doi.org/10.26398/IJAS.0029-013

CORRESPONDENCE ANALYSIS OF MASSIVE DATA: KEY ROLE OF RESOLUTION SCALE OF THE ANALYTICS

Fionn Murtagh[1]

School of Computing & Engineering, University of Huddersfield, Huddersfield, UK

Abstract: *Resolution scales of analytics can be coupled in the following way. The principal or main analysis is carried out on a low resolution data encoding. This simply expresses that data are aggregated. We consider where data aggregation is interpretationally of value; and where it is of computational benefit. Once the principal or main analysis is carried out, using supplementary elements, we can locate or map rows or columns, i.e. individuals, or attributes or attribute modalities, in the semantic, factor space. From the overall perspectives we proceed to address specific aspects of our data. While the above provides focus in analytical processing, what is so very important also is context. Here to be described is how context gives rise to qualitative as well as quantitative effectiveness and impact assessment.*

Keywords: *Big Data, High dimensionality, Semantics, Ontology, Twitter social media*

1. INTRODUCTION

The resolution scale of observation and consequent analyses are quite fundamental in many domains. One example of a domain where such computational methodology is prominent is image and signal processing, employing multiresolution transforms, that include the wavelet transform, the curvelet transform, and many related transforms. In this article, we wish to demonstrate such perspectives, in general, in data analytical processes. In analytics, resolution scale can aid with computational time performance, and with interpretation and impactful outcomes. However, in regard to this, there are also ethical aspects of data analytics to be considered. The many cases that are described in O'Neill (2016) are effectively due to the resolution scale of the data processing. In Le Roux and Lebaron (2015), it is noted how Correspondence Analysis (CA) in the general context of Geometric Data Analysis (GDA) is most appropriate for: "Rehabilitation of individuals. The context model is always formulated at the individual level, being opposed there-fore to modelling at an aggregate level for which the individuals are only an 'error term' of the model."

[1] Corresponding author: Fionn Murtagh, email: fmurtagh@acm.org

Two related issues are central in this work: firstly, relating our analytics to contextual information and data; and secondly, relating high resolution data and information to the principal, baseline, analysis that is carried out on low resolution data and information.

The former of these themes, viz. relating our analytics to contextual information and data, is our starting point. We seek to preserve contextual, and hence semantic, properties of data analysis outcomes. This is in order to make use of aggregated data and information at varying resolution scales. We must take into account the geometry and topology of data and information.

An application study based on the semantics of aggregated data is as fol lows. We study the semantic mapping of communicative processes. We want to map out qualitative aspects of such activity (or event) processes. We semantically map a Twitter discourse, using the CA platform. Our case study is a set of eight carefully planned Twitter campaigns relating to environmental issues. The aim of these campaigns was to increase environmental awareness and behaviour. Each campaign was launched by an initiating tweet. The semantic distance between an initiating act and the aggregate semantic outcome is used as a measure of process effectiveness.

The aggregate semantic outcome can follow also the need for process efficiency. An increasingly widely used principle for processing large data volumes is to allow for distributed computational processing work. This processing becomes therefore, the mapping stage. Data is defined, or structured, as key-value pairs. Then in the reduce or combine stage, the distributed outcomes are aggregated. This approach to processing data, that may include some stages of the analytics, is core to the widely used software environment, Apache Hadoop that supports distributed processing, and includes MapReduce for its key-value pairbased processing. As briefly described, this has become a prominent contemporary approach to analytical processing. In the case studies in this article, we aim to add some further important perspectives in contemporary analytics. The two fundamental concepts that motivate and inspire this work are scale and context.

We aim to take into account fully the important underlying and underpinning semantic structures. There is piling or concentration of data, with increase in dimensionality (cf. Hall et al. (2005)) and this can be of major benefit for our analytics. The particular benefit is observed for CA of random projection based orthonormal mapping, or scaling, (Critchley and Heiser, 1988), of power law distributed data that are found in many domains (Murtagh and Contreras, 2015).

Motivation for this work includes the CA of an infinite (unbounded) number of rows or observations, crossed by 1000 attributes, discussed in Benzécri (1982). Other relevant cases of use of CA for massive data sets, using, for example, iterative solution of the eigen-decomposition, are in Benzécri (1997); Lebart et al. (1984);

Murtagh (1996). While it is so very evident that we can handle massive data volumes, the essential question becomes: why, when or how is this useful?

