Big Textual Data: Lessons and Challenges for Statistics

Grandi dati testuali: le lezioni e le sfide per le statistiche

Fionn Murtagh

Abstract At issue are a few early stage case studies relating to: research publishing and research impact; literature, narrative and foundational emotional tracking; and social media, here Twitter, with a social science orientation. Central relevance and importance will be associated with the following aspects of analytical methodology: context, leading to availing of semantics; focus, motivating homology between fields of analytical orientation; resolution scale, which can incorporate a concept hierarchy and aggregation in general; and acknowledging all that is implied by this expression: correlation is not causation. Application areas are: research publishing and qualitative assessment, narrative analysis and assessing impact, and baselining and contextualizing, statistically and in related aspects such as visualization.

Key words: mapping narrative, emotion tracking, significance of style, Correspondence Analysis, chronological hierarchical clustering

1 Underlying Themes in Methodology, Introduction

Clearly, through integration of analytical methodology and domain of application, the choice of methodology or even its development is dependent on the specific requirements. However the following general aspects of contemporary analytics, including textual data analytics, are useful to be noted.

An interview with Peter Norvig, Google, in C. Anderson [1] contained the following controversial perspectives: “Petabytes allow us to say: ‘Correlation is enough’. We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where

Fionn Murtagh
Institute of Mathematics and Data Science, University of Huddersfield, UK e-mail: fmurtagh@acm.org

science cannot." "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

To counteract this automation of all analytical reasoning, and to accept the need for inductive reasoning, there is: (1) Importance of: context (for our analytics); integration of data and domain; leading to the following. (2) Semantic analytics, and analytical synthesis in, and from, data and information. (3) Qualitative as well as quantitative evaluation and related analytics. All in all, this is leading to the Correspondence Analysis platform as an inductive reasoning framework for other analytical methodologies also.

Interestingly, the focus on regions of interest in information space is stressed by [21]. An article about the Internet of Things and Big Data by John Thornhill in the newspaper, Financial Times, on 9 January 2017 had this comment: "Sir Nigel Shadbolt, co-founder of the Open Data Institute ... The next impending revolution, he argues, will be about giving consumers control over their data."

Ethical consequences of Big Data mining and analysis may be associated with the following, from [10]: "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."

In [6], "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. 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, ...". 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.

Furthermore there is this: "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".

Hence our motivation for the following framework for analytical processes: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. A further useful case is when the hierarchy respects chronological or other sequence information.

2 Towards: Qualitative as well as Quantitative Research Effectiveness and Impact

For analysis of research funding, of publishing, and of commercial outcomes, account needs to be taken of measures of esteem. Also account is taken of research impact, through impact of research products: (1) research results, (2) organisation of science (journal editing, running conferences), (3) knowledge transfer, supervision, (4) technology innovations.

Correspondence Analysis when based on part of an ontology or concept hierarchy can be considered as "information focusing". Correspondence Analysis provides simultaneous representation of observations and attributes. We project other observations or attributes into the factor space: these are supplementary or contextual observations or attributes. A 2-dimensional or planar view is an approximation of the full cloud of observations or of attributes. Therefore there can be benefit in the following: define a small number of aggregates of either observations or attributes, and carry out the analysis on them. Then project the full set of observations and attributes into the factor space.

In support of "The Leiden Manifesto for research metrics", DORA (San Francisco Declaration on Research Assessment), Metrics Tide Report (HEFCE, Higher Education Funding Council England, 2015), qualitative judgement is primary. Research results may be assessed through first determining a taxonomic rank by mapping to a taxonomy of the domain (a manual action). There there will be unsupervised aggregation of criteria for stratification.

Research impact should be evaluated, first of all, based on qualitative considerations. Evaluation of research, especially at the level of teams or individuals can be organized by, firstly, developing and maintaining a taxonomy of the relevant subdomains and, secondly, a system for mapping research results to those subdomains that have been created or significantly transformed because of these research results. Of course, developing and/or incorporating systems for other elements of research impact, viz., knowledge transfer, industrial applications, social interactions, etc., are to be taken into account also.

