Addressing New Challenges Including Ethics, of Big Data Analytics, with the Correspondence Analysis and Geometric Data Analysis Platforms

The Geometry and Topology of Data and Information for Analytics of Processes and Behaviours: Building on Bourdieu and Addressing New Societal Challenges

JSTAR2016, October 2016

Fionn Murtagh

- Part 1: New challenges and opportunities in the context of Big Data analytics.
- Part 2: Social media analytics. (Example of use of Jürgen Habermas's convergent processes.)
- Part 3: Analytical focus and contextualization. (Example: mental health analytics.)
- Part 4: Semantic mapping. (Large scale Twitter analytics.)

Part 1

The following slides will discuss the importance for Big Data analytics, and associated paradigm shifts.
Then: Open Data and lots of other sources.
Essential analytics methodology uses the geometry and topology of data and information.

"This is my motto: Analysis is nothing, data are everything. Today, on the web, we can have baskets full of data ... baskets or bins?"

Jean-Paul Benzécri, 2011.

N. Keiding and T.A. Louis, Perils and potentials of self-selected entry to epidemiological studies and surveys, Jnl.Royal Statist. Soc. A, 179, Part 2, pp. 319-376, 2016

- This comprehensive survey (118 citations) sets out new contemporary issues of sampling and population distribution estimation. An interesting conclusion 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.

Keiding and Louis (contd.)

- "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 is interesting.
- “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.”

- In what follows: semantic contextualization, aggregation, metric (i.e. geometrical) and ultrametric (i.e. topological) mapping that incorporates correlation.

AAPOR Report on Big Data

AAPOR Big Data Task Force

February 12, 2015

Prepared for AAPOR Council by the Task Force, with Task Force members including:

***Lilli Japec**, Co-Chair, Statistics Sweden*
***Frauke Kreuter**, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB*
***Marcus Berg**, Stockholm University*
***Paul Biemer**, RTI International*
***Paul Decker**, Mathematica Policy Research*
***Cliff Lampe**, School of Information at the University of Michigan*
***Julia Lane**, American Institutes for Research*
***Cathy O'Neil**, Johnson Research Labs*
***Abe Usher**, HumanGeo Group*

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

3 April 2013 Last updated at 12:42

The Great British Class Survey – Results

Middle class? | Class calculator | US view | Reader reactions | 'Huge survey'

Mike Savage and Fiona Devine examined class in a brand new way

Mike Savage from the London School of Economics and Fiona Devine from the University of Manchester describe their findings from The Great British Class Survey. Their results identify a new

Related Stories

How do you identify

3 April 2013 Last updated at 12:43

How do you identify new types of class?

Middle class? | Class calculator | US view | Reader reactions | 'Huge sur

The Great British Class Survey is the biggest scientific investigation into social class in the UK

Sociologists are interested in the idea that class is about your cultural tastes and activities as well as the type and number of people you know.

These factors are important when put alongside people's economic position.

Professors Mike Savage and Fiona Devine explain how **a BBC Lab UK experiment** allowed them to better understand class in the 21st Century.

Related Stories

The Great British Class Survey – Results

The Great British class calculator

UK 'now has seven social classes'

Measuring Class

Understanding classes as amounts of different types of 'capitals' helps us to see class across a number of dimensions.

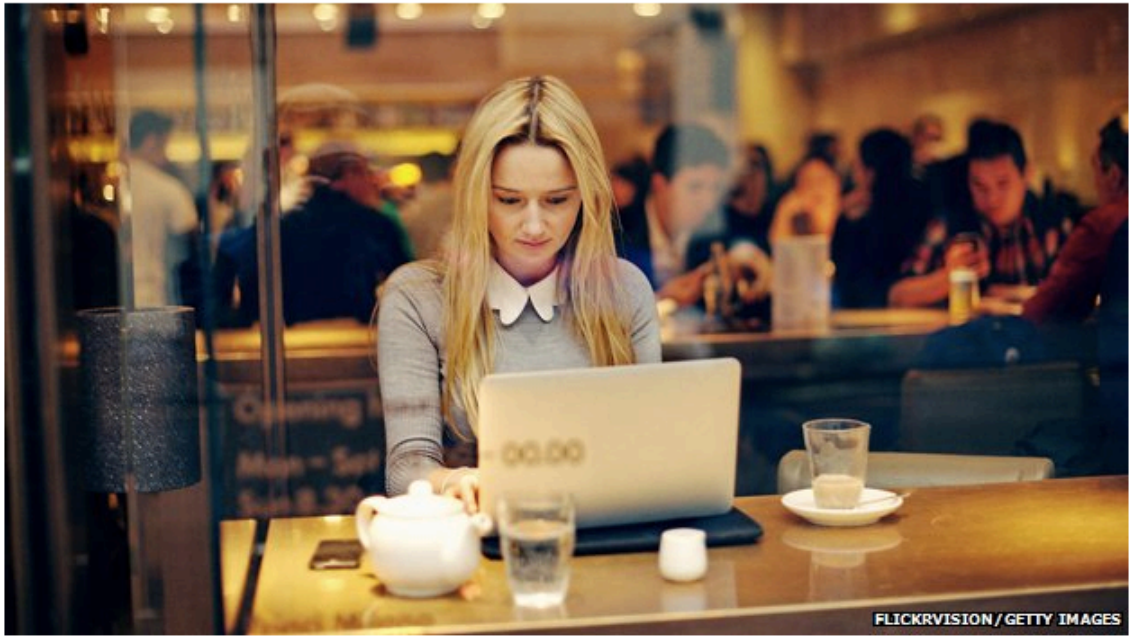
The French sociologist, Pierre Bourdieu first developed this approach in 1984, suggesting there are different types of capitals which give people an advantage in life. Economic, cultural and social capital may overlap but they are different. Using this approach, we distinguished between people with different amounts of each of

