MESE 1

Measurement in STEM Education (MESE1)

Naples, 30-31 January, 1 February

DISP Dipartimento di Scienze Politiche Università degli studi di Napoli Federico II

Università degli Studi di Palermo

DiFC Dipartimento di Fisica e Chimica - Emilio Segrè

Clustering methods to study student reasoning lines: theoretical aspects and experimental results

Onofrio Rosario Battaglia

Dipartimento di Fisica e Chimica – Emilio Segrè, University of Palermo, Palermo, Italy

INTRODUCTION

I recent years, many studies in Physics Education have explored and discussed the possibility to investigate student answers

- to obtain information about the reasoning lines students deploy dealing with problematic situations;

- to investigate students' conceptual understanding.

INTRODUCTION

Some of these studies try to find groups of homogeneous or “similar” students by considering the ways in which they answer a questionnaire.

The goal is to efficiently partition a student sample to obtain groups so that the elements of each group are similar to each other while being substantially different from elements in other groups.

This goal may become increasingly difficult to achieve as the sample size increases.

CLUSTER ANALISYS – AN OVERVIEW

Cluster analysis (CLA) is one of the methodologies used for this purpose.

CLA methodologies are common in many fields of research, such as Information Technology, Biology, Medicine, Archaeology, Econophysics, and Market Research.

These methodologies allow a researcher to locate subsets, or clusters, within a set of objects of any nature, which tend to be homogeneous “in some sense”, without any prior knowledge of what forms those groups take (unsupervised classification).

CLUSTER ANALISYS – AN OVERVIEW

Several research papers show that the use of cluster analysis leads to identifiable groups of students that make sense to researchers and are consistent with previous results obtained using more traditional methods.

CLA can be carried out using various algorithms and techniques that differ significantly in their notion of what constitutes a cluster and how to effectively find them.

It is worth noting that in the literature the various techniques have seldom been explored and compared when applied to study a student sample, to reveal their mutual coherence, points of strength and weakness.

CLUSTER ANALISYS – AN OVERVIEW

Many aspects of CIA have been underexplored, especially in the education research, and require further study.

For instance

- The choice of criteria of similarity between students;
- The choice of clustering algorithms;
- The criteria to find the best clustering solution among all possible ones.

CLUSTER ANALISYS – EDUCATION RESEARCH

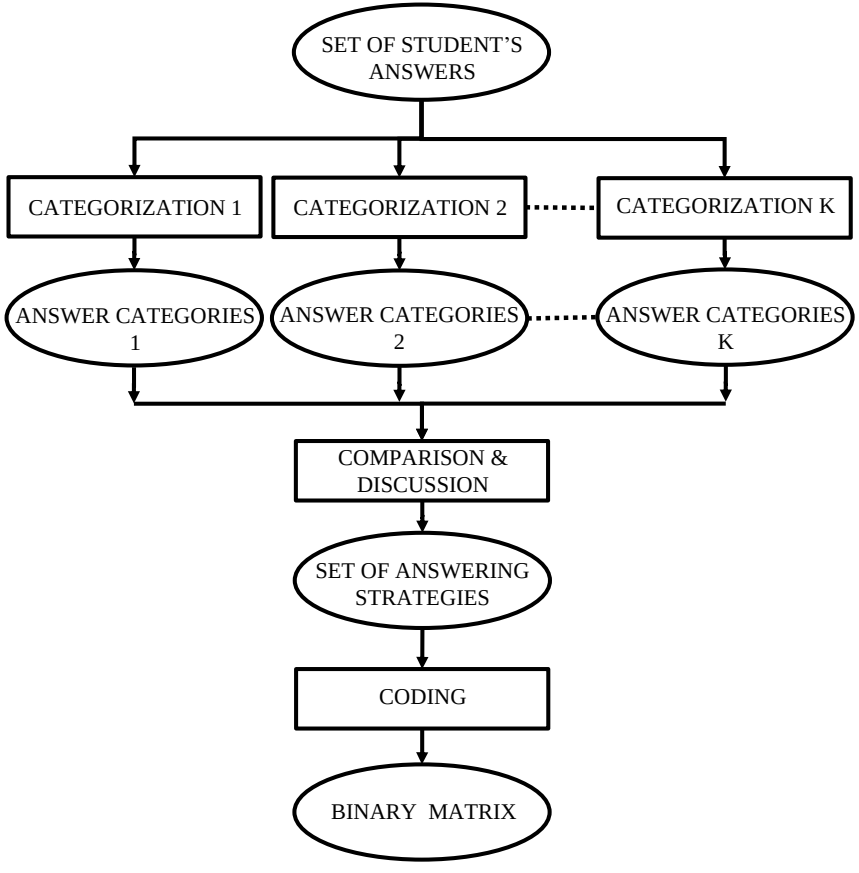
We want to statistically study a set of elements (for example, a set of students) characterised by several properties.