See [19] for such work. Generally also see [5]. The latter maps out evolving vocabulary and associates this also with influential published articles.

3 Qualitative Style in Narrative for Analysis and Synthesis of Narrative

For [11], the composition of the movie, Casablanca, is "virtually perfect". Text is the "sensory surface" of the underlying semantics.

Here there is consideration as to how permutation testing and evaluation can be very relevant for qualitative appraisal. Considering the Casablanca movie, shot by Warner Brothers between May and August 1942, and also some early episodes of the CSI Las Vegas, Crime Scene Investigation, television drama series, from the year 2000, the attributes used were as follows, [15].

All is based on the following: Euclidean geometry for semantics of information; hierarchical topology for other aspects of semantics, and in particular how a hierarchy expresses anomaly or change. The hierarchy respects chronological or other sequence information. Chronological hierarchical clustering, also termed contiguity constrained hierarchical clustering, is based on the complete link agglomerative clustering criterion [12, 2, 7].

1. Attributes 1 and 2: The relative movement, given by the mean squared distance from one scene to the next. We take the mean and the variance of these relative movements. Attributes 1 and 2 are based on the (full-dimensionality) factor space embedding of the scenes.
2. Attributes 3 and 4: The changes in direction, given by the squared difference in correlation from one scene to the next. We take the mean and variance of these changes in direction. Attributes 3 and 4 are based on the (full-dimensionality) correlations with factors.
3. Attribute 5 is mean absolute tempo. Tempo is given by difference in scene length from one scene to the next. Attribute 6 is the mean of the ups and downs of tempo.
4. Attributes 7 and 8 are, respectively, the mean and variance of rhythm given by the sums of squared deviations from one scene length to the next.
5. Finally, attribute 9 is the mean of the rhythm taking up or down into account.

For permutation testing, assessment was carried out relative to uniformly randomized sequences of scenes or sub-scenes.

4 Statistical Significance of Impact

Underlying [18] is the testing of social media with the aim of designing interventions, associated with statistical assessment of impact. The application here is to environmental communication initiatives. Measuring impact of public engagement theory, in the sense of the eminent political scientist, Jürgen Habermas, involves public engagement centred on communicative theory; by implication therefore, discourse as a possible route to social learning and environmental citizenship.

The case study here, was directed towards:

1. Qualitative data analysis of Twitter.
2. Nearly 1000 tweets in October, November 2012.
3. Evaluation of tweet interventions.
4. Eight separate twitter campaigns carried out.

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

Impact was algorithmically specified in this way: semantic distance between the initiating action, and the net aggregate outcome. This can be statistically tested through the modelling of semantic distances. It can be further visualized and evaluated.

A fundamental aspect of the Twitter analysis was how a tweet, considered as a "campaign initiating tweet", differed from an aggregate set of tweets. The latter was the mean tweet, where the tweets were first mapped into a semantic space. The semantic space is provided by the factor space, which is endowed with a Euclidean metric. For very high dimensions, we find "data piling" or concentration. That is, the cloud of points becomes concentrated in a point. Now that could be of benefit to us, when we are seeking a mean (hence, aggregate) point in a very high dimensional space. A further aspect is when it is shown that the cloud piling or concentration is very much related to the marginal distributions.

Here we show how we can test the statistical significance of effectiveness.

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 space is equal to 3.670904.

Compare that to all pairwise distances of non-initiating tweets. We verified that these distances are normal distributed, with a small number of large distances. By the central limit theorem, for very large numbers of such distances, they will be normal distributed. Denote the mean by μ, and the standard deviation by σ. Mean and standard deviation are defined from distances between all non-initiating tweets, in the full dimensionality semantic (or factor) space. We find $\mu = 12.64907$, $\mu - \sigma = 8.508712$, and $\mu - 2\sigma = 4.368352$.