In section 2, aggregating our data for interpretation-related objectives is at issue. In this section, some general background description is provided. In section 3, aggregation in general is shown, and how this lends itself very well to carrying out the main analysis at lower resolution, but continuing in the analytics to fine, high resolution, semantic mapping. Twitter data is used as an example. In section 4, the potential for such aggregation to be used for the analytical resolution is at issue. A small example is described of discipline-related national science research funding.

2. DATA AGGREGATION FOR QUANTIFYING EFFECTIVENESS, USING AGGREGATE OUTCOME, OF A HUMAN OR MACHINE ACTION

In this first case study, we look at the computational requirements of the processing. Our application is where data aggregation has a relevant interpretation.

Traditionally the "impact" of an action is considered in computational terms as what happens when an algorithm is executed. In general, an algorithm is a chain of processing actions. Now consider some specified action, with some desired or targeted outcome. In order to express and model (mathematically, computationally) general human or social, or other, scenarios, we will relax what we take as "impact", to instead be "effectiveness". We define effectiveness as the general, and hence aggregate, outcome. In the space of all actions, our initial action will be a point. Then all the actions considered are all points in the space of actions. Together, they comprise a cloud of points. Finally, the aggregate outcome of all these actions is the centroid (mean, centre of gravity) of the cloud of actions.

The initiating tweets are taken as "instigational" actions and we examine them relative to an aggregate outcome. The latter is an average profile. In that work, the actions were tweets (so-called micro-blogs), in a Twitter context. We were studying the process of communicative action, with Jürgen Habermas' socialpolitical theory of communicative action as motivation for that work. We used successive Twitter campaigns (relating to environmental citizenship, i.e. socially and personally good environmental practice and behaviour). We wanted to see how well an initiating tweet would be matched against an overall campaign average. We used a semantic embedding of all of our tweets that were studied. This semantic embedding was based on the textual content of the tweets. CA provides such a latent semantic embedding.

Here we wish to use this case study as a nice example for (i) both qualitatively through the semantics of the tweet content, and quantitatively too, even to the stage of specifying a statistical hypothesis test, assessing impact and effectiveness of particular actions; and (ii) how the set-up that unfolds here, based on the Euclidean metric endowed CA factor space, i.e. the factor space embedding or mapping, this supports what we may refer to as generalization and contextualization, for further, related data and information.

The almost one thousand tweets were from eight weeks, each week being an experimental "campaign", that was started with a provoking or instigation tweet. These initiating tweets are listed in Murtagh et al. (2016). The category expressing the campaign that each tweet was associated with led to these categories, the experimental campaigns, being supplementary attributes. The principal attributes were the selected word corpus. Then in regard to the tweets, since the initiating tweets were expected to be effective in instigating, or initiating, resulting and consequent weets, the eight initiating tweets, for the eight experimental campaigns, were considered as supplementary rows.

The corpus, i.e. the set of words or terms, was determined, starting with adjacent character strings, then with deletion of accented characters and nonalphabetic symbols, deletion of punctuation, and setting all upper case characters to lower case characters.

In Figure 1, the principal factor plane is shown, with arrows pointing from the tweet initiating a campaign, e.g. tic1 for campaign 1, to the centre of gravity of all tweets other than the initiating tweet, in a given experimental campaign.

Effectiveness or impact was considered as follows: proximity between the initiating tweet and the net outcome. While the original data, comprising all tweets, consisted of 985 tweets, the corpus, to start with, was of cardinality 3056. Through selection from the corpus, including requiring that terms be shared by a few tweets, it resulted that there were 339 sufficiently often used terms. This further had the consequence that the non-initiating tweet set dropped to 968 nonempty tweets.

The principal factor plane in Figure 1 accounts for a relatively small percentage of inertia. In Murtagh et al. (2016), there is the consideration of impactful tweets, seen in Figure 1, especially campaigns 3, 5, 8. That is because of the initiating tweet being close to the net effect of that tweet. An alternative, and more well-based perspective is to consider the full factor space dimensionality. In Murtagh et al. (2016), it is described how campaign 7 is the best here. Then, looked at was the distribution of inter-tweet distances. Being Gaussian-distributed, that led to an approach for testing the significance of the most impactful campaign, here campaign 7.

