Universal Journal of Computational Analysis **9 (2021), 30-41**
www.papersciences.com **EU Academy of Sciences (EUAS) e-Symposium**

Clustering Exceptionality and Commonality: DivisiveTernary Hierarchical Clustering

Fionn Murtagh
University of Huddersfield
United Kingdom
fmurtagh@acm.org

Abstract
Textual data analysis is innovatively pursued here by having a hierarchical clustering method that distinguishes a typical and exceptional relationships and hence the cluster memberships determined in that way. What is termed divisive ternary hierarchical clustering is of linear computational time, and therefore advantageous for analyzing large data sets. For data sources that are mostly qualitative data, being frequency of occurrence data or, for example, Likert scale response categories, for example, in questionnaires, Correspondence Analysis maps the data clouds onto the Euclidean-metric endowed factor space, that also can be termed semantic space. It is shown how in the factor space, using p-adic, with p prime, representation, here with p=3, representation can lead naturally to hierarchical clustering. From this, here by focusing the analysis on what is defined as typical in the factor space, and, in particular, what is far less typical in the data, these are to be core themes here for pattern recognition and data mining. This is an innovative approach with clear interpretation oriented objectives, for data science, here text mining.

Key Word and Phrases
Correspondence Analysis, Geometric Data Analysis, Semantic Mapping, Ultrametric Topology, Qualitative and Quantitative Analysis

1. Introduction

Here there will be for interpretation and all that follows from the analysis, the learning of partitions derived from the hierarchical clustering, hence ultrametric topology, with that derived from the Euclidean metric endowed geometric factor space, and that derived from the chi-squared metric endowed clouds of qualitative and possibly also quantitative data. The smaller of the two clouds, i.e. column set or row set, is computationally at issue for the singular value decomposition, determining the factor space from the obtained eigenvalues and eigenvectors. Implementation is fully described in Murtagh [8] and terming Correspondence Analysis as Geometric Data Analysis is described in Le Roux and Rouanet [7]. If there are n rows, and m columns, the dimensionality of the factor space is $\min(n, m) - 1$, i.e. this is the number of factors. In practice, for interpretation, active column variables are determined, and these may be modalities of response, in a survey or questionnaire, and perhaps there can be also active rows, and these are used to determine the factor space, with their mappings into the factor space. So the active variables, and possibly also the active rows, these are the conceptual data that determine the factor space, which can also be expressed as the semantic space. Then all other variables, and possibly all other rows, these can be mapped in the factor space. These variables are called supplementary variables, and that can be very relevant for contextual data sources. For example, Murtagh and Farid [16] analyze surveying of mental health with the factor space having medical active variables and with supplementary variables being gender, age, work, and such demographic variables.

For the new hierarchical clustering methods, major themes in this paper include, firstly, distinction between very extensive commonality for all the data mapped into the factor space, relative to all that is separately located from other data mapped into the factor space, and then how what are very related to the factors, how they are to be well handled in the clustering. In addition for this methodology, there is the relevance of ternary number encoding. Historically, while binary and digital computing was developed, there was also in the Soviet Union in the

1950s, ternary computing was developed. An interesting description of ternary data coding is in Stakhov [19], although that work is essentially oriented towards Fibonacci number theory.

Another major theme in this paper is that the outliers in the data, here mapped into the Correspondence Analysis factor space, they may be both strong and important exceptionality in the data. How data outliers are to be handled in clustering is a major theme in Kuwil et al. [6].

There is the searching for clusters based on a ternary, 3-adic, data encoding. Starting with the mapping of data into a Euclidean-metric endowed factor space, the following is predominant in the interpretation of the factor space: what projections are on the positive factor axis, what projections are on the negative factor axis, and what projections are close to the origin, i.e. close to zero on the factor axis. Associated with an interpretative characterization of the most important factors, if not all, positive and negative projections, and near origin projections, all are quite fundamental for pursuing the analytical interpretation. A ternary hierarchy, an alternative to standard binary hierarchy resulting from agglomerative hierarchical clustering, can be constructed. However if the data under investigation is very large, then instead of displaying the ternary hierarchy, in this paper it is intended to develop a useful way, to determine the clustering, that is computationally efficient, generally to be of linear computational complexity. The ternary hierarchy would be a divisive hierarchical clustering, with each level of the hierarchy determined from the decreasing sequence, eigenvalue-based and hence percentage inertia based, of factors.

It may be noted that the data under analysis here is considered as being text data, or other categorical data and with quantitative attributes. The Correspondence Analysis endows categorical data with the chi squared metric, on both the data clouds associated with the row observations or individuals, etc., and with the column attributes, or variables, or modalities of the attributes. From that there is mapping in the Euclidean metric endowed factor space.