We find the distance between initiating tweet and mean outcome, for campaign 7, in terms of the mean and standard deviation of tweet distances to be: $\mu - 2.168451\sigma$. Therefore for $z = -2.16$, the campaign 7 effectiveness is significant at the 1.5% level (i.e. $z = -2.16$, in the two-sided case, has 98.5% of the normal distribution greater than it in value).

In the case of campaigns 1, 4, 5, 6, their distances between initiating tweet and mean outcome are less than 90% of all tweet distances. Therefore the effectiveness of these campaigns is in the top 10% which is not greatly effective (compared to campaign 7).

In the case of campaigns 3 and 8, we find their distances to be less than 80% of all tweet distances. So their effectiveness is in the top 20%.

Finally, campaign 2 is the least good fit, relative to initiating tweet and outcome.

5 Tracking Emotion

This relates to determining and tracking emotion in an unsupervised way. This is as opposed to machine learning, like in sentiment analysis, which is supervised. Emotion is understood as a manifestation of the unconscious. Social activity causes emotion to be expressed or manifested. This can lead to later discussion of psychoanalyst, Matte Blanco. See [14].

The foundation of this tracking of emotion, and determining the depth of emotion, is using the methodology of metric space mapping and hierarchical topology. The former here maps the textual data into a Euclidean metric endowed factor space, and the latter may be chronologically constrained hierarchical clustering.

The examples to follow are based on: in the Casablanca movie, dialogue (and dialogue only) between main characters Ilsa and Rick, having selected this dialogue from the scenes with both of these protagonists (scenes 22, 26, 28, 30, 31, 43, 58, 59, 70, 75 and last scene, 77); and chapters 9, 10, 11, 12 of Gustave Flaubert's 19th century novel, Madame Bovary. This concerns the three-way relationship between Emma Bovary, her husband Charles, and her lover Rodolphe Boulanger.

Following [16], in Figure 1 in the full dimensionality factor space, based on all interrelationships of scenes and words, the distance between the word "darling" in this space, was determined with each of the 11 scenes in this space. The same was done for the word "love". The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with boxes.

Then in Figure 2, hierarchical clustering, that is sequence constrained, is carried out on the scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59 70, 75, 77 (using the dialogue, between Ilsa and Rick). See how the big changes in scenes 30 and 70 are indicated in the previous figure.

Now there is consideration of the novel Madame Bovary, with the 3-way interrelationships of Emma Bovary, her husband Charles, and her lover, Rodolphe.

Figure 3 presents an interesting perspective that can be considered relative to the original text. Rodolphe is emotionally scoring over Charles in text segment 1, then again in 3, 4, 5, 6. In text segment 7, Emma is accosted by Captain Binet, giving her qualms of conscience. Charles regains emotional ground with Emma through Emma's father's letter in text segment 10, and Emma's attachment to her daughter, Berthe. Initially the surgery on Hippolyte in text segment 11 draws Emma close to Charles. By text segment 14 Emma is walking out on Charles following the botched surgery. Emma has total disdain for Charles in text segment 15. In text segment 16 Emma is buying gifts for Rodolphe in spite of potentially making Charles indebted. In text segments 17 and 18, Charles' mother is there, with a difficult mother-in-law relationship for Emma. Plans for running away ensue, with pangs of conscience for Emma, and in the final text segment there is Rodolphe refusing to himself to leave with Emma.

In Figure 4, there is display of the evolution of sentiment, expressed by (or proxied by) the terms "kiss", "tenderness", and "happiness". We see that some text segments are more expressive of emotion than are other text segments.