The properties might come from the answers given to a questionnaire.

The questionnaire could be built to investigate the lines of reasoning implemented by students when they are proposed problematic situations.

CLASSIFICATION OF STUDENT ANSWERS AND DATA CODING

Flow chart of the steps that can be followed by researchers when processing data coming from student answers to an open-ended questionnaire.



CLASSIFICATION OF STUDENT ANSWERS AND DATA CODING

Answering Strategy	Student			
	S_1	S_2	…	S_N
AS_1	0	0	…	0
AS_2	1	0	…	1
AS_3	1	...	...	...
AS_4	0	…	…	…
AS_5	1	...	...	...
…	0	…	…	…
AS_M	0	1	…	0

CORRELATION COEFFICIENT FOR BINARY DATA

CIA requires the definition of quantities as “similarity” or “distance” indexes that are used to build the clusters.

These indexes are defined by starting from the MxN binary matrix before discussed.

CORRELATION COEFFICIENT FOR BINARY DATA

If these variables are non-numeric, we may use, for instance, a modified form of the Pearson's correlation coefficient, $R_{bin}(a_i, a_j)$.

$$R_{bin}(a_i, a_j) = \frac{C(a_i, a_j) - \frac{p(a_i) \cdot p(a_j)}{M}}{\sqrt{p(a_i) \cdot p(a_j) \cdot \left(\frac{M - p(a_i)}{M}\right) \cdot \left(\frac{M - p(a_j)}{M}\right)}}$$

where

- $p(a_i)$, $p(a_j)$ are the numbers of 1s in the arrays a_i and a_j,
- M is the total number the answering strategies,
- $C(a_i, a_j)$ is obtained by counting how many times the symbol 1 is present in the same position in the arrays a_i, and a_j. $[p(a_i) \cdot p(a_j)]/M$ is the expected value of $C(a_i, a_j)$.

DISTANCE MATRIX

The similarity between students *i* and *j* can be defined by choosing a metric and calculating a distance d_{ij}.

Such a choice is often complex and depends on many factors.

If we want that two students, represented by arrays a_i and a_j negatively correlated, be more dissimilar than two uncorrelated students, a possible definition of the distance between a_i and a_j, making use of the modified correlation coefficient, $R_{bin}(a_i, a_j)$, is:

$$d_{ij} = \sqrt{2 \cdot \left(1 - R_{bin}\left(a_i, a_j\right)\right)}$$

(Gower Distance)

NON-HIERARCHICAL CLUSTERING

Non-hierarchical clustering (NH-ClA) methods partition the data space into a number of non-overlapping subsets (clusters) containing data similar to each other according to the given criteria.

Among the currently used NH-ClA algorithms, we will consider the k-means one.

K-MEANS GRAPH

To easily visualize the clusters, we would like to graphically represent in a Cartesian plane the set of students according to all the distances characterizing each student.

It is worth noting that the data input of the k-means algorithm could be the MxN binary matrix. However, a formally correct application of this algorithm strictly requires the use of a Euclidean metric, that can not be defined with binary data.