This work encompasses the following: the semantics of the tweet content (contrast this with the standard approaches that only use quantity of tweets, and retweets, and networks of vertices being hashtags); impact or effect of action, in the

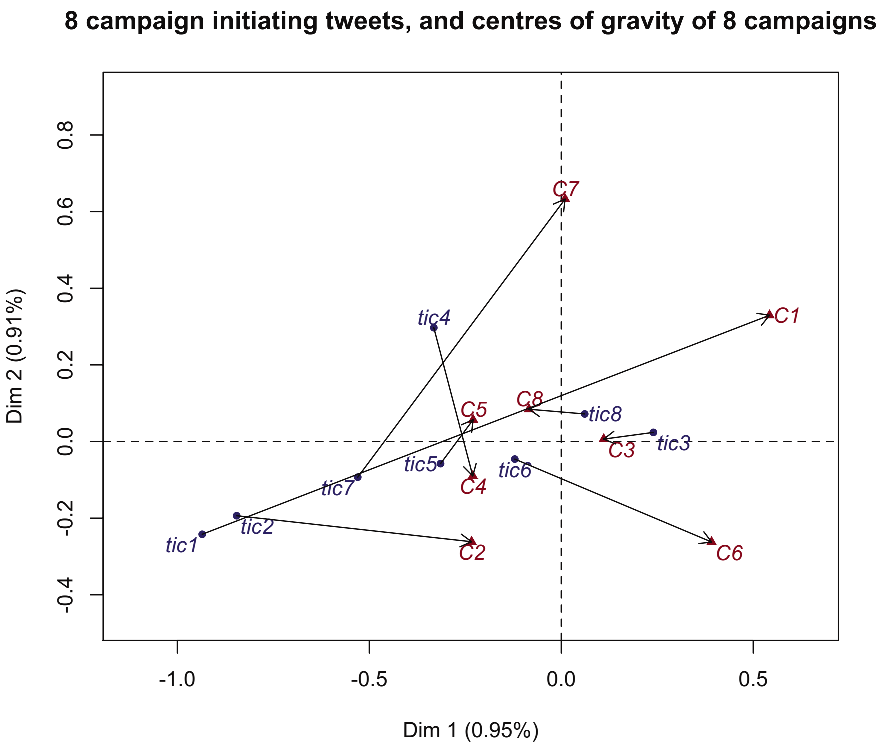


Figure 1: The campaign initiating tweets are labelled "tic1" to "tic8". The centres of gravity of the campaigns, i.e. the net aggregate of the campaigns, are labelled "C1" to "C8". In each case, the tweet initiating the campaign is linked with an arrow to the net aggregate of the campaign. The percentage inertia explained by the factors, "Dim 1" being factor 1, and "Dim 2" being factor 2, is noted.

manner described above (i.e. distance between instigation or initiation, relative to net outcome); availing of full dimensionality of the Euclidean metric endowed factor space, of the cloud of tweets, and the word corpus cloud; and intuitively having both visualization and statistical modelling straightforwardly supported.

Having both the mapping of very particular elements and having the mapping also for net effect, this can lead to further mapping of new cases or perhaps even newly arising, cases of exceptional interest. We expect to pursue such mapping, with where beneficial, statistical hypothesis testing (and this could well be in regard to application domains of forensics and security, or recommender systems, or behavioural analytics in such burgeoning fields as smart cities, and Internet of Things). The following is a most important consideration here: in CA, there is invariance between the chi squared metric endowed dual data clouds, that then becomes mapped into the Euclidean metric endowed dual data clouds. That is to say, there is invariance between the chi squared distances on the individual or observation (or row) set, and the attribute (or column set), on the one hand, and on the other hand, the Euclidean distance in the factor space. We could draw the following inspiring and rewarding conclusions: carry out the CA to both qualitatively and quantitatively structure our information space; then set up all manner of linkage with new data sources, using for that the easily and directly applicable chi squared distance.

A potentially important point follows from this, for our analytical processing chain, for example in dynamic and evolving contexts. We can establish our well-structured CA-based contextual framework. This can be followed by linear computational time integration of newer data. As noted above, important domains of application to have such insightful analytics for streaming data include smart cities and Internet of Things.

3. ANALYZING MASSIVE DATA SETS WITH CORRESPONDENCE ANALYSIS: DATA AGGREGATION, RESOLUTION LEVEL OF ANALYSIS

CA is based on dual spaces endowed with the chi squared metric, and these dual clouds are mapped into a dual space that is endowed with the Euclidean metric. In this framework, at issue are profiles of individuals or observations (i.e., rows) and of attributes (encompassing modalities of response, variables in complete disjunctive form and fuzzy coding and other forms of coding). The principle of distributional equivalence, related to the aggregrating of similar profiles, leads to the following informal remark: aggregating similar or identical profiles is welcome. Aggregation, when profiles are identical or nearly identical, is a very desirable processing action, in particular for interpretational reasons.

