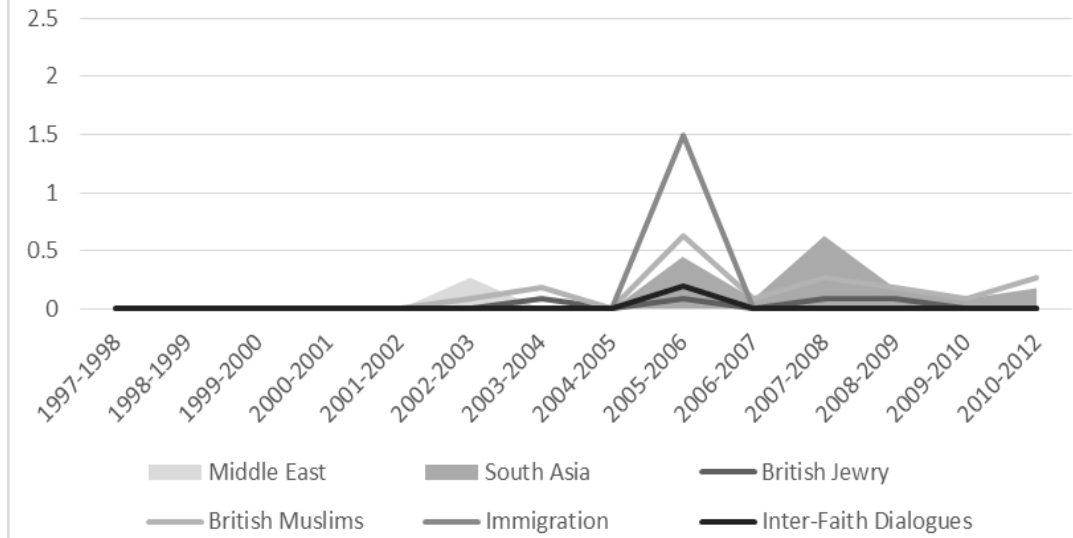
- Party of MPs
- Gender of MPs
- Opposition/Government
- Frontbencher/backbencher

And have a stab at the data analysis in any statistical package!