K-MEANS GRAPH – MULTIDIMENSIONAL SCALING

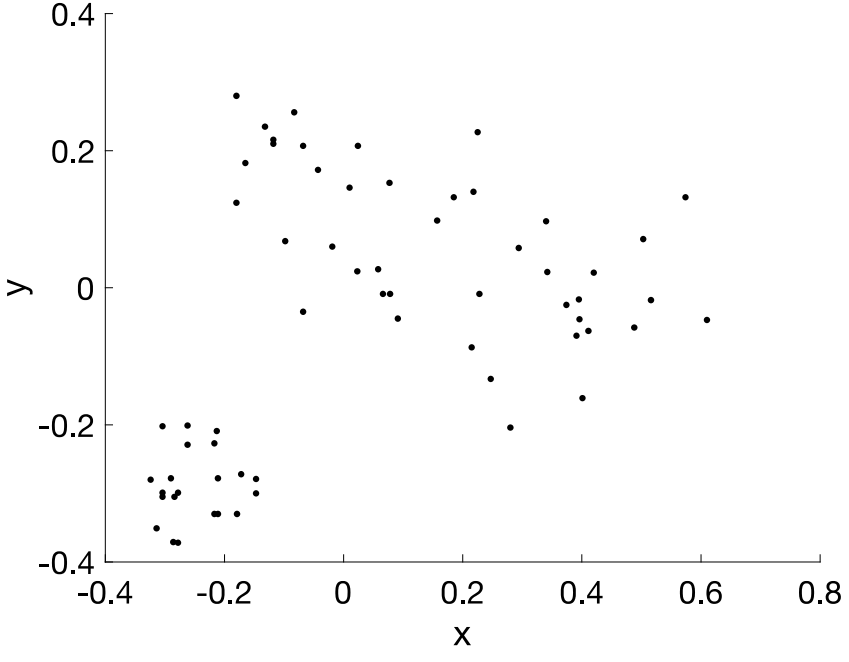
The great number of properties (the number of answers) associated with each student makes not possible to graphically represent in a Cartesian plane the set of students.

A well known procedure in the specialized literature called **Multidimensional Scaling** can be used to reduce the dimensionality associated with each student.

K-MEANS GRAPH – MULTIDIMENSIONAL SCALING

	Student			
Student	S_1	S_2	…	S_N
S_1	0	d_1	…	d_{1n}
S_2	d_{21}	0	…	d_{2n}
…	…	…	…	…
S_n	d_{n1}	…	…	0

MDS



MULTIDIMENSIONAL SCALING

Multidimensional Scaling (MDS) methods allows one to move from a space with a number of dimensions (one for each property), usually much larger than 3, to one with a smaller number of dimensions (usually two).

MULTIDIMENSIONAL SCALING

The new representation in the reduced space is a function of the initial representation and tries to preserve the distances between pairs of elements.

This MDS approach makes easier the interpretation of the data and gives a representation based on a smaller number of properties very often sufficiently faithful to the initial one.

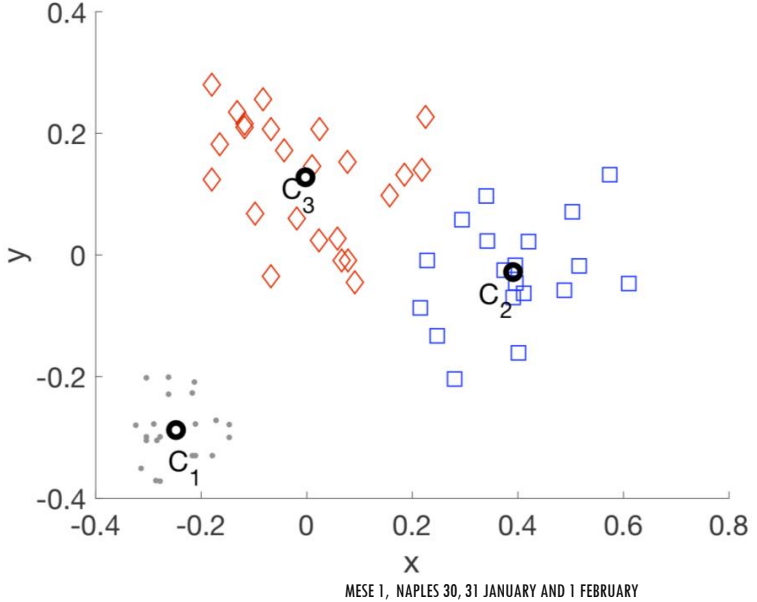
MULTIDIMENSIONAL SCALING

There are many MDS methodologies.