The clustering is related to, and in effect derived from, the ternary hierarchical clustering. To be noted is that the clustering criterion is to have identity of the ternary encoding properties. What is mapped very close to the origin is to be considered as common and shared, while what is mapped either positively or negatively on the factor axis is quite specific. Since we are dealing with a set of factors, and perhaps all, pairwise identity of the ternary coding, over many factors, this is termed being exceptional, here.

While the ultimate aim of a great deal of data analytics is to have clusters formed and studied, some open questions may need to be: to define the dissimilarity or distance measure to use; then to define the cluster optimization criterion, or the hierarchical agglomerative clustering criterion. This provides both motivation and justification for the following, and here this follows Murtagh [12]. Our approach here is to assume a fac- tor or principal component space, thoroughly taking semantic relationships into account, and that is endowed with the Euclidean metric. For original data that is comprised of categorical (qualitative) and quantitative values, Correspondence Analysis is most suitable.

Since the factor space is constructed through eigenvalue, eigenvector decomposition of the source data, it follows that if the number or rows, $n >> m$, the latter here being the number of columns, and these may be the active variables, then the computational requirement is for $O(m^3)$ processing time. This is likely to be achievable for Big Data sources.

A particular benefit of Correspondence Analysis is its suitability for carrying out an orthonormal mapping, or scaling, of power law distributed data. Power law distributed data are found in many domains. Correspondence factor analysis provides a latent semantic or principal axes mapping. Much is described in Murtagh [11].

This paper takes Murtagh and Iurato [18] further, in particular towards the orientation of the analysis being undertaken, that is semantic interpretation of clusters derived now from the divisive ternary hierarchical clustering. In section 2, there is how clusters are obtained here. In section 3 and in section 4, further analytical work is carried out on the dream text reports. The aim is to determine the best way to specify clusters, in a manner that is complementary to what can always be the predominant interest and motivation for the analytics being undertaken. This latter, predominant analytical focus, in the application

domain here, is in essence an oriented, directed and therefore focused analytical processing. Besides, Section 5 is the Conclusion.

2. Multidimensional Baire Distance

An example of the Baire distance for two numbers > 0, < 1, is as follows. Having $x = 0.425$ and $y = 0.427$, the base is defined as 10, being suitable for real values, and for the first digit here, $k = 1$, 4, $x_k = y_k$; for the second digit here, $k = 2$, 2, $x_k = y_k$; and for the third digit here, $k = 3$; $x_k \neq y_k$, since here $5 \neq 7$. So, having the two numbers with precision, $|k| = 3$, the Baire distance is here $d_{10} = 10^{-2}$, with the superscript from the two equal digits, here the "common prefix", or the number of initial digits that are the same.

The Baire metric, that is also an ultrametric, this can be termed the "longest common prefix metric", and it is both important for linear computational time hierarchical clustering construction, and it can allow the hierarchy to be naturally mapped from a decimal or 10-adic numbering context to a p-adic representation. See Contreras and Murtagh [2]-[4] and see also Murtagh [11]. One further small note is that for such a metric to be applied to our data, then for multidimensional data, a computationally efficient and effective manner to carry out the hierarchical clustering is to use random projection. See Murtagh and Contreras [14].

The Baire distance can be used to have a hierarchical clustering based on decimal, i.e. 10-adic numbering, that is natural for real number systems, and this can be transformed into other p-adic numbering, see Murtagh [10]. Of course, $p = 2$, binary numbering can be taken into consideration in particular for the standard and common agglomerative hierarchical clustering, which carries out pairwise cluster creation. In general, related to such a multidimensional Baire distance is the Baire distance formulated for multi-channel data, i.e. hyperspectral images, and used for machine learning (Support Vector Machine, supervised classification) in Bradley and Braun [1].

3. Application in Text Mining

We take the dream reports that are openly accessible (DreamBank, [5]), with text mining carried out, so as to provide a basis for, or a framework for psychoanalytical context analysis: Murtagh and Iurato [18]. This here is an extension of that work, in order to both have the new hierarchical clustering carried out, with deriving of clusters, with computational benefits and also to have proof of both relevance and importance for the previous objectives. That reference has figures of the factor space mapping, and for all this work, R is used.