Pierre Bourdieu investigated what propelled people into the upper strata

22 April 2013 Last updated at 13:35

How can the results from a web survey represent a whole society?

Middle class? | Class calculator | US view | Reader reactions | 'Huge survey'



Sam Friedman, Daniel Laurison and Andrew Miles (2015), "Breaking the 'class' ceiling? Social mobility into Britain's elite occupations", The Sociological Review, Vol. 63, Issue 2, pp 259–289, May 2015

- "... 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 ..."

D. Laurison in an associated blog posting; and
Züll, Cornelia, and Evi Scholz. 2015. "Who is Willing to Answer Open-ended Questions on the Meaning of Left and Right?"
Bulletin of Sociological Methodology 127 (1): 26-42.

- "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".

F. Lebaron, "How Bourdieu "quantified" Bourdieu: the geometric modelling of data", chapter 2 in K. Robson and C. Sanders, Eds., Quantifying Theory: Pierre Bourdieu, Springer 2009.

- 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).

- Through analysis of social fields and homology, and beyond.
- General and broad relevance and applicability.
- Learning and drawing inspiration from the work of P. Bourdieu.
- Underpinning analytics, we may consider that we have: data → information → knowledge → wisdom.
- Geometric data analysis, and based on GDA, topology, is the essential basis for all such work.

Part 2

- Social media analytics.
  (Example of use of Jürgen Habermas's convergent processes.)

- Take some work of noted social / political science theorist, Jürgen Habermas
- Motivated by: Theory of Communicative Action (Theorie des kommunikativen Handelns)

1) Testing social media with the aim of designing interventions

2) Application here to environmental communication initiatives

3) Measuring impact of public engagement theory

(in the Jürgen Habermas sense – public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship)

1) Qualitative data analysis of twitter.

2) Nearly 1000 tweets between 1.10.2012 and 24.11.2012.

3) Evaluation of tweet interventions.

4) Eight separate twitter campaigns carried out.

1) Testing social media with the aim of designing interventions

2) Application here to environmental communication initiatives

3) Measuring impact of public engagement theory

(in the Jürgen Habermas sense – public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship)

1) Qualitative data analysis of twitter.

2) Nearly 1000 tweets between 1.10.2012 and 24.11.2012.

3) Evaluation of tweet interventions.

4) Eight separate twitter campaigns carried out.

Background and aims

- Analyse the semantics of the discourse in a data-driven way. (Pianosi, Bull and Rieser: "top-down communication campaigns both predominate and are advised by those involved in "social marketing" …. However, this rarely manifests itself through measurable behaviour change …") Thus our approach is, in its point of departure and vantage point, bottom-up.

- Mediated by the latent semantic mapping of the discourse, we develop semantic distance measures between deliberative actions and the aggregate social effect. We let the data speak in regard to influence, impact and reach.

- Impact: semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested. It can be visualized. It can be further visualized and evaluated.