The one presented here refers to the Principal Component Analysis.

It is used to obtain a graphical representation of a set of students in a two-dimensional space.

Through the K-means algorithm, this representation is then partitioned into clusters.



MULTIDIMENSIONAL SCALING

Multidimensional Scaling (MDS) is a method used in the field of Natural Sciences, Engineering and Economics when you want to deal with problems related to multivariate analyses.

MULTIDIMENSIONAL SCALING

Usually, in a multivariate analysis, the data are represented by a matrix.

Let us consider a matrix X (n x k) (n individuals and k quantitative variables or properties).

We want to obtain a reduction of the columns of the matrix X, by finding a number q ($q < k$) of properties or artificial variables.

MULTIDIMENSIONAL SCALING

We want to get a representation of the students in a bi-dimensional space, where the two dimensions represent two artificial properties dependent on real properties.

It is necessary to find a function that put in relationship the space with *k* dimensions with the two-dimensional one.

MULTIDIMENSIONAL SCALING

The function can be a linear application

Let's take into account a generic artificial variable $\boldsymbol{y_j}$ as a linear combination of the real variables $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ..., $\boldsymbol{x_k}$.

$\boldsymbol{y_j}$ is a vector made of *n* components.

The generic element is y_{ij} (with $i = 1 \ldots n$) and can be expressed as

$$y_{ij} = a_{1j}x_{i1} + a_{2j}x_{i2} + \cdots + a_{kj}x_{ik}$$

or in matricial terms

$$\mathbf{y_j} = X\mathbf{a_j}$$

with a_{1j}, a_{2j}, ..., a_{kj} the coefficients in the linear combination and $\boldsymbol{a_j}$ the vector (k x 1).

MULTIDIMENSIONAL SCALING

We want to find the $\boldsymbol{a_j}$ **array**, also called principal component.

We want that the new artificial variables have maximum variance.

In this way, it will have the maximum possible information contribution.

It is easy to demonstrate that

$$\mathrm{var}(y_j) = a_j^T S\, a_j$$

where *S* is the variance/covariance matrix of the matrix *X* with dimension (k x k).

MULTIDIMENSIONAL SCALING

It is important to note that, to have a unique solution, a constraint must be imposed on the coefficient array $\boldsymbol{a_j}$.

The principal component can be determined by solving the following constrained maximum problem

$$\max\left(\mathrm{var}(\mathrm{y}_1)\right)$$

where $\mathrm{a}_1^{\mathrm{T}}\,\mathrm{a}_1 = 1$ is the constrain.

MULTIDIMENSIONAL SCALING

The method of the Lagrangian function allows one to obtain the constrained maximum.

The Lagrangian function L to be maximized is therefore given by

$$\mathrm{L} = \mathrm{a}_1^{\mathrm{T}} \mathrm{S}\, \mathrm{a}_1 - \lambda\left(\mathrm{a}_1^{\mathrm{T}}\, \mathrm{a}_1 - 1\right)$$

where λ is the lagrangian multiplier.

The solution is found by the following expression

$$\frac{\partial \mathrm{L}}{\partial a_{i1}} = 2S\, \mathrm{a}_{\mathrm{i1}} - 2\lambda\, \mathrm{a}_{\mathrm{i1}} = 0$$

MULTIDIMENSIONAL SCALING

The previous equation identifies a linear and homogeneous system that admits solutions if and only if

$$\det(S - \lambda I) = 0$$

The *k* solutions of the previous equation are the eigenvalues λ of the matrix *S*.

For instance

$$S\, a_1 = \lambda_1\, a_1$$

The array a_1, which we are looking for, is an eigenvector of the matrix *S*.

MULTIDIMENSIONAL SCALING

The results of the principal component analysis depend on the unit of measurement used for the variables.

It is a not negligible drawback because by changing the units of measurement we could obtain completely different results.

To get around this problem, it is customary to conduct an analysis on standardized variables

MULTIDIMENSIONAL SCALING

The standardized variables z_1, z_2, z_k where the generic z is

$$z = \frac{x - \langle x \rangle}{\sigma(x)}$$

It has zero mean value and variance equal to 1.