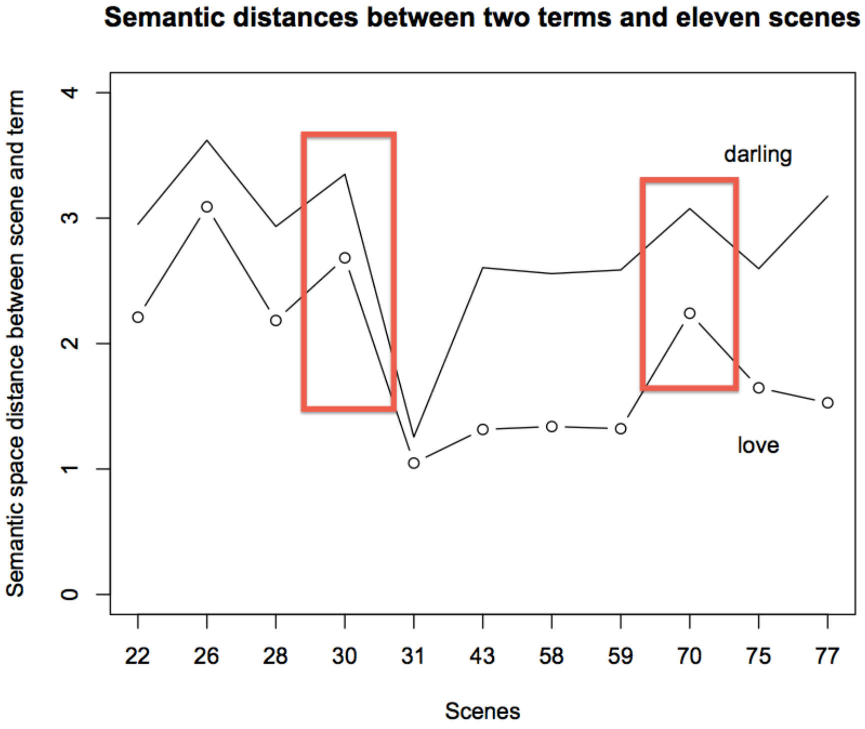

Fig. 1 In the full dimensionality factor space, based on all interrelationships of scenes and words, we determined the distance between the word "darling" in this space, with each of the 11 scenes in this space. We did the same for the word "love". The semantic locations of these two words, relative to the semantic locations of scenes 30 and 70 are highlighted with boxes.

6 Analyses of Mapping of Behavioural or Activity Patterns or Trends

This concerns semantic mapping of Twitter data relating to music, film, theatre, etc. festivals. 75 languages were found to be in use, including Japanese, Arabic and so on, with the majority in Roman script. As indicative association to language, because the labelled language may be partially used or not in fact used, we take the following: English, Spanish, French, Japanese, Portuguese. We consider the days 2015-05-11 to 2016-08-02, with two days removed, due to lack of tweets. The numbers of tweets for these languages were as follows (carried out on 11 August 2016): en, 37681771; es, 9984507; fr, 4503113; ja, 2977159; pt, 3270839

The tweeters and the festivals are as follows. Tweets characterized as French, 4913781 tweets. (For user, date and tweet content, the file size was: 667 MB.) The following were sought in the tweets: Cannes, cannes, CANNES, Avignon, avignon, AVIGNON. Upper and lower case were retained in order to verify semantic prox-

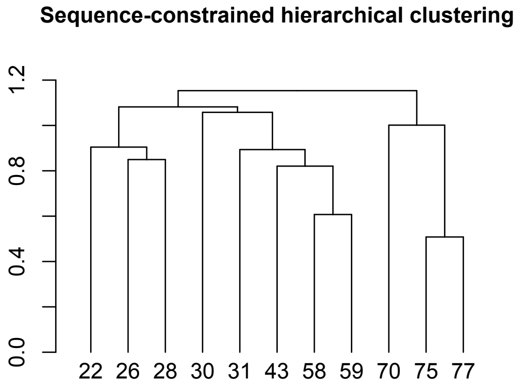

Fig. 2 Hierarchical clustering, that is sequence constrained, of the 11 scenes used, i.e. scenes 22, 26, 28, 30, 31, 43, 58, 59 70, 75, 77 (all with dialogue, and only dialogue, between Ilsa and Rick). Rather than projections on factors, here the correlations (or cosines of angles with factors) are used to directly capture orientation.