3.1 RESOLUTION LEVEL OF THE ANALYSIS CARRIED OUT

Very often in analytical frameworks, and quite generally across lots of domains, there is a role to be played by the ontology or taxonomy (i.e., concept hierarchy) that encapsulates the general context that is at issue. If there is focus of interest in the set of attributes related to the upper level of a taxonomy of the attributes, then it is very clear that the number of such attributes leads to a lower dimensional Euclidean distance endowed, factor space mapping. Furthermore, attributes that are at a finer resolution level, i.e. at a lower level in the taxonomy, or concept hierarchy, these can be retained as supplementary elements. These finer resolution level attributes can be projected into the analysis. By having our CA factor space mapping based on the lower resolution, and hence top level taxonomy data, we can state that this is a means towards information focusing in our analytics. We may even state that this is analytical focusing.

With due and appropriate justification, we can implement such information focusing as follows. We define a small number of aggregates of either observations or attributes, and we carry out the analysis on them. Then we project the full set of observations and attributes into the factor space. A benefit of this, for interpretation, is that a low-dimensional factor space, i.e. using just the first, second, etc. principal axes, is likely to be a very good approximation to the inertia of the clouds of observations, or, identically, of the cloud of attributes. As noted above, benefiting from such information focusing is especially relevant when our data is taken as coming from, or being associated with, an ontology or concept hierarchy. An example of this will now follow.

3.2 LOW RESOLUTION ANALYTICS WITH HIGH RESOLUTION MAPPING

The Twitter analytics had about 12 million tweets. We chose the daily sets of tweets as a most useful main or principal resolution level to begin the analytics. We also extracted, in order to project following the principal analysis, the hashtags (terms preceded by the hash character, #), tweeter names (terms preceded by the @ character), URLs (web addresses, always in abbreviated format). We retained 5820 terms. Our data set comprised 233 days of our tweets, relating to festivals, including the Cannes Film Festival, Fèis íle, Islay (Scotland), the Berlin Film Festival, CMA, Country Music Association, the Avignon Theatre Festival, and the (disputed) Yulin Dog Meat Festival.

Figure 2 displays the first and second planes, i.e. formed by factors 1,2 and by factors 3,4. Factors 1 and 2 are associated very much with the Cannes Film Festival, and with the Avignon Theatre Festival, and factors 3 and 4 with the latter. The analysis was carried out on the 233 days crossed by 5820 terms, including the main word in the festival title. What is also displayed in Figure 2 is just one example, here, of where on 18 August, later than these festivals, there was Twitter dialogue. Similarly even an individual tweet or any other available contextual attribute can be projected into the factor space.

This example illustrates also an ethical aspect of Big Data analytics. This is following the noting in Section 1 of the need for “Rehabilitation of individuals”. We emphasize therefore how aggregation of data in our work is not simplifying the analysis from the interpretation point of view. Rather it is providing the baseline, which fully allows high resolution semantic mapping of individuals, attribute modalities, and any and all combinations of these.