The variance/covariance matrix of the standardized variables is the correlation matrix.

Therefore, it is possible to perform a principal component analysis by finding the eingvalues of the matrix R.

MULTIDIMENSIONAL SCALING

Education Research field

- We calculate the distances matrix from the correlation matrix.

Each element in the matrix represents a distance (similarity) between pairs of students.

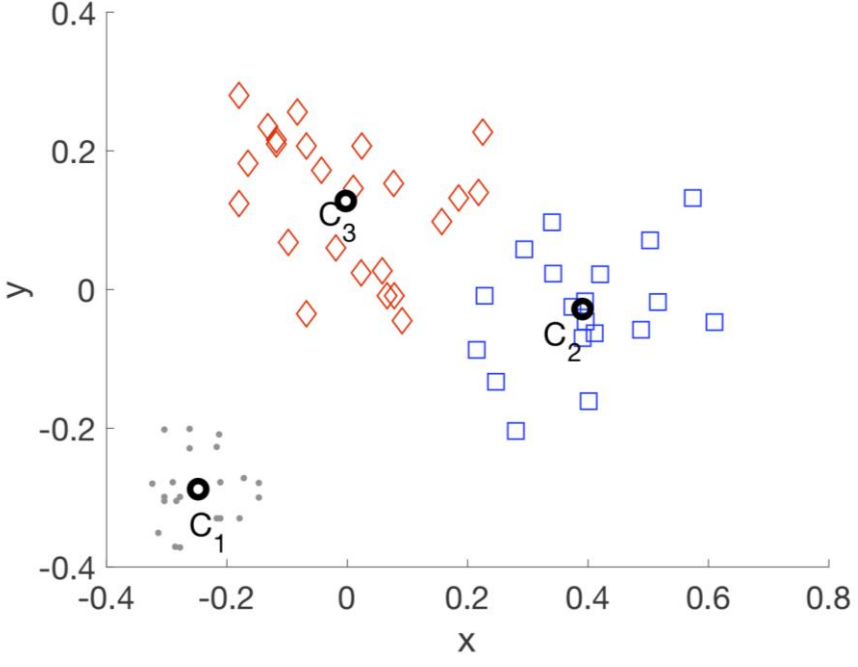
- We apply the method previously described **by finding the eingvalues of the distance matrix.**

K-MEANS ALGORITHM

The starting point of the k-means algorithm is the choice of the number, *q*, of clusters one wants to populate and of an equal number of “seed points”.

Data (students) are then grouped on the basis of the minimum distance between each student and the seed points.

Starting from an initial classification, students are iteratively attributed from one cluster to another one, until no significant improvement can be made.



K-MEANS ALGORITHM

The k-means algorithm has some points of weakness.

- the a-priori choice of the initial positions of the centroids.

  This is usually fixed by repeating the clustering procedure for several values of the initial conditions and selecting those that lead to the minimum values of the distances between each centroid and the cluster elements.

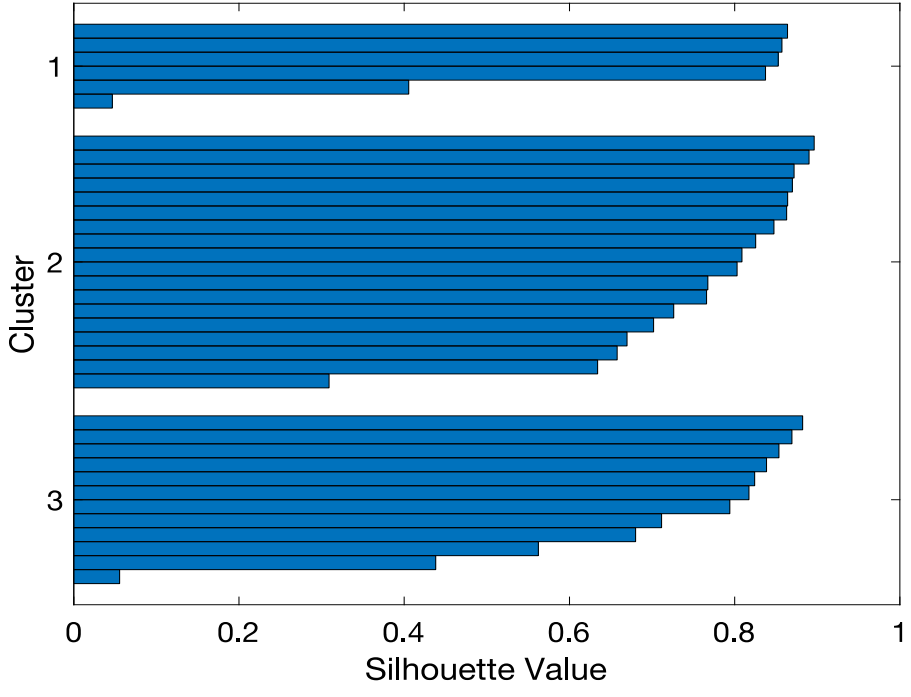
- at the beginning of the procedure, it is necessary to **arbitrarily** define the number, *q*, of clusters.

