Communicator

Special Supplement: Science Communication

Autumn 2016

Science is ageless

Outreach and the Science Museum

At-Bristol dissections case study

Statistics and surnames

Q&A with a science illustrator
The UK held a referendum on 23 June to provide an answer to the question “Should the United Kingdom remain a member of the European Union or leave the European Union?” The option to leave was supported by a slim majority: 51.9% voted for “Brexit”, while 48.1% voted to Remain. The official groups campaigning for the UK to remain in or leave the EU were called “Britain Stronger in Europe” and “Vote Leave”. Both these groups made extensive use of social media such as Facebook to pursue their campaigns. Social media can provide us with an enormous amount of data about a vast range of topics, but is it possible to understand, summarize and communicate such data in a way that provides real insight?

Monitoring Facebook activity
A range of information is available for each post to Facebook public pages including its message, when it was created and the number of likes, comments and shares it has received. In the run-up to the referendum we monitored the number of likes, comments and shares per day for posts to the “Britain Stronger in Europe” and “Vote Leave” pages. We can communicate this information by an appropriate graph (Figure 1).

We can immediately see that there was increasing activity as the referendum date approached and that the average number of likes and shares for “Vote Leave” was considerably higher than for “Britain Stronger in Europe”.

Producing graphs using the R software
Figure 1 was produced using the ggplot2 package running in R, a free software environment for statistical computing and graphics. ggplot2 is freely available and allows the user to produce highly customizable graphs, including graphs such as Figure 1 that make use of faceting. Faceting allows similar plots to be produced for different subsets of the data, such as “Britain Stronger in Europe” with the monthly average number of likes, or “Vote Leave” with

---

**Figure 1.** Graphs with facets can be used to communicate different measures of activity on contrasting Facebook pages
the monthly average number of shares. In this way the communication of quite complicated information that changes across time is rendered simpler. Of course, the use of ggplot2 and the other R packages needed to download data from Facebook and to produce Figure 1 requires software-specific knowledge, together with data handling and coding experience. Producing plots that communicate the message that underlies data, especially the complicated and unstructured data that is increasingly available these days, can be a time consuming task. We should involve our audience in this task, seeking feedback and being prepared to act upon criticism. A graph that we create for our own use may be produced in five minutes, but we should allow considerable time for a graph that may be viewed by thousands of people. We may not always get it right, and we should always learn by experience.

Figure 2. Word clouds can be used to summarise text data posted on Facebook pages

Figure 3. We can represent how sentiments expressed on Facebook pages change over time
Text data and sentiment analysis

The most substantive part of a Facebook post is the message that it contains. Communicating the underlying views contained in a large number of Facebook posts can be challenging. A simple approach is to produce and display word counts using bar charts of the most commonly occurring words, or using word clouds. With word clouds, the more often a word is used, the larger and more centrally it is displayed. Word clouds for our two examples (Figure 2) can be used to highlight the main issues posted on the referendum campaign pages. If we had to communicate a few phrases from each campaign based on each word cloud, we could write “Sharing is better for the economy and important for families.” The veracity of these phrases will become clear as time passes! Several R packages are available to facilitate the production of word clouds.

It is also possible to perform a simple Sentiment Analysis on the text data to determine whether the opinions expressed on pages are generally positive, negative or neutral. A straightforward approach involves matching posts to dictionaries of positive and negative words. A phrase such as “This is good” has one positive match and so scores 1, while the phrases “This is not good” (a negated positive match) and “This is bad” (a negative match) score 0 and -1. Because we know the date of each post, we can investigate how the sentiment scores change over time. Average sentiment scores for each month are shown (Figure 3) using dots. The error bars indicate how much variation there is in sentiment scores within each month, with wider bars meaning that the sentiment scores for posts in that month take on a broader range of values. Posts on the “Britain Stronger in Europe” page seem to be relatively constant in their positivity, while those on “Vote Leave” show a slight decline in both positivity and range as the referendum date approached.

Surname diversity and the referendum results

An analysis of the surnames of 45.6 million people resident in the UK in October 2001 was published in 2008. Because a 2001 change in the law allowed voters to “opt out” of the publicly available version of the electoral register, we believe that this data set is the best source of UK surname data presently available. From these data, the percentage of people in each UK administrative district with unique surnames (indicated by HL, after the term *hapax legomenon* used in literature vocabulary analysis to denote a word that appears once in a text) can be calculated and used as a measure of surname diversity in that district. Higher values of HL indicate an increased number of unique surnames and therefore a higher surname diversity in the district. HL values can be displayed as a map (Figure 4) which shows, for example, that Northern Ireland, South Wales and much of Scotland stand out for having more districts with lower percentages of people with unique surnames. One of the challenges of constructing such a map is to choose the colour palette in such a way that area differences appear clearly. Very compact regions can be shown separately, as has been done here for London.
A natural question would be whether surname diversity could have been used to gain some insight into the referendum results, so suggesting where campaign resources could have been targeted. To answer this we can produce a graph which shows for each UK region the percentage of voters opting for Leave and the percentage of people with unique surnames in each region (Figure 5). We use coloured text to indicate the regions and the four countries that comprise the UK. We have also adopted a logarithmic scale for the percentage of people with unique surnames in which the gap between values becomes smaller as the values get bigger. We need this scale because the percentage of people with unique surnames in Greater London is so relatively large that, if it were plotted on an ordinary scale, the data points corresponding to the other regions would be squashed towards the left of the plot rendering them unreadable. We have also added a horizontal line at 50%: regions above this line opted for Leave, while those below it opted for Remain. The message that Figure 5 provides it that as the percentage of people with unique surnames increases, the percentage of people voting Leave decreases, as illustrated by the added sloping line. Scotland and Northern Ireland are exceptions to this trend, as those countries expressed strong Remain sentiments but have low surname diversity. The use of colour, text and suitably designed scales enable us to present the underlying message in a complicated data set in an easily understood way.