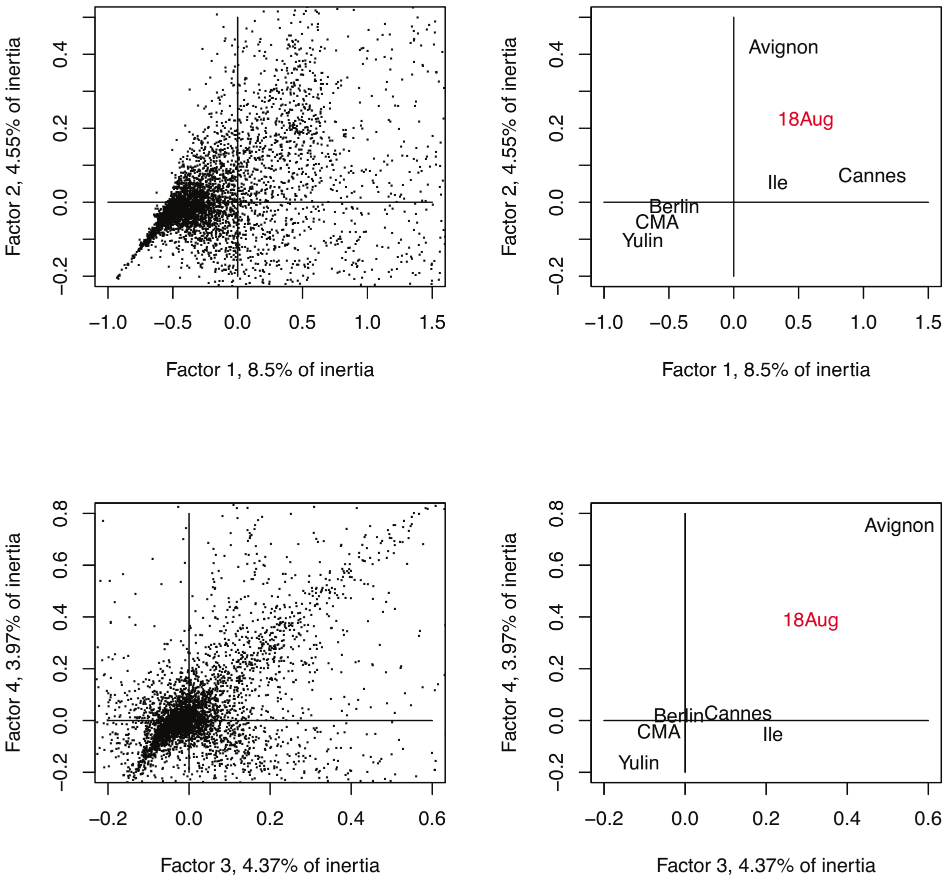


Figure 2: The principal factor plane, in the top two panels, and the plane of factors 3,4 in the bottom two panels. The left panels display all words, with a dot at each word location. The right panels display the selected festivals.

4. THROUGH AGGREGATION: RESOLUTION SCALE OF THE ANALYSIS

4.1 RESOLUTION SCALE OF ANALYSIS

In this section, a small example is used to explain and illustrate how aggregation-based analytics is a very practical approach for scaling up our analytics. Such scaling up is both the motivation and justification for our principal analysis to be carried out at lower or higher resolution scales. Finer, higher resolution issues are easily addressed through use of supplementary elements.

A future analytics objective of ours is qualitative evaluation of research impact. Quantitative measures are also to be considered, possibly, though, as supplementary attributes. The following subsections are relating to an initial, and quite general, assessment of national research funding.

National research funding impact will typically have a range of quantitative measures such as number of company start-ups, number of completed PhDs and, of course, all that is related to publications, and citation counts. Increasingly noted also are social media manifestations or commentary, e.g. using Twitter microblogging, with consequent counts, cf. Casey et al. (2016). The work at issue in this section is also very relevant for journal editorial work, and related scholarly and archival publishing. This is to be a major focus of our work, based on our editorial roles, as well as previously directing national research funding agency work. The most immediate and direct implications of the short accounting for resolution scale in analytics will be used in our journal editorial activities. This will be based on the comprehensive approach, described in Murtagh et al. (2017), for qualitative analysis, together with its quantitative relations, of content. A very new planned analytics activity is to qualitatively and quantitatively analyse the content of research themes that both reach a successful outcome, and those that attempted and proposed, but that do not, apparently, reach a successful outcome.

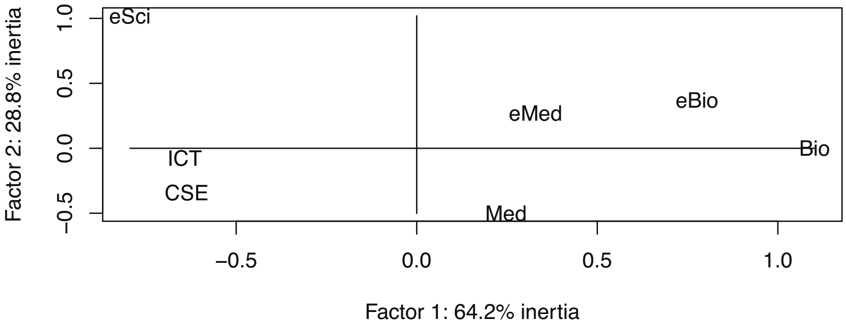
4.2 SMALL EXAMPLE OF A TAXONOMY OF ATTRIBUTES

Using national funding agency data, where the author was a director, 20 major funded projects are considered. These were CSETs and SRCs, respectively Centres for Science, Engineering and Technology, and Strategic Research Clusters. Respectively these were funded at about € 20 and 7.5 million.