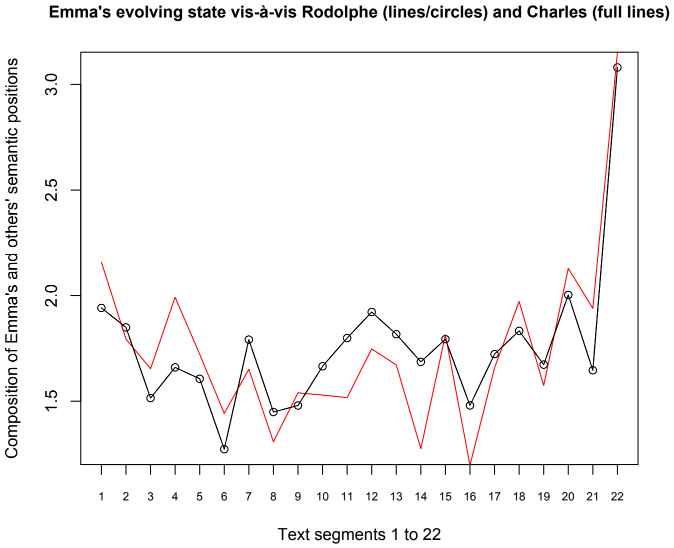

Fig. 3 The relationship of Emma to Rodolphe (lines/circles, black) and to Charles (full line, red) are mapped out. The text segments encapsulate narrative chronology, that maps approximately into a time axis. Low or small values can be viewed as emotional attachment.

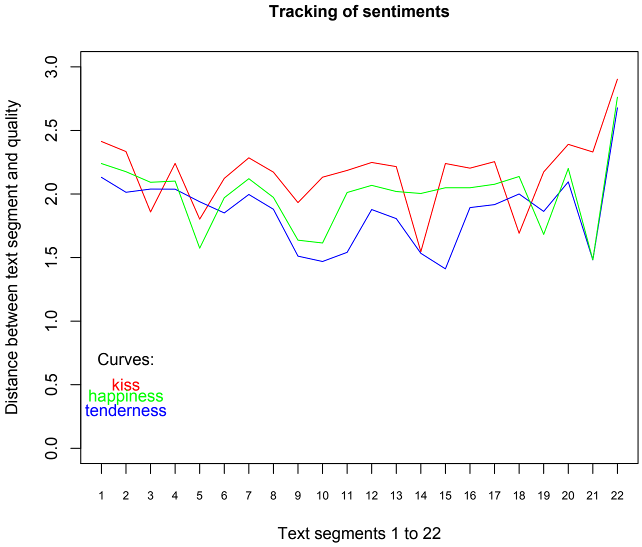

Fig. 4 A low value of the emotion, expressed by the words “kiss”, “happiness” and “tenderness”, implies small distance to the text segment. These curves, “kiss”, “happiness” and “tenderness” start on the upper left on the top, the middle, and the bottom, respectively. The chronology of sentiment tracks the closeness of these different sentimental terms relative to the narrative, represented by the text segment. Terms and text segments are vectors in the semantic, factorial space, and the full dimensionality of this space is used.

imity of these variants. 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, i.e. by a tweeter: Cannes, 1230559 and 3388; cannes, 145939 and 4024; CANNES, 57763 and 829; Avignon, 272812 and 4238; avignon, 39323 and 2909; AVIGNON, 14647 and 900.

The 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%

In Figure 6, mapped are: 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.

Current considerations, relating to approximately 55 million tweets per year (from May 2015), are as follows. 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. Carry out year by year trend analysis.