We can further investigate the regional dependencies that we saw (Figure 5) on the finer scale of administrative districts. The results (Figure 6) that we present do not include Northern Ireland (and some other areas) as we were unable to match past Northern Ireland administrative districts with the referendum results. Again, we have used a different colour for country, but, instead of printing text, we used a different size dot to indicate the total number of votes cast. Many of the England points lie above the 50% line, indicating that these areas opted for Leave. The situation is markedly different for Scotland, all areas of which opted for Remain. As before, a special, logarithmic scale has been used for the percentage of people with unique surnames to enable the full range of values to be clearly displayed. A smooth curve has also been added for each country to show the underlying dependencies. In general, the percentage of voters opting for Leave decreases as surname diversity increases, with the effect being less strong in Scotland, probably due to the generally strong Remain feeling there. Of course, here our conclusions must be regarded tentatively both because the surname data used to calculate the surname measures is fifteen years old and so there may be a bias towards underestimating surname variability, and because the matching between previous administrative districts and 2016 voting areas may not be precise. However, the way in which we have presented our analyses clearly communicates that looking at surname diversity can provide additional useful insights to political thinking and strategic planning.

Figure 6. The message underlying complicated data can be communicated by using faceting, colour, dots of different sizes and curves that indicate underlying dependencies
Conclusions
We have seen a variety of ways of communicating information about the 2016 UK European Union membership referendum and its results. We have presented examples of how social media activity can be monitored, of how text data can be summarized, and of how sentiments can be analysed.

Moreover, we have discussed the presentation of: area-based data using a map with a suitable colour palette; data showing complicated dependences using faceted graphs that employ colour, text and plotting symbols of meaningful sizes; and logarithmic scales and curves indicating the underlying relationships. Although the data with which we have worked relate to the recent referendum, the visualization techniques discussed could be applied to a massive amount of data coming from areas such as medicine, business, industry, computing, literature, climate change and sport.

All the plots that we have presented have been produced using R and a number of contributed packages, and all of this software can be downloaded from the internet free of charge. We do not claim that designing graphs that communicate complicated ideas well is a quick or easy task. Indeed, considerable technical skill and data experience may be required, and it is for this reason that Data Science type undergraduate and graduate programmes are now being offered. We do, however, believe that these days many tools exist to produce high quality, insightful data presentations that allow us to communicate in a relatively simple and attractive way, the messages that underlie sometimes quite complicated data sets.

References

Further reading

Julian Stander is Associate Professor (Reader) in Mathematics and Statistics in the School of Computing, Electronics and Mathematics, Plymouth University. He has an MA in Mathematics from the University of Oxford, a Diploma in Mathematical Statistics from the University of Cambridge and a PhD in Statistics from the School of Mathematical Sciences at the University of Bath. He served on the Council of the Royal Statistical Society from 2012 to 2015.
E: j.stander@plymouth.ac.uk
W: www.plymouth.ac.uk/staff/julian-stander

Luciana Dalla Valle is a Lecturer in Statistics in the School of Computing, Electronics and Mathematics, Plymouth University. She has a MSc degree from the University of Pavia and a PhD from the University of Milan–Bicocca. She has served as Secretary of the South West Local Group of the Royal Statistical Society since 2011.
E: luciana.dallavalle@plymouth.ac.uk
W: www.plymouth.ac.uk/staff/luciana-dalla-valle

Julian Stander and Luciana Dalla Valle have considerable teaching and consultancy experience in many areas of statistics and have recently designed and taught modules on data manipulation, visualization and modelling at Plymouth University.

Mario Cortina Borja is chairman of the Significance editorial board. He has been at the Centre for Paediatric Epidemiology and Biostatistics in the UCL Institute of Child Health since October 2000, where he is Professor of Biostatistics. He has a BSc in Actuarial Science and an MSc in Statistics from the Universidad Nacional Autónoma de México, and a PhD in Statistics from the School of Mathematical Sciences at the University of Bath. Before coming to UCL, Mario was a Statistician at the Instituto de Investigaciones Antropológicas in Mexico City, a Research Officer in applied mathematics at the School of Chemical Engineering in the University of Bath, a Lecturer, then Senior Lecturer in Statistics at the Instituto Tecnológico Autónomo de México, and, for five years, Consulting and Teaching Officer in the Department of Statistics, University of Oxford.
E: m.cortina@ucl.ac.uk
W: www.ucl.ac.uk/ich/research/population-policy-practice/People/cortina-borja-Mario

Mario Cortina Borja has vast experience of developing and applying applied statistics methodology in a wide range of areas including child health, medicine, engineering, epidemiology and psychology. He is a noted expert on the seasonal patterns of disease, especially the effect of date of birth on adult health.