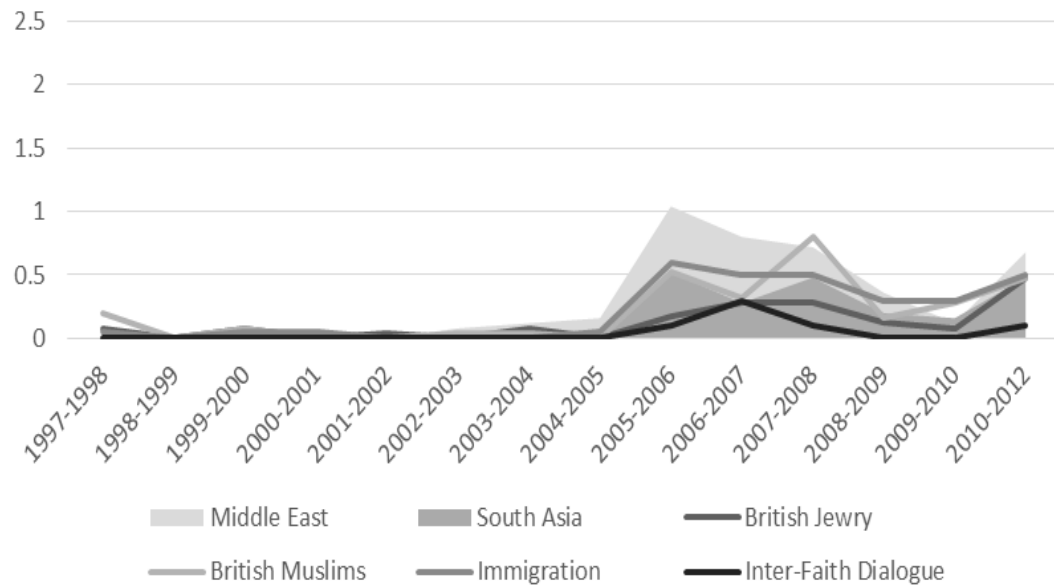
MPs from a Jewish background



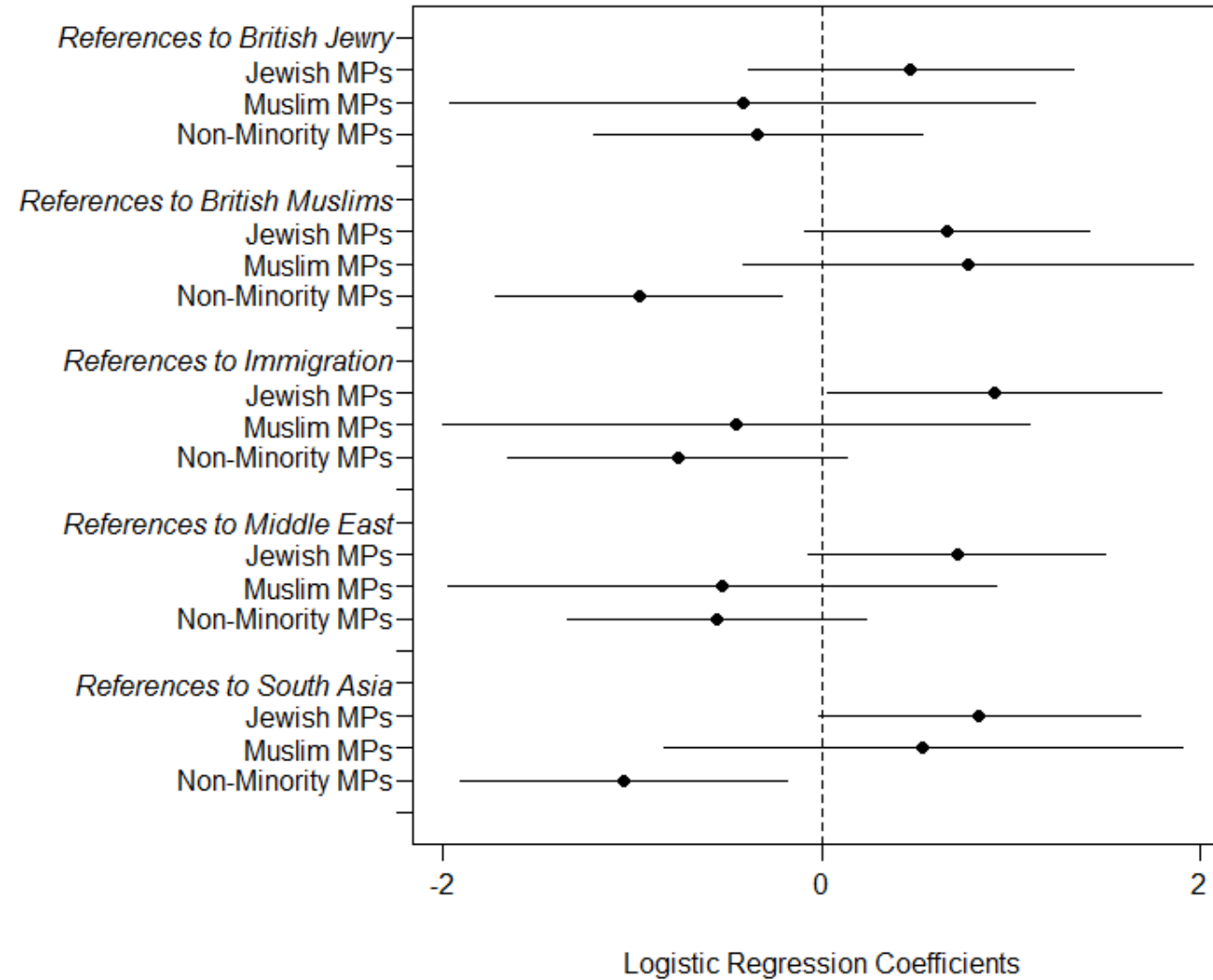
MPs from a Muslim background



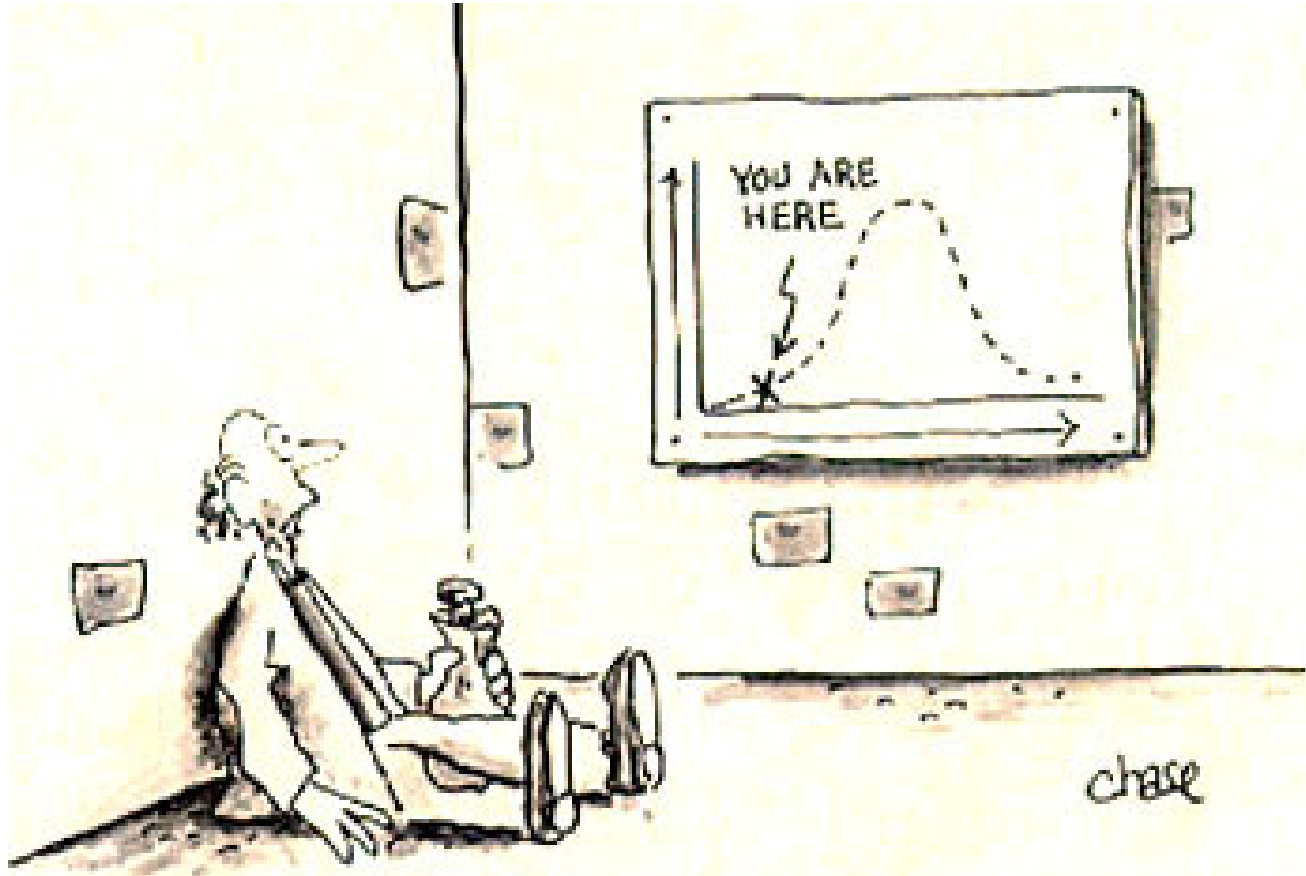
MPs from a non-minority background



Effect from religious background on the likelihood of referring to minority issues in EDMs



Stats Help Desk



Monday, 2:30-4:30pm in XFI Seminar Room C

Make an appointment at <http://statshelpdesk.eventbrite.co.uk>