From the Barbara Sanders dream reports, we take 421 dream reports, recorded by day, although sometimes with up to a year between the successive reports, and these days ranged from 2 December 1960 to 13 January 1976. The word set from these 421 dream reports was such that every word had to be used five times, or more than five times. That threshold is motivated by requiring commonality at issue here, i.e. shared word associations between dream reports. Stopword removal and lemmatization were not carried out, since natural language processing, as such, is not an objective here, but rather, emotional expression, and, especially, the symmetric logic which is the basis of human unconsciousness (from Ignacio Matte-Blanco's bi-logic, cf. chapter 5 of Murtagh [11], and Murtagh [9], Murtagh and Iurato [15]). For standardizing the texts, all upper case letters were converted to lower case, and all punctuation was replaced by a blank space. This yields such words as follows: "tv" (TV in lower case), "xxx" (perhaps a person's name substituted by this); "ll" (for, e.g., we'll, i.e. we will); "couldn" (couldn't). Our objective is to find close association between words. The number of words in this derived word corpus here is 1568. For the 421 dream reports, the minimum number of words was 5, and the maximum number of words was 755. In the contingency table, or matrix crossing the 1568 words by the 421 dream reports, this matrix is quite sparse, with many 0, i.e. non-presence, values. The percentage of non-zero values was: 5.7%.

For now, we will consider all 420 factors. The eigenvalues are plotted in Figure **1**. The first eigenvalues have these values, 0.142, 0.115, 0.093, 0.089, 0.085, 0.083, 0.081, 0.080,

0.079977. This may lead, in interpretation, to retention of the first nine factors here.

While the factor space interpretation can be largely focused on the positive related to the negative axis projections, in effect it can be important to see what is close to a coordinate of zero. Such is having proximity in the factor space to the origin, that expresses commonality and typicality and lack of exceptionality, for the factor under consideration.

The following may be noted, and may also be clear from practical application and consequent interpretation. A default clustering, for a factor, might be with the assumption of uniform distribution of projections on the factor. That could lead to having a decimal, 10-ary, segmentation of the factor projections for the factors at that level, in the hierarchy associated with the factor. Therefore factor by factor in sequence leads to a regular 10-fold hierarchical or tree structure. In a particular sense, this is a divisive hierarchical construction.

A regular 10-fold hierarchical structure can be constructed at the root level, related to the factor resulting from the first in the sequence of eigenvalues; each of these 10 clusters, then have 10 further clusters at the next level, obtained from the second factor, from the second in the sequence of eigenvalues. At the next level, linked to the third factor, the third eigenvalue, the 10 × 10 clusters create, for each cluster, the set of 10 further clusters from the segmentation of factor 3 coordinates. Continuing for factor 4, at the 4th level from the tree root, there are 10^4 clusters formed.

However it has been seen from Figure **2** just how irrelevant in practice such a decimal, 10-ary based segmentation per factor, would be not particularly useful here, given the dominance of commonality, and typicality, in that many projections for each factor, may be, are, projected to, or close to, the origin.

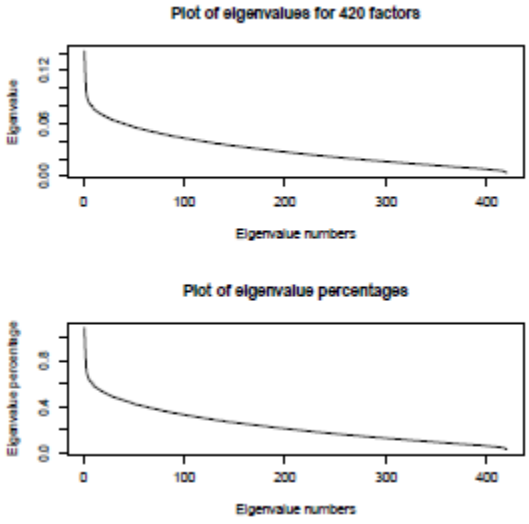

Fig. 1 A plot of the eigenvalues, the values and the percentages, yielding the percentage inertia associated with the axes, i.e. the factors, the principal components. The first nine eigenvalues have percentage values greater than 0.6.

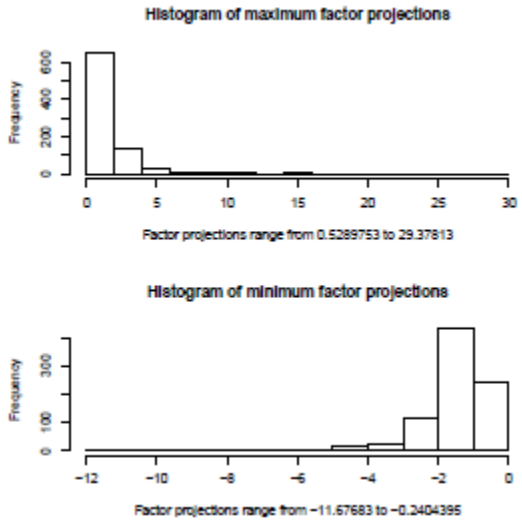

Fig. 2 Overall all 420 factors, these are histogram plots of the maximum and the minimum factor projections. Here the maximum values range from 0.5289753 to 29.37813; the minimum values range from −11.67683 to −0.2404395.