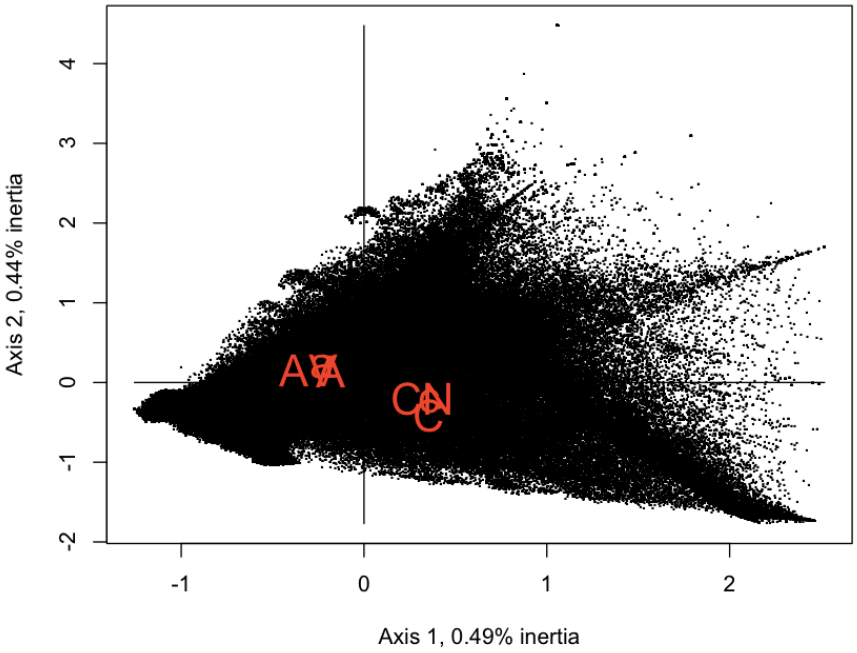

Fig. 5 880664 Twitter tweets projected on the principal factor, i.e. principal axis plane. Attributes are projected.

For further analyses and description of the data, see [4] and [17].

6.1 Baselining or Contextualizing Analysis

The following is in regard to such baselining, i.e. contextualizing, against healthy reference subjects, from a case study in [9]. This repeats some of the description in [13], in regard to testing through statistically baselining or contextualizing in a multivariate manner.

In [20], there is an important methodological development, concerning statistical inference in Geometric Data Analysis, i.e. based on MCA, Multiple Correspondence Analysis. At issue is statistical "typicality of a subcloud with respect to the overall cloud of individuals". Following an excellent review of permutation tests, the data is introduced: 6 numerical variables relating to gait, body movement, related to the following; a reference group of 45 healthy subjects; and a group of 15 Parkinsons illness patients, each before and after drug treatment. [8] (section 11.1) relates to this analysis, of the, in total, 45 + 15 + 15 observation vectors, of subjects between the ages of 60 and 92, of average age 74.

First there is correlation analysis carried out, so that when PCA of standardized variables is carried out, it is the case that the first two axes explain 97% of the variance. Axis 1 is characterized as "performance", and axis 2 is characterized as "style". Then the two sets of, before treatment, and after treatment, 15 Parkinsons patients are input into the analysis as supplementary individuals. [20] is directly addressing statistically the question of effect of treatment. Just as in [8], the healthy subjects are the main individuals, and the treated patients, before and after treatment, are the supplementary individuals. This allows to discuss the subclouds of the before, and of the after treatment individuals, relative to the first, performance, axis, and the second, style, axis. The test statistic, that assesses statistically the effect of medical treatment here, is a permutation-based distributional evaluation of the following statistic. The subcloud's deviations relative to samples of the reference cloud are at issue. The Mahalanobis distance based on covariance structure of the reference cloud is used. The test statistic is the Mahalanobis norm of deviations between subcloud points and the mean point of the reference cloud.

In summary, this exemplifies in a most important way, how supplementary elements and the principal elements are selected and used in practice. The medical treatment context is so very clear in regard to such baselining, i.e. contextualizing, against healthy reference subjects.

7 Conclusion

Much that is at issue here is close to what is under discussion in [3]. The integral association of methodology and application domain will, of course, have shared and common methodological perspectives. However the application of statistical models, and other analytical stages such as feature selection, data aggregation with the various implications of this, and what is often termed data cleaning or data cleansing, all of these issues require analytical focus, and account to be taken of the analytical context. The latter may well include baselining, or benchmarking in an operational manner. In a sense, we might state that combinatorial inference is so paramount because of its applicability.