The 8 Twitter campaigns in late 2012

1.10-7.10: Climate change: The big picture and the global consequences

8.10-14.10: Climate change: The local consequences

15.10-22.10: Light and electricity

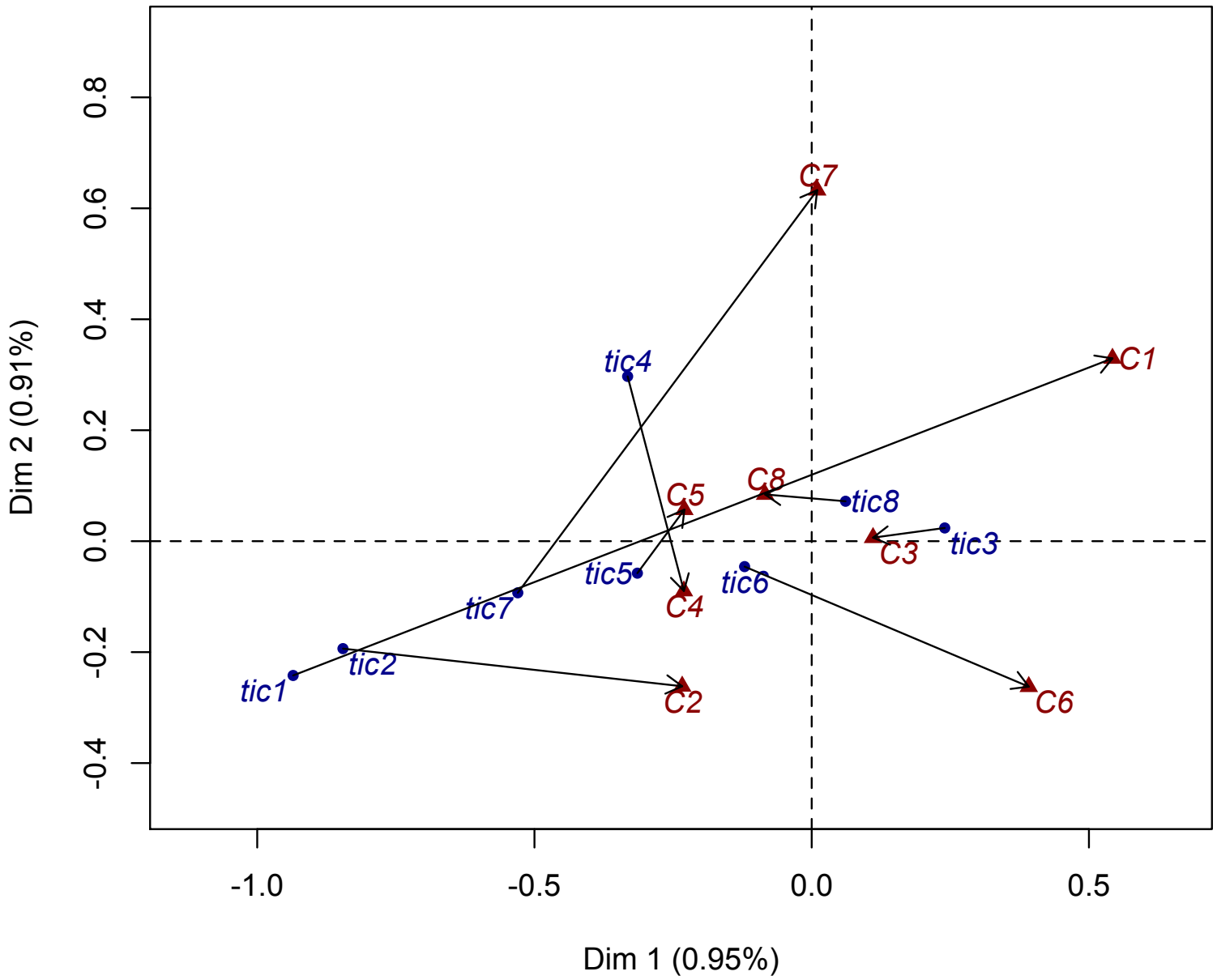
23.10-28.10: Heating systems

29.10-4.11: Sustainable Food choices

5.11-11.11: Sustainable Travel choices

12.11-18.11: Sustainable Water use

19.11-25.11: Sustainable Waste


8 campaign initiating tweets, and centres of gravity of 8 campaigns
Dim 2 (0.91%)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-1.0
-0.5
0.0
0.5
Dim 1 (0.95%)
C7
C1
tic4
C8
C5
tic8
C3
tic3
tic5
tic6
C4
tic7
tic2
tic1
C2
C6

Statistical significance of impact

- The campaign 7 case, with the distance between the tweet initiating campaign 7, and the mean campaign 7 outcome, in the full, 338-dimensional factor (semantic) space equal to 3.670904.
- Compare that to all pairwise distances of non-initiating tweets. (They are quite normally distributed, with a small number of large distances.)

- F. Murtagh, M. Pianosi, R. Bull, "Semantic Mapping of Discourse and Activity, Using Habermas's Theory of Communicative Action to Analyze Process", **Quality and Quantity**, 50(4), 1675-1694, 2016.