A set of theme areas, and the host institutions, were available for each one of these projects. With frequency of occurrence in these characterisations, the first 8 of these terms were as follows: `physsystems 10, logsystems 6, body 5, mobile 5, cs 4, disease 4, health 4, sensors 4`. Abbreviations used in these are: physical systems, logical systems, computer science. Host institutions were: UCD, TCD, UCC, TNI (Tyndall National Institute, located in UCC, University College Cork), UL, NUIG, NUIM. That lists 6 of the 7 universities in Ireland. The final 4 of the terms were as follows, with an abbreviation here for “software engineering”: `se 1, semanticweb 1, systems 1, telecoms 1, transmission 1, vaccines 1.`

Motivation is as follows: to use keyword or associated characterisation of the objects, here the research centres; then to use conceptual terms based on the initially given terms.

From the data for the 20 funded projects, i.e. research centres, we retain the following for analysis: 5 main themes of interest, `physsystems, logsystems, body, health, datainformation,` where again it is noted that there are compound and abbreviated expressions: physical systems, logical systems, data and information. So while these are of interest, we lower the resolution level of our analysis by aggregating, i.e. summing values, for each one of the 20 funded projects. Rewriting, for convenience, our 5 main themes of interest, as `Phys, Log, Body, Health, Data,` we define the lower resolution terms, or terms that are higher in what can be taken as a taxonomy: e-Science, Biotechnology, Computer and Software Engineering, Medical, Information and Computing Technologies, e-Medicine (digital medicine), and e-Biology (digital biology). Aggregating our terms, we have: `Log,Data = eSci; Body, Health = Bio; Phys,Log = CSE; Body,Health, Phys = Med; Phys,Log,Data = ICT; Body,Health,Log = eMed; Body,Health,Data = eBio.`



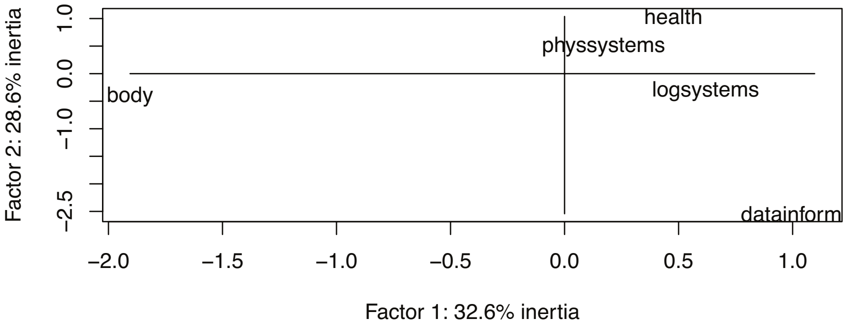


Figure 3: Top: 7 lower resolution aggregated terms; bottom: 5 basic terms used. For both, 20 research centres are crossed by the term set.

Considering now our basic themes, and our higher taxonomic level themes, the principal factor plane is shown in Figure 3. Clearly enough, both axes are reversed in their semantics, in these two principal factor planes. Cf. how `Bio, eBio` are positively on the first axis, and `body` is negatively projected on the first axis in the lower plot. In Figure 4, the host institutes are projected as supplementary elements. In regard to interpretation, this would be motivated by seeing what and where there is specialisation. Further contextualization can follow, e.g. amount of the funding, year of funding when appropriate, that will lead to trend analyses. Also publication counts, company start-ups and other measures can be contextually located in the semantic, factorial space.

In order to further assess the similarity or comparability of host institutes, we will check their relations through hierarchical clustering. Again it is to be noted how