Factor	One	Minus one	Zero
1	95	2	1471
2	0	20	1548
3	14	2	1552
4	54	33	1481
5	35	25	1508
6	43	4	1521
7	18	41	1509
8	50	45	1473
9	14	42	1512

Table 1 Frequency of occurrence of projections in this ternary coding. The ternary coding, one if factor projection is ≥ 1; minus one if factor projection is ≤ 1; zero if close to the origin.

Taking into account the mapping context here, a ternary hierarchical tree, hence p-adic with $p = 3$, this may be considerably more appropriate for negative coordinates on the factor, coordinates at or close to 0, and positive coordinates on the factor. This would then lead, from the root of the hierarchical tree to a ternary, or 3-adic, hierarchical tree, with, after the tree's root, 3 clusters, next level 3^2 clusters, next level 3^3 clusters, and so on.

One easy manner to construct the ternary tree could be to have the range of factor projections to be segmented from the 33.33 and 66.66 quantiles, to that the clusters for each factor are: low-valued, negative on the factor; close to zero on the factor; and high- valued, positive on the factor.

However, from Figure **2**, the one third and two third quantiles are, in effect, more bypassed by the clear observation here: firstly, that the proximity to the origin, hence commonality and typicality in the factor projections, these dominate in the semantic, Euclidean metric endowed, factor space.

In a sense, the following is at issue here. From the decimal, 10-ary numbering, that is real-

valued, we are seeking an appropriate manner to map the 10-ary factor space into a 3-adic, ternary factor space. In the very standard agglomerative hierarchical clustering, it can be stated that the 10-ary, real space is mapped into a 2-adic, binary space that is most commonly represented by the ultrametric or 2-way tree topology. In all respects, of course, with the 2-way tree, or, as referred to above, the 3-way tree or the 10-way tree, all are represented by, and displayed by, the ultrametric topology.

From a close consideration of the histograms of the factor projections, the following ternary encoding was carried out: assigned value 1 if having factor projection ≥ 1; assigned value - 1 if having factor projection ≤ 1; and assigned value 0 otherwise. Table **1** displays the frequency of occurrence for the selected set of the initial nine factors.

From Table **1**, it is recognized that the most dominant mapping into the factor space has led here, and generally would do so, to have the mapping into the factor space at or near the origins of the factors, i.e. at or near the zero values of the factors, the axes or, as also termed in related principal components analysis, the principal components. Words here, that are commonly used, and that are typical, hence mapped into the factor space origin, for many factors, the following are the first ten in the frequency ranking: "the", "and", "to", "it", "in", "he", "my", "of", "me", "is". Some of these would be termed stopwords (and often removed from the text to be analyzed.). It was determined that from the 1568 word set in the dream reports textual data, that 579 words were at or very close to the origin of all the 420 factors. Thus there were 989 words that were not at the origin of all factors. Therefore these 989 words were selected in order for their cluster properties to be studied. Rather than setting these up in a ternary hierarchical clustering, it was decided to study the relationships and proximity of the 989 works, from their factor space mapping. The following was now undertaken: to have, for every factor, the words considered either for their very close proximity to the origin, or their exceptionality in factor projections, encompassing both positive and negative factor projections.

This is motivated by the most basic characteristic of factor space mapping being: proximity to the commonality of the factor's origin, i.e. zero value; and exceptionality in being quite prominent, here now, either positively or negatively on the factor axis. It is to be acknowledged that common analytics can be based on the most prominent contributions to inertia of the factors. Furthermore common analytics, can take a few clusters into account, that are on factor planes. The novel work under investigation here is quite complementary to such standard practice by taking, to begin with, all factors into consideration, and having an orientation for the analytics here that gives priority to commonality (near the factor origins) versus exceptionality (i.e., with factor projections that are not extremely close to the origin). Thus the analytical work in this paper is based here on commonality versus exceptionality.

For the word set of 989 words, and simply for concentrating on, or focusing on, the most important factors, these being determined, here from the full factor space, of 420 factors (i.e. axes, dimensions). For these words and for that set of predominant factors, we had each word as a vector with "exceptionality", i.e. quite positive or quite negative projections, and also with some, at least, i.e. for some factors, "commonality" mappings. Then for the 989 words, a distance matrix was determined. This was to determine the clusters, comprising identical characteristics, i.e. have the same role played in the full set of 420 factors. The L_1 distance is used to determine the word distances, and an L_2, i.e. Euclidean distance, was also assessed.