Part 3
Analytical Focus and Contextualization

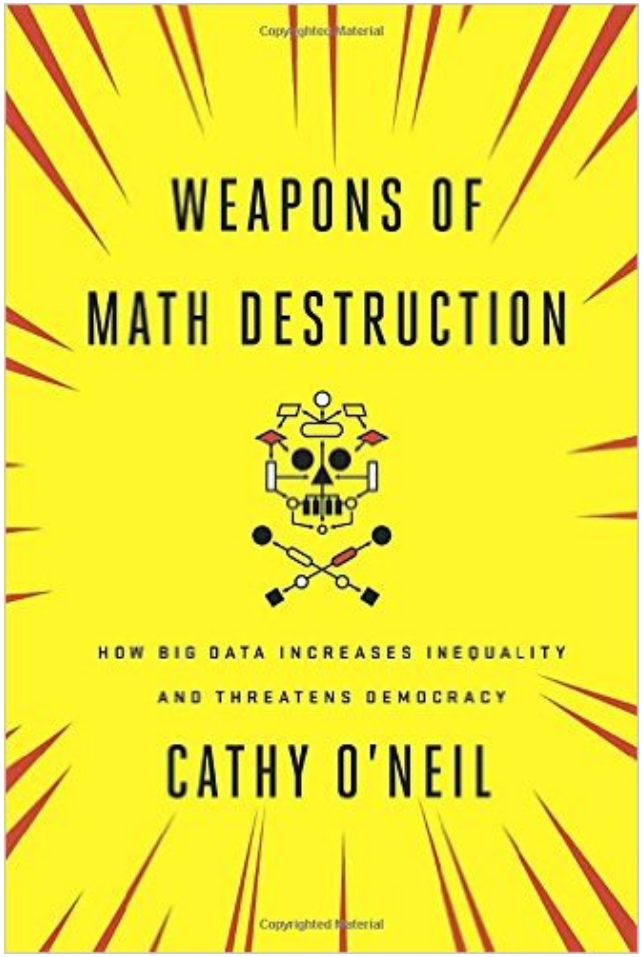
- In the following preliminary study of mental health, the following is described:
- Choice and selection of main and supplementary variables.
- Therefore: our main focus of analysis, and an explanatory context.

- "Rehabilitation of individuals. The context model is always formulated at the individual level, being opposed therefore to modelling at an aggregate level for which the individuals are only an 'error term' of the model."

- B. Le Roux and F. Lebaron. "Idées-clefs de l'analyse géometrique des données" (Key ideas in the geometric analysis of data). In F. Lebaron and B. Le Roux, editors, La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données, pages 3-20. Dunod, Paris, 2015.

Cathy O'Neil (cf authors of AAPOR Report on Big Data)
Crown Publishing Group (NY), Sept. 2016

At issue:
replacing
individuals
by a cluster or
group.



Mental Health: Adult Psychiatric Morbidity Survey, England, 2007

- HSCIC, Health and Social Care Information Centre (National Health Service, UK) (2009). National Statistics Adult Psychiatric Morbidity in England – 2007, Results of a household survey, Appendices and Glossary. 174 pp.
- Available at: http://www.hscic.gov.uk/pubs/psychiatricmorbidity07