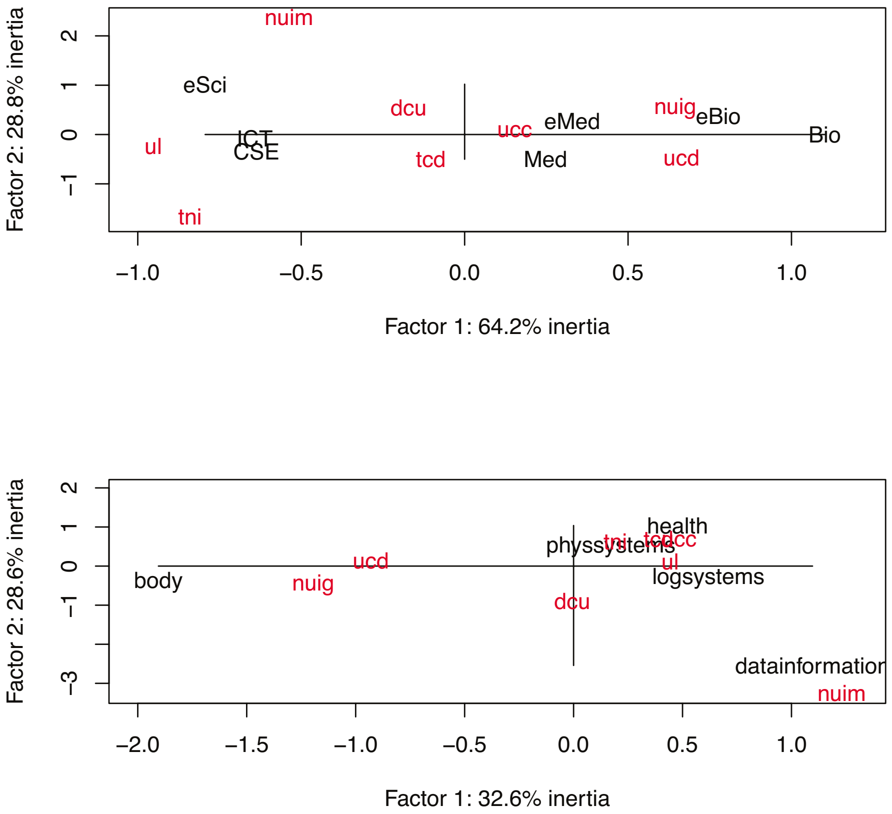
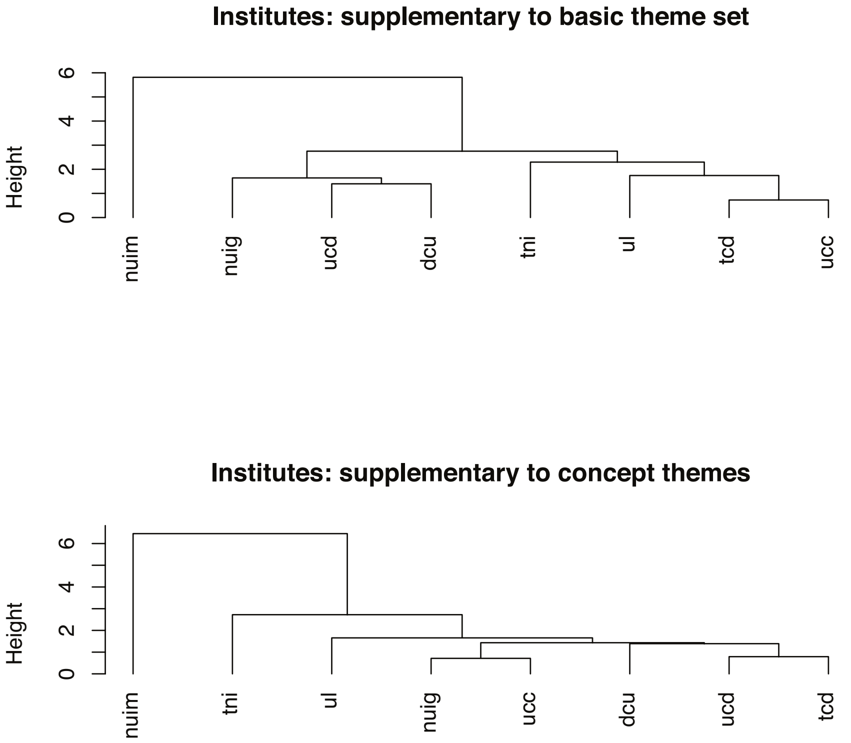


Figure 4: As Figure 3, top and bottom biplots, with the host institutes as supplementary elements.


Institutes: supplementary to basic theme set

Institutes: supplementary to concept themes

Figure 5: Hierarchical clustering, minimum variance or Ward method used, from the full factor spaces, for which the principal factor planes were displayed in the previous figures. Respectively, the top and bottom plots are associated with the top and bottom plots in Figure 4, but note that full dimensionality factor space coordinates are at issue here.

this small case study is used for illustrative and expository purposes. Figure 5 shows the hierarchies. The full factor spaces were of dimensions, respectively, 4 and 6, given the inputs: 20 research centres crossed by either 5 basic terms, or 7 higher taxonomic level terms. So the factor space dimensions were, respectively, 4 and 6. Eigenvalue percentage rates are, respectively, 32.6, 28.6, 22.1, 16.7, and 64.2, 28.8, 6.7, 0, 0, 0.