Having the thereby determined distance matrix for the 989 words, the following step was to determine the zero distance relationships here, which are to be read off as clusters here. Thus the clusters are comprised of subsets of the word corpus that have the same mapping into the factor space characteristics. As noted already, the very dominant com- monality relationships, for word mappings, were bypassed and not taken into account at this stage of the analytical processing, simply due to lack of particular interest in these 579 set of words, that were found to be lacking in any form of exceptionality. Also, as noted already, rather than studying every one of the 420 factors, it was agreed that the nine most important factors would be at issue now.

The next step was to see, for each of the 989 word set, how many zero distances, hence here

identical relationships there were. It was found that many words had just one zero distance, which was a self-distance equal to zero. The maximum number of zero distances between words, including of course the self-distance, this was found to be 33. So now, in the next stage of the processing, the clusters to be determined were for having distances between the word members of these clusters, having $1, 2, 3, \ldots, 33$ zero distance values. So this stage of the analytical processing was to determine the cluster cardinalities for each of these distance value properties of the clusters. An interesting finding was that clusters were determined with the number of their zero distance values being equal to: 1, 2, 3, ..., 33. Of course just one zero distance implies a singleton cluster, i.e. just one word in the cluster. For the other sets of zero valued distances, the following were found. Clusters with just one zero valued distance (i.e. self-distance), 920 words; with 20 zero- valued distances: 20 words; with 3 zero-valued distances: 3 words; with 6 zero-valued distances: 6 words; with 7 zero-valued distances, 7 words; and finally, with 33 zero-valued distances, 33 words.

The following are the determining of the words for these clusters.

- For having just one zero-valued distance, hence self-distance, rather than listing 920 words here, these are the first eleven of them. The 1st and the 5th in this listing being individual names, here with the first letter being reconfigured to being upper case: “Howard”, “brother”, “store”, “cat”, “Darryl”, “face”, “building”, “cousin”, “pool”, “children”, “night”.

- Next is the 2 zero-valued distances clusters (with “xxx” being an often used reference word, and the name “Sanders” here): “getting”, “drive”, “seat”, “dance”, “edge”, “both”, “xxx”, “job”, “loves”, “written”, “careful”, “rooms”, “cliff”, “straight”, “faces”, “stopped”, “liked”, “Sanders”, “bare”, “wig”.

- In all of the following cases of this specification of clusters, it was checked that all words were members of the same cluster. But this did now hold for the above, 2 zero-valued distances, multiple clusters. From the word set above, these were the clusters with zero-valued distances of the members: “getting”, “both”; “drive”, “seat”; “dance”, “loves”; “edge”, “cliff”; “xxx”, “Sanders”; “job”, “written”; “careful”, “straight”; “rooms”, “wig”; “faces”, “bare”; “stopped”, “liked”. So, therefore, ten clusters were obtained from these word pairs with zero-distance relationships. Some of these clusters are so very clear as to why the words should be semantically clustering, i.e. identically mapped based on the “exceptionality” properties that are at issue here. Cf., for example, the cluster consisting of “drive” and “seat”; the cluster consisting of “edge” and “cliff”; and the cluster with the written word “xxx” and the dream reporter here, whose name was Barbara and “Sanders”.

- For the 3 zero-valued distances cluster: “husband”, “color”, “position”.

- For the 6 zero-valued distances cluster: “which”, “running”, “until”, “silly”, “free”, “bottom”.

- For the 7 zero-valued distances cluster: “chair”, “lady”, “these”, “office”, “foot”, “without”, “himself”.

- Finally for the 33 zero-valued distances cluster: “was”, “had”, “said”, “were”, “got”, “could”, “felt”, “looked”, “went”, “started”, “came”, “thought”, “asked”, “saw”, “wanted”, “told”, “couldn”, “knew”, “took”, “tried”, “wouldn”, “walked”, “kept”, “gave”, “turned”, “picked”, “handle”, “laughed”, “noticed”, “kissed”, “talked”, “refused”, “assist”.

This work has been directed here at the sequence of stages, and these are oriented towards general patterns in the analytical mapping carried out, patterns that can be informally described as "commonality" versus "exceptionality". On that basis, clusters were determined.

This will be, or can be, complementary to the major analytical tasks that, here, for example can be related to named individuals or to other stated activities or actions. Rather, for the clusters here, the cluster characteristics are very largely resulting from avoiding "commonality". In text mining as is the case here, such "commonality" can be comprised of stopwords and quite standard and often repeated terms.

If a ternary hierarchy were constructed here, then, for non-empty clusters, there would be the maximum of, from the root node, three clusters for the first factor, then for the next factor, in total nine clusters, then for the next factor, in total twenty seven clusters and so on, for the 420 factors, ultimately to have, subject to having every cluster being non-empty, 3^{420}, i.e. $2.459954e + 200$ clusters. It may be noted however that from the set of clusters that were determined, a lot more were found to be empty. So the number of terminal nodes in the hierarchy would very likely to be a lot smaller than the maximum that has been indicated. One further consideration might be that for a large dataset, as is the case here, that a hierarchy perhaps is best to be predefined in its structure. In Murtagh [11], it is noted how it can be very desirable to have a hierarchy taking resolution scale, and data aggregation, into consideration. That can very well be relevant and important for the use of a hierarchical clustering as a conceptual hierarchy, i.e. an ontology.