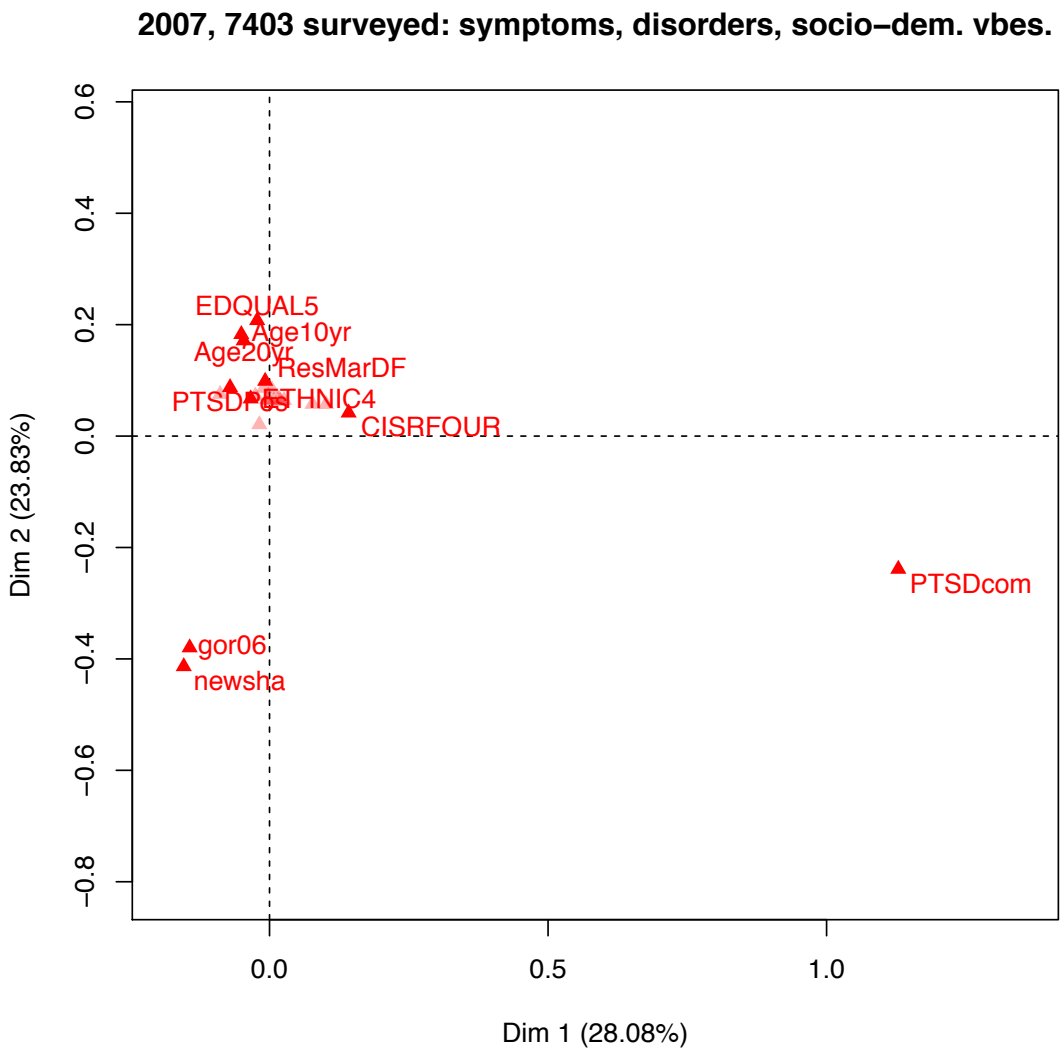
- 1704 variables, including questioning of the subjects about symptoms and disorders, psychoses and depression characteristics, anti-social behaviours, eating characteristics and alcohol consumption, drug use, and socio-demographics, including gender, age, educational level, marital status, employment status, and region lived in.

First analysis: relating to "Neurotic symptoms and common mental disorders"

- 14 questions, hence 14 categorical variables, 14 categorical variables, relating to this.
- Almost all of these variables had as question responses, whether or not there were symptoms or disorders in the past week, one question related to one's lifetime, and one question related to the age of 16 onwards.
- Another question set was selected, relating to 9 socio-demographic variables.

- It was checked whether neurotic symptoms and common mental disorders data should be jointly analysed with the socio-demographic data. The following figure shows this outcome.
- On the positive first factor, what is particularly important, from the contribution to the inertia, is PTSDcom, “TSQ (Trauma Screening Questionnaire) total score".
- The negative second factor, is highly influenced by these two variables: “gor06", “newsha", respectively: “Government office region" and “Strategic Health Authorities". These were both sets of geographic regions in England, respectively 9 and 10. (The 7th such region in each of these variables was London.)
- A summary interpretation derived from the following figure is how factor 1 accounts for recorded trauma, and factor 2 accounts for region of the respondent.

Adult psychiatric morbidity survey 2007, England, household survey.
The analysis has the neurotic symptoms and common mental disorders, and the socio-demographic variables. Displayed are the 10 highest contributing variables to the principal plane.



Future work relating to Bourdieu's field and homology analytics: Mental capital

Kleinman A., Lockwood Estrin G., Usmani S., Chisholm D., Marquez P.V., Evans T.G., and Saxena S. (2016), "Time for mental health to come out of the shadows". The Lancet vol: 387, 2274-2275.

Cooper C., Wilsdon J., and Shooter M. (2016), "Making the Case for the Social Sciences, No. 9 Mental Wellbeing". 28 pp. BACP, British Association for Counselling and Psychotherapy.

- It is noted in Kleinman et al. (2016) how relevant and important mental health is, given the integral association with physical health. There is the following: "... parity between mental and physical health conditions remains a distant ideal". "The global economy loses about $1 trillion every year in productivity due to depression and anxiety". "Next steps include ... integration of mental health into other health and development sectors".

- In Cooper et al. (2016), under the heading of "Five Ways to Wellbeing", reference is made to "mental capital and wellbeing". A section (page 14) is entitled "The `mental capital' values of the outdoors". At issue here: going for a walk, therefore taking exercise, in the countryside or parkland.

Part 4: Semantic Mapping

Semantic mapping of Twitter data, relating to festivals – music, film, art, parades.