Just how similar the hierarchical clusterings are, shown in Figure 5, is addressed as follows. The cophenetic, or ultrametric, distances are obtained. The correlation between the host institutes' ultrametric distances, obtained from the two dendrograms, is 0.925. We may consider this outcome to be satisfactory and to be motivated, fundamentally, by the interpretational value, and potential benefits for policy-making, of the taxonomic level or resolution scale that will be primarily retained. Here, this would be the conceptual categorization using the terms or concepts: e-Science, Biotechnology, Computer and Software Engineering, Medical, Information and Computing Technologies, e-Medicine (digital medicine), and e-Biology (digital biology).

4.3 LOW RESOLUTION ANALYTICS BASELINING FOR BIG DATA

Our aim has been to develop taxonomic relations, and to show, using a small case study, just how approximation in regard to certain tasks, can be noted whenever it is to our benefit, from the interpretation point of view, primarily, as well as with computational benefit.

Major benefits are likely to come about in such areas as quality assessment of research outputs, rather than just quantitative, citation-based, assessment alone. In our work in Murtagh et al. (2017), we present the case for stratification rather than basic ranking. Secondly, the qualitative analytics are based on a taxonomy, appropriate for a given discipline or domain, and quite possibly generated and maintained through open dialogue. Qualitative analytics that are open and transparent, and changeable and dynamic, have implications in regard to expertise and ethics.

Further potential is as follows. In Keiding and Louis (2016), our contribution to this milestone work includes the following. It is noted that there is need for the “formulation of abstract laws” that bridge sampled data, subject to bias effects through selectivity, and calibrating Big Data. This can be addressed, for the data analyst and for the application specialist, as geometric and topological. The bridge between the data that are analysed 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. Eminent quantitative and qualitative sociologist Pierre Bourdieu’s concept of field is a prime exemplar.

5. CONCLUSION

As indicated by Jean-Paul Benzécri in the following remark in 2011, the contemporary climate of Big Data analytics is both motivating and it is all very nicely in tune with the theory and practice of CA.

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

Clearly the CA platform, and associated GDA, and statistical and mathematical methods, are most appropriate for obtaining patterns and trends, and all manner of information from data. In addition to the case studies in this article, the following are major application domains in Murtagh (2017): cinema, literature, cosmology and psychoanalysis. Many other application domains certainly are of great relevance and importance, and will also come to the fore. These include mental health, security, lifestyle, and more.

REFERENCES

Benzécri, J.P. (1982). L'approximation stochastique en analyse des correspondances. In *Les Cahiers de l'Analyse des Données,* 7 (4): 387–394.

Benzécri, J.P. (1997). Approximation stochastique, réseaux de neurones et analyse des données. In *Les Cahiers de l'Analyse des Données,* 22 (2): 211–220.

Casey, A., Ahmadi, S. and Murtagh, F. (2016). Exploring the relationship article level metrics with linguistic patterns and quantities. In *Archives of Data Science, KIT Scientific Publishing.* Proceedings ECDA 2015, European Conference on Data Analysis, forthcoming.

Critchley, F. and Heiser, W. (1988). Hierarchical trees can be perfectly scaled in one dimension. In *Journal of Classification,* 5: 5–20.

Hall, P., Marron, J. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. In Journal of the Royal Statistical Society Series B, 67: 427–444.

Keiding, N. and Louis, T. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. In *Journal of the Royal Statistical Society Series A,* 179: 319–376.

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

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis.* Wiley. (Chapter 6, Direct Reading Algorithms).

Murtagh, F. (1996). Application de l'analyse factorielle et de l'analyse discriminante à des données colligées pour être soumises à des réseaux de cellules. In *Les Cahiers de l'Analyse des Données,* 21 (1): 53–74.

Murtagh, F. (2017). *Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics.* Chapman & Hall, CRC Press. Accompanying data and software web site: http://www.DataScienceGeometryTopology.info.

Murtagh, F. and Contreras, P. (2015). Random projection towards the Baire metric for high dimensional clustering. In A. Gammerman, V. Vovk, and H. Papadopoulos, eds., *Proceedings SLDS 2015, Symposium on Learning and Data Sciences, Lecture Notes in Artificial Intelligence Volume 9047,* 424–431. Springer, Heidelberg.

Murtagh, F., Orlov, M. and Mirkin, B. (2017). Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research. In *Journal of Classification.* In press. Preprint: https://arxiv.org/abs/1607.03200.

Murtagh, F., Pianosi, M. and Bull, R. (2016). Semantic mapping of discourse and activity, using Habermas's theory of communicative action to analyze process. In *Quality and Quantity,* 50 (4): 1675–1694.

O'Neill, C. (2016). *Weapons of Math Destruction.* Crown/Archetype.