4. Again the Exceptionality Clusters, from Important Factors, Here 9D

It can be the case of examining and interpreting the factor space by taking dimensionality reduction into account, if a set of the most dominant factors are to be considered as constituting a sufficient percentage of the data clouds' inertia. From the above noting of the first nine factors being the most important, and constituting effective dimensionality reduction, it can be the case of finding more relationships, here between the words, compared to the full dimensionality factor space.

As in the previous section, for the ternary encoding, applying an L_1, Manhattan, distance, and to ensure that a zero distance would not be formed from, for example, a positive factor projection, with ternary encoding equal 1, and a negative factor projection, with ternary encoding equal -1, to disallow a zero distance to be thus formed, the negative factor projection, ternary encoding was reset to 2. For proximity to the origin, having noted some maximum and minimum factor projections to be respectively more than 22, and less than -27, the interval was set for proximity to the origin, i.e. 0-valued, as follows: < 1 and > -1.

Above for the full dimensionality factor space, 420 factors, it was found that there were 2171 zero distances, i.e. 0.22%. For the best 9 factors, it was found that there were 453169 zero distances, i.e. 46.23% of the distances between the 989 word set. Both percentages here are without including the zero, self-distances.

For the 420-dimensional (full dimensional) factor space, there was a maximum of 33 zero distances for any one of the 989 words. For the 9-dimensional factor space, there is a maximum of 667 zero distances for any one of the 989 words. Many words did not have non-zero distances, apart of course from self-distances. The finding of the number of words with non-zero, here from the 9-dimensional factor space, this was:

1 zero distance, 48 words; 2 zero distances, 28 words; 3 zero distances, 18 words; 4 zero distances, 12 words; 5 zero distances, 15 words; 6 zero distances, 12 words; 7 zero distances, 7 words; 9 zero distances, 9 words; 10 zero distances, 10 words; 11 zero distances, 11 words; 17 zero distances, 17 words; 18 zero distances, 18 words; 19 zero distances, 38 words; 79 zero distances, 79 words; and 667 zero distances, 667 words.

Now, we investigate these to find clusters, defined here by all cluster members being of zero distance, semantically from the factor space.

- For the 28 words each with two zero distances, the following are the pairs with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. fourteen clusters: “stage”, “piano”; “hide”, “fake”; “loved”, “woke”; “Fletcher”, “Abner”; “wedding”, “lawyer”; “bowl”, “tea”; “metal”, “bag”; “killed”, “Gus”; “roof”, “earthquake”; “George”, “football”; “gold”, “information”; “Rosemary”, “Elliot”; “supportive”, “facing”; “catches”, “numbers”. (Included here are the names of uncle Gus, cousin Abner, other individuals are Fletcher and Elliot; so, from the dream reports, there could be interest in going back to the texts, with a focus on, and an interest in, relations’ and other peoples’ names, and expressive themes like, here, piano, earthquake, what can be a colour, gold, etc.).

- For the 18 words each with three zero distances, the following are the triplets with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. six clusters:
“delighted”, “mirror”, “barely”; “grown”, “alive”, “Naomi”; “preparing”, “background”, “pretending”; “babies”, “frozen”, “horses”; “lead”, “monkey”, “gorilla”;“Dylan”, “Josh”, “detective”.

- For the 12 words each with four zero distances, the following are the quadruples with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. three clusters:
“mountain”, “block”, “sadly”, “buildings”; “machine”, “cost”, “size”, “seem”; “drawer”, “items”, “cups”, “apple”.

- For the 15 words each with five zero distances, the following are the quintuples with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. three clusters:
“Peter”, “disgusted”, “wanting”, “roses”, “Dee”; “steep”, “sand”, “easy”, “manual”, “apart”; “ladder”, “jeep”, “bow”, “waves”, “hits”.

- For the 12 words each with six zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. two clusters: “grandmother”, “fashioned”, “plump”, “Patricia”, “bought”, “evening”; “Valerie”, “cigs”, “hers”, “bank”, “fur”, “Diane”.

- For the 7 words each with seven zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. one cluster: “hill”, “shower”, “suggests”, “rope”, “dangerous”, “helicopter”, “view”.

- For the 9 words each with nine zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. one cluster:
“getting”, “night”, “both”, “Lydia”, “almost”, “last”, “guilt”, “neat”, “french”.