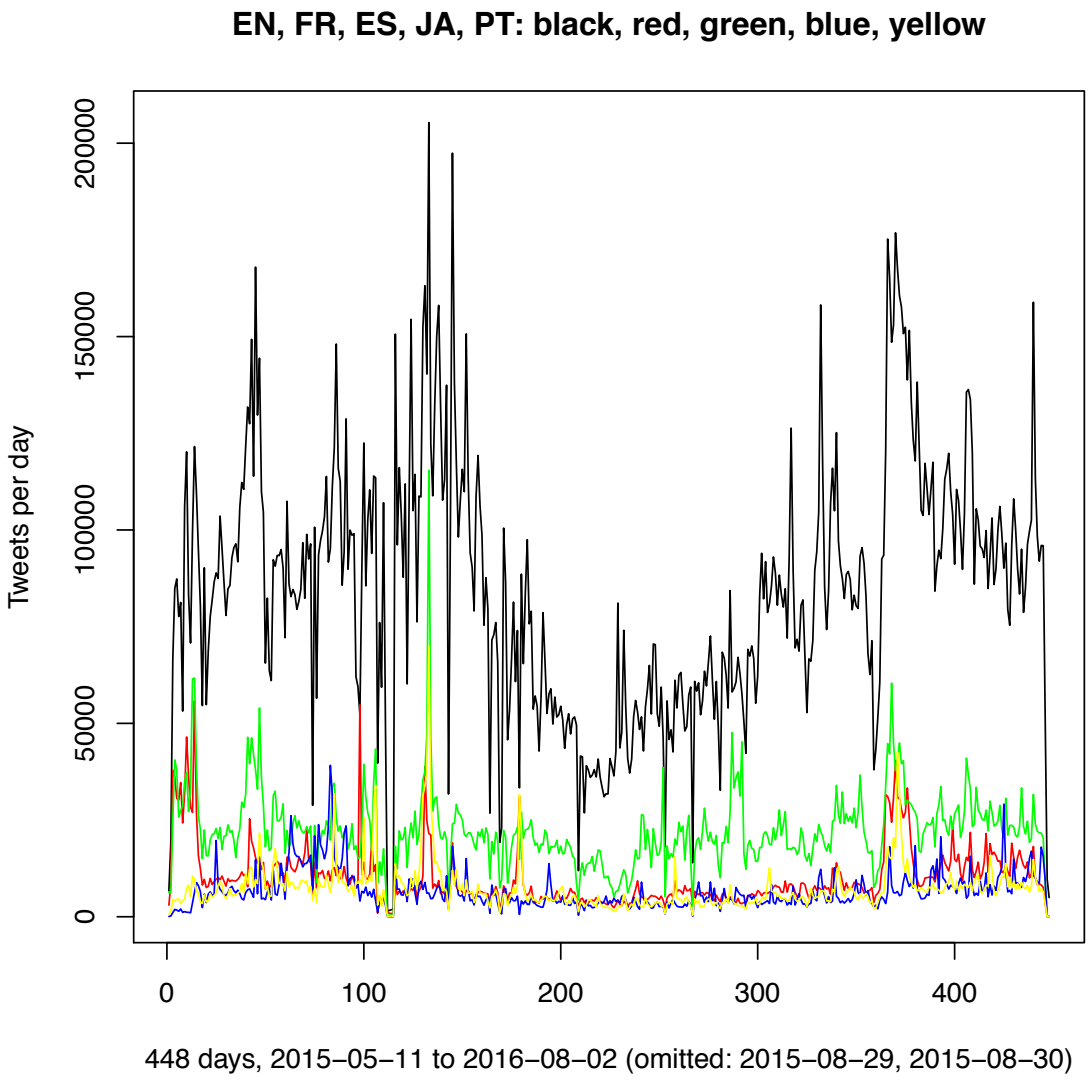
Analysis: towards behavioural or activity patterns or trends

- Consider the tweeter, variable: from_user
- As an important characteristic of the tweeter, consider the associated language, variable: iso_language_code
- This language indicated is indicative only. We assume it to be an indicative, and relevant, characteristic of the tweeter.
- From 2015-05-11 to end 2015 data: 75 languages in use, including Japanese, Arabic and so on, with the majority in Roman script.

For a language label, the activity over time

- As indicative association to language, we take the following: English, Spanish, French, Japanese, Portuguese.
- We consider the days 2015-05-11 to 2016-08-02, with two days removed, 2015-08-29, 2015-08-30, due to lack of tweets.
- The numbers of tweets for these languages were as follows (carried out on 11 Aug. 2016): en, 37681771; es, 9984507; fr, 4503113; ja, 2977159; pt, 3270839
- Following removal of the null dates, and the two days with 0 tweets for some case, there were these totals of tweets:
- en, 37681567; es, 9984485; fr, 4503077; ja, 2977156; pt, 3270837

The dominant labels of tweets for these languages: English, French, Spanish, Japanese, Portuguese. There are numbers of tweets per day from 2015-05-11 to 2016-08-02, with two days removed.