A good deal of the case studies reported on here made use of preliminary functionality, to be part of the R package, `Xplortext`. This package makes use of these R packages, and add greatly to their functionality: `tm`, `FactoMineR`.

The software system, `SPAD`, is also extending greatly into support for text processing.

References

1. Anderson, C.: "The end of theory: the data deluge makes the scientific method obsolete", Wired Magazine (16 July 2008),
2. Bécue-Bertaut, M., Kostov, B., Morin, A., Naro, G.: "Rhetorical strategy in forensic speeches: Multidimensional statistics-based methodology", Journal of Classification, 31, 85–106 (2014)
3. Gelman, A., Hennig, C.: "Beyond subjective and objective in statistics", *Journal of the Royal Statistical Society Series A*, 180, Part 4, 1–31 (2017).
4. Goeuriot, L., Mothe, J., Mulhem, P., Murtagh, F., SanJuan, E.: "Overview of the CLEF 2016 Cultural Micro-blog Contextualization Workshop". In: Editors: N. Fuhr, P. Quaresma, T. Goncalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 7th International Conference of the CLEF Association, CLEF 2016, vora, Portugal, September 5-8, 2016, Proceedings, Lecture Notes in Computer Science, volume 9822, pp. 371–378 (2016)
5. Hernández, D.M., Bécue-Bertuat, M., Barahona, I.: "How scientific literature has been evolving over the time? A novel statistical approach using tracking verbal-based methods", *JSM Proceedings, 2014, Section on Statistical Learning and Data Mining*, American Statistical Association, 1121–1132 (2014)
6. Keiding, N., Louis, T.A.: "Perils and potentials of self-selected entry to epidemiological studies and surveys", *Journal of the Royal Statistical Society A*, 179, Part 2, 319–376 (2016)
7. Legendre, P., Legendre, L.: *Numerical Ecology*, 3rd edn., Elsevier, Amsterdam (2012)
8. Le Roux, B.: *Analyse Géométrique des Données Multidimensionelles*, Dunod, Paris (2014)
9. Le Roux, R., Rouanet, H.: *Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis*, Kluwer, Dordrecht (2004)
10. Le Roux, B., Lebaron, F.: "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)
11. McKee, R.: *Story: Substance, Structure, Style, and the Principles of Screenwriting*, Methuen (1999)
12. Murtagh, F.: *Multidimensional Clustering Algorithms*, Physica-Verlag, Würzburg (1985)
13. Murtagh, F.: "Contextualizing Geometric Data Analysis and related data analytics: A virtual microscope for Big Data analytics", JIMIS, submitted (2017)
14. Murtagh, F.: *Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics*, Chapman and Hall, CRC Press (2017)
15. Murtagh, F., Ganz, A., McKie, S.: "The structure of narrative: the case of film scripts", *Pattern Recognition*, 42, 302–312 (2009)
16. Murtagh, F., Ganz, A.: "Pattern recognition in narrative: Tracking emotional expression in context", *Journal of Data Mining and Digital Humanities*, vol. 2015 (published May 26, 2015).
17. Murtagh, F.: "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-1609/16091207.pdf
18. Murtagh, F., Pianosi, M., Bull, R.: "Semantic mapping of discourse and activity, using Habermas's Theory of Communicative Action to analyze process", *Quality and Quantity*, 50(4), 1675–1694 (2016
19. Murtagh, F., Orlov, M., Mirkin, B.: "Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research", *Journal of Classification* (in press, 2017). Preprint: https://arxiv.org/abs/1607.03200
20. Bienaise, S., Le Roux, B.: "Combinatorial typicality test in Geometric typicality test in geometric data analysis", preprint (2016)
21. Wessel, M.: "You dont need Big Data – You need the right data". *Harvard Business Review* (3 Nov. 2016). https://hbr.org/2016/11/you-dont-need-big-data-you-need-the-right-data.