- For the 10 words each with ten zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. one cluster:
“which”, “running”, “until”, “silly”, “green”, “free”, “cover”, “haven”, “bottom”, “sticks”.

- For the 11 words each with eleven zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. one cluster:

"dance", "loves", "race", "likes", "kid", "course", "leader", "Harrison", "trick", "salesman", "grabs".

- For the 17 words each with seventeen zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor sub- space, i.e. one cluster:
  "row", "clock", "candy", "breakfast", "sink", "chocolate", "shelf", "cooked", "potatoes", "pregnant", "beads", "program", "photo", "cook", "traveling", "lower", "mo tor".

- For the 18 words each with eighteen zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor sub- space, i.e. one cluster:
  "chair", "person", "lady", "these", "office", "foot", "meeting", "without", "himself", "bench", "counseling", "co", "worker", "self", "confused", "kittens", "pianos", "herself".

- For the 38 words each with nineteen zero distances, the following are the groupings with semantically identical mappings here in the 9-dimensional main factor subspace, i.e. two clusters:
  First cluster here: "Lucy", "real", "late", "game", "sweet", "than", "disabled", "mail", "dreams", "community", "reluctantly", "skirt", "loud", "longer", "radio", "performing", "department", "bathtub", "share".
  Second cluster here: "audience", "end", "lines", "order", "fingers", "lawn", "bit", "interesting", "conference", "telephone", "opening", "aside", "noisy". "showing", "ex", "speaks", "actors", "chute", "curtain".

- For the 79 words each with seventy nine zero distances, the following are the first sixteen of the 79 words, which constitute one cluster, since all 79 words are related with zero distances.
  "was", "had", "said", "were", "got", "could", "felt", "looked", "went", "started", "came", "thought", "asked", "saw", "wanted", "told".

For the 667 words each with 667 zero distances, there is found to be one cluster with all of these 667 words, all of which are here with zero distance to every other word. The first ten of these words are as follows: "brother", "drive", "store", "cat", "Darryl", "face", "building", "cousin", "pool", "children".

For this large cluster, the factor projections were examined, and for the 667 words, the minimum for the 9-dimensional factor space was -0.9918297, and the maximum was 0.9934012. Therefore, with none of the 667 words here in the 9-dimensional space, being <-1 or > 1, it is the case for such a large cluster here that the words are of a commonality nature. They are projected in the 9-dimensional factor space, very close to the origin, i.e. the zero on the nine factor axes.

Let us see if this is different when the full factor space, 420-dimensional, is considered. For the 667 words, and their projections on 420 factors, the minimum and maximum values were found to be:-4.252654 and 4.790118, respectively. Unlike the 9-dimensional factor space, here there are cases of negative factor projections and positive factor projections. Since these are not close to zero on those factors, it may be stated: taking into account the full factor space here has made the great majority of the words to be exceptional. This of course, results from having not just the more important nine factors, but from 420 factors, i.e. the full 420-dimensional factor space.

It may be very interesting to conclude from this: with the full factor space dimensionality, we obtain a small number of clusters, from the exceptionally semantic mapped words. This

therefore allows these words to be found, and their clusters, and this may be possible to result in a focus for our analysis being undertaken.

Contrary to this, it can also be the case that one seeks the majority of what is under investigation to be common and un-exceptional. For that, larger clusters can be determined.

5. Conclusions

Here it has been observed and commented on, that exploratory and studying analysis, also termed data mining, including text mining, and expressed as unsupervised machine learning, or also pattern recognition in data, or possibly also related to inductive reasoning in data analytics (cf. Murtagh and Devlin, [17], Murtagh, [11], and also [13]). The major perspective here is to distinguish between what can be clusters of exceptionality and not of commonality.

The objective is to determine what can be at the basis of, and underpinning, the causality that is manifested by the data properties and characteristics. Thus, it can be very beneficial to distinguish between what is common and shared, relative to what is exceptional and, just in a practical, application-oriented, sense, anomalous.

Most of the processing here is with linear computational time. While distances are generated for attributes, here words, this processing is carried out on words to be conceptualized as "exceptional" and not expressing "commonality"; therefore the word set can be, in practical situations, much reduced, here from 1568 words to 989, where 579 words are all mapped into the full dimensional factor space origin.

It can be noted that the very original processing stages obtained words that had more than, or equal to, five occurrences. This is just so that ultra-exceptionality and extreme anomalousness can be handled and addressed differently, by machine learning if this is required. However, for narrative analysis, and for well-based, contextualized analysis, all that is at issue is both well-established and potentially relevant.

This work (i) focuses on "exceptionality" and not "commonality", (ii) forms clusters of the former here, based on full dimensional, semantic identity, (iii) has computational efficiency, and (iv) is and must be fully based on the given data source.