- For a given language label – which serves as a preliminary selection that is potentially relevant for one or more specific festivals – carry out the following.
- Determine the set of tweeters (variable from_user), crossed by the day (variable created_at). In this matrix there is the number of tweets by each tweeter on each day.
- Just consider days that have non-zero tweet activity.

- Compactification – piling – is quite clearly the case here.
- This can be due to very high dimensionality. Or sparsity.
- The following slide just expresses this Euclidean distance endowed, factor space (latent variable space) mapping.
- That motivates the behavioural analysis carried out here: to look for tweeters and days that are exceptional for factors. I.e. tweeters and days that are the most important and predominant in defining the factors.

Portuguese, ISO Label PT

- For iso_lang_code as PT, Portuguese, there were:
- 449 days, 744584 tweeters, maximum of 21 tweets by a tweeter per day, and total number of tweets: 3166418
- (Carried out on 29 August 2016.)
- Very sparse matrix, tweeters x days: 0.63% non-zero values.
- The first eigenvalues explained the following % inertia: 0.49, 0.46, 0.42, 0.39, 0.38, 0.37, 0.36, 0.34, 0.33 …

- (Top contributions to factors are in following slide.)

The first few factors, the highest contributions (inertia, that defines the factor) by tweeter, and by day

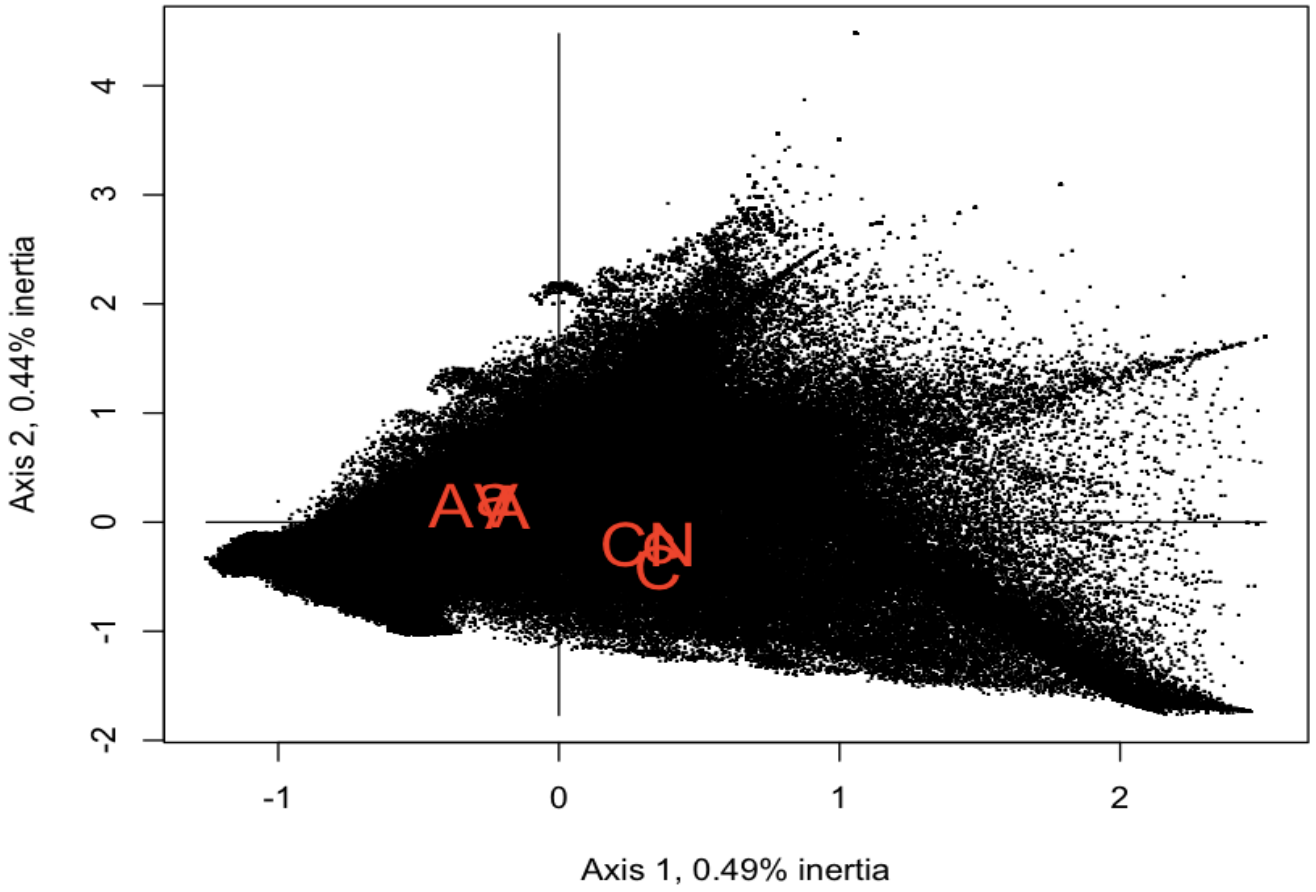
Factor	tweeter (contr.)	tweeter (contr.)	tweeter (contr.)
1	GiannaAdamson (0.002)	gumartinslive (0.001)	CamilaMacielSer (0.001)
2	newslarts (0.0005)	herdersonfile (0.0004)	biebersmaniabr (0.0004)
3	Onlifebiebsr (0.0005)	justindrewluz (0.0005)	sandybieber1 (0.0005)
4	AlissonNobrien (0.0007)	Music_Industry3 (0.0006)	PAULINAUSURPA (0.0005)
5	cacolizei (0.002)	knowhoneymoon (0.001)	Tayanediass (0.001)

Factor	day (contr.)	day (contr.)	day (contr.)
1	2015-09-22 (0.018)	2016-05-17 (0.013)	2015-09-21 (0.012)
2	2015-11-07 (0.111)	2015-08-24 (0.086)	2015-08-23 (0.060)
3	2015-11-07 (0.161)	2015-09-22 (0.143)	2015-08-24 (0.084)
4	2015-09-22 (0.108)	2015-11-07 (0.059)	2015-09-23 (0.042)
5	2015-12-17 (0.656)	2016-01-25 (0.017)	2015-08-24 (0.013)

The Tweeters and the Festivals

- Tweets characterized as French, accessed on 11 Sept. 2016 (dating from 11 May 2015).
- 4913781 tweets. (For user, date and tweet content, the file size was: 667242607 bytes.)
- The following were sought in the tweets: Cannes, cannes, CANNES, Avignon, avignon, AVIGNON.
- Upper and lower case retained in order to verify semantic proximity of these variants.
- By the way, these related to the Cannes Film Festival, and the Avignon Theatre Festival.
- The following total numbers of occurrences of these words were found, and the maximum number of occurrences by a user: Cannes, 1230559 and 3388; cannes, 145939 and 4024; CANNES, 57763 and 829; Avignon, 272812 and 4238; avignon, 39323 and 2909; AVIGNON, 14647 and 900.
- Total number of tweeters, also called users here: 880664; total number of days retained, from 11 May 2015 to 11 Sept. 2016, 481
- Cross-tabulated are: 880664 users by 481 days. There are 1230559 retained and recorded tweets. The non-sparsity of this matrix is just: 0.79%

C, c, CA (Cannes, cannes, CANNES) and A, a, AV (Avignon, avignon, AVIGNON)
They are supplementary variables in the Correspondence Analysis principal factor plane.
Semantically they are clustered. They are against the background of the Big Data, here the 880664 tweeters, represented by dots.



Projections of supplementary variables on the principal factors (to 2 digits of precision)

word	Factor 1	Factor 2
Cannes	0.357	-0: 426
cannes	0.355	-0: 244
CANNES	0.320	-0: 2131
Avignon	- 0.182	0.136
avignon	- 0: 226	0.193
AVIGNON	- 0: 303	0.141

So Cannes is positive F1, negative F2; Avignon is negative F1, positive F2. We see their opposition or polarity, against the background of the large set of tweets.

Current Considerations

- Determine some other, related or otherwise, behavioural patterns that are accessible in the latent semantic, factor space.
- Retain selected terms from the tweets, and, as supplementary elements, see how they provide more information on patterns and trends.
- Year by year trend analysis.

Some references

- (1) F. Murtagh “Semantic mapping: towards contextual and trend analysis of behaviours and practices”, in K. Balog, L. Cappellato, N. Ferro, C. MacDonald, Eds., **Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum**, Évora, Portugal, 5-8 September, 2016}, pp. 1207--1225, 2016. http://ceur-ws.org/Vol-1609.
- (2) F. Murtagh, “**The geometry and topology of data and information for analytics of processes and behaviours: building on Bourdieu and addressing new societal challenges**”, 2016, submitted.
- (3) F. Murtagh, “**Sparse p-Adic Data Coding for Computationally Efficient and Effective Big Data Analytics**”, p-Adic Numbers, Ultrametric Analysis and Applications, 8(3), 236-247, 2016.
- (4) F Murtagh, “**Big Data Scaling through Metric Mapping: Exploiting the Remarkable Simplicity of Very High Dimensional Spaces using Correspondence Analysis”**, Proc. IFCS 2015 – International Federation of Classification Societies, forthcoming, 2016.
- (5) F Murtagh and P Contreras, “**Clustering through High Dimensional Data Scaling: Applications and Implementations”**, ECDA 2015 – European Conference on Data Analysis, Archives of Data Science (KIT Scientific Publishing), in press, 2016.