Future objectives are to have this methodology for repeated surveys, to determine evo-lution and exceptionality and commonality, with domains of application, such as countries, companies, etc.

References

1. *Bradley P.E., Braun A.C.,* 'Finding the asymptotically optimal Baire distance for multi-channel data', *Appl. Math., **6**(2015), 484–495*
2. *Contreras P., Murtagh F.,* 'Linear time Baire hierarchical clustering for enterprise information retrieval',. *Int. J. Software Informat., **6**(2012), 363–380*
3. *Contreras P., Murtagh F.,* 'Fast, Linear Time, m-Adic Hierarchical Clustering for Search and Retrieval using the Baire Metric, with linkages to Generalized Ultrametrics, Hashing, Formal Concept Analysis, and Precision of Data Measurement', *p-Adic Numbers, Ultrametric Analysis Applic. **4**(2012), 45–56*
4. *Contreras P., Murtagh, F.,* 'Fast, linear time hierarchical clustering using the Baire metric', *J. Classification, **29**(2012), 118–143.*
5. *DreamBank Repository of dream reports. Retrieved from https://www.dreambank.net, 2004*
6. *Kuwil F.H., Shaar F., Topcu A.E., Murtagh F.* 'A new data clustering algorithm based on critical distance methodology', *Expert Systems Applic., **129**(2019), 296–310.*
7. *Le Roux B., Rouanet H.,'Geometric Data Analysis, From Correspondence Analysis to Structured Data Analysis',* Kluwer Academic Publishers, Amsterdam, *2004.*
8. *Murtagh F., 'Correspondence Analysis and Data Coding with Java and R',* Chapman and Hall, CRC Press, London, *2005*.
9. *Murtagh F.,* 'Mathematical representations of Matte Blanco's bi-logic, based on met- ric space and ultrametric or hierarchical topology: towards practical application', *Language Psychoanalysis*

3*(2014), 40–63.*

10. *Murtagh F.,* 'Sparse p-Adic Data Coding for Computationally Efficient and Effective Big Data Analytics', *p-Adic Numbers, Ultrametric Analysis Appl. **8**(2016), 236–247.*
11. *Murtagh F., 'Data Science Foundations: Geometry and Topology of Complex Hierarchic Systems and Big Data Analytics',* Chapman and Hall, CRC Press, London, *2017*
12. *Murtagh F., 'Big data scaling through metric mapping: exploiting the remarkable simplicity of very high dimensional spaces using Correspondence Analysis'. In: Palumbo R et al. (eds.) Data Science, Studies in Classification,* Data Analysis, and Knowledge Organization. Springer, Wien, pp. 279–290, *2017*.
13. *Murtagh F., (2017c) 'Linear Time Visualization and Search in Big Data using Pixellated Factor Space Mapping', in press: Editors, Tadashi Imaizumi, Akinori Okada, Sadaaki Miyamoto, Fumitake Sakaori, Yoshiro Yamamoto and Maurizio Vichi, IFCS2017 Post-Proceedings, volume in Springer (Wien) series "Studies in Classification, Data Analysis and Knowledge Organization" (ISSN: 1431-8814), 2017.*
14. *Murtagh F., Contreras P., 'Random projection towards the Baire metric for high di-mensional clustering'. Gammerman A., Vovk V., Papadopoulos H. (eds.) Statistical Learning and Data Sciences,* Springer (Wien, Austria) *Lecture Notes in Artificial Intelligence (LNAI) Vol. 9047, p.p. 424-431, 2015.*
15. *Murtagh F., Iurato G.,* 'Human behaviour, benign or malevolent: understanding the human psyche, performing therapy, based on affective mentalization and Matte-Blanco's bi-logic', *Annals Translational Medicine, **4**(2016), 1–10.*
16. *Murtagh F., Farid M. (2017)* 'Contextualizing Geometric Data Analysis and Related Data Analytics: A Virtual Microscope for Big Data Analytics', *J. Interdisciplinary Methodologies Issues Science, **3**(2017).*
17. *Murtagh F., Devlin K.,* 'The Development of Data Science: Implications for Education, Employment, Research, and the Data Revolution for Sustainable Development', *Big Data Cognitive Computing, **2**(2018), 14.*
18. *Murtagh F., Iurato G., '*Core Conflictual Relationship: Text Mining to Discover What and When', *Language Psychoanalysis,**7**(2018), 1–26.*
19. *Stakhov A.,* 'Mission-Critical S,ystems, Paradox of Hamming Code, Row Hammer Effect', 'Trojan Horse' of the Binary System and Numeral Systems with Irrational Bases', *The Computer Journal, **61**(2018),1038–1063.*