  A method widely used to select this number *q*, is the calculation of the so-called Silhouette Function, S.

THE SILHOUETTE FUNCTION

This function allows one to decide how good is the partition into *q* clusters

For each number of clusters, *q*, and for each student, *i*, assigned to a cluster *k*, with *k=1, 2,..q*, the value of the *Silhouette Function*, $S_i(q)$, is calculated.

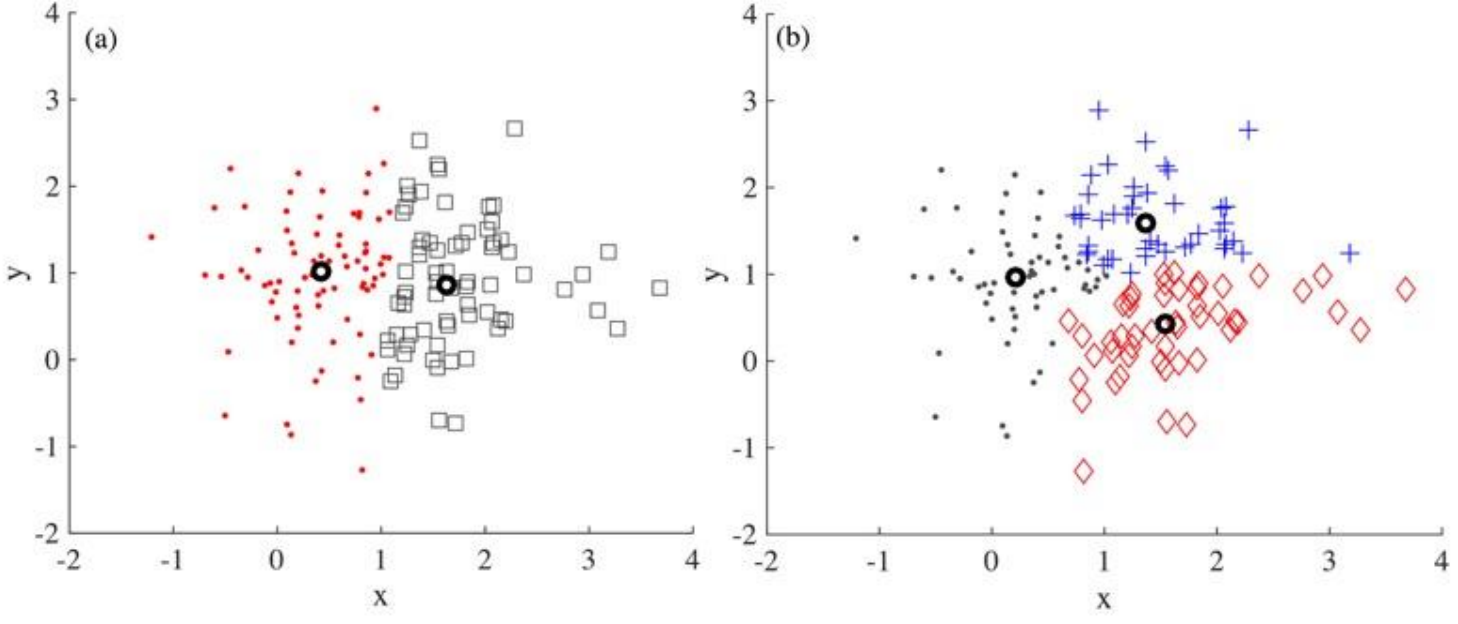


THE SILHOUETTE FUNCTION

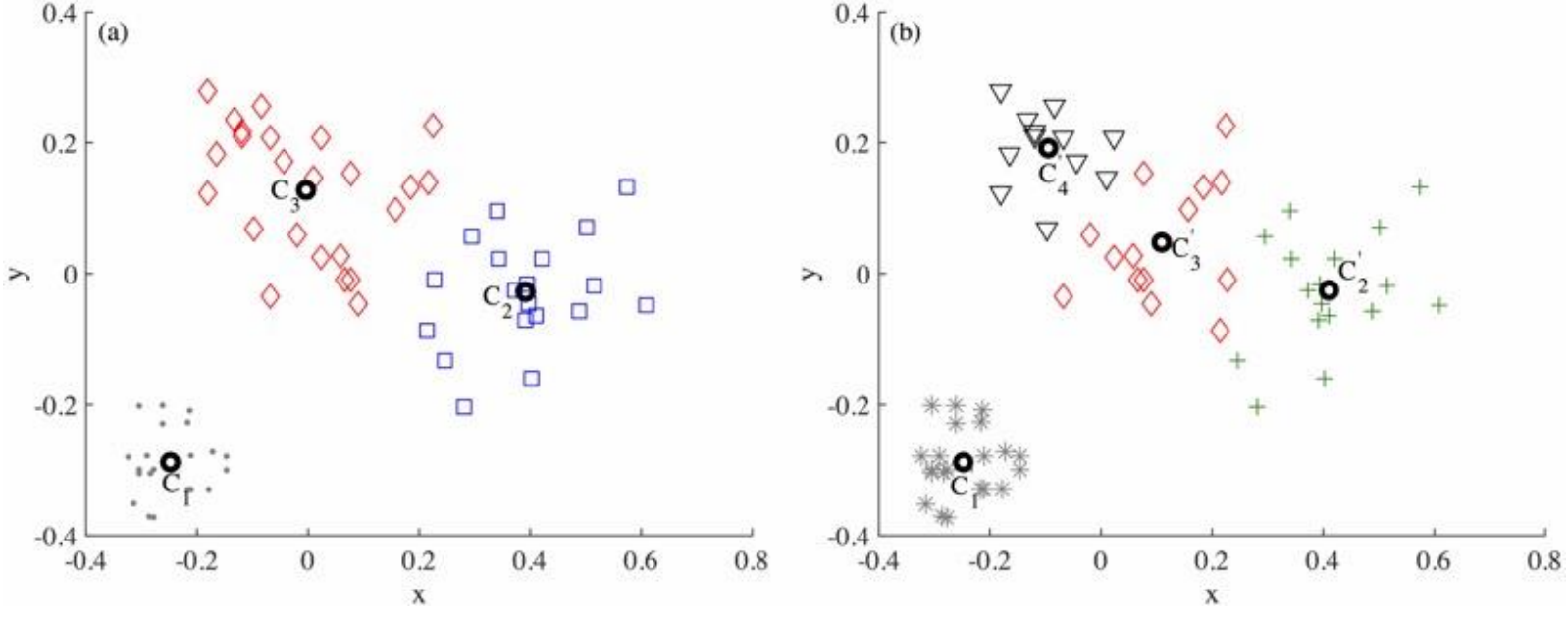
$$S_i(q) = \frac{\min\limits_{p,p\neq k}\left[\sum_{l=1}^{N-n_k}\frac{d_{il}}{N-n_k}\right] - \sum_{j=1}^{n_k}\frac{d_{ij}}{n_k}}{\max\left[\sum_{j=1}^{n_k}\frac{d_{ij}}{n_k}, \min\limits_{p,p\neq k}\left[\sum_{l=1}^{N-n_k}\frac{d_{il}}{N-n_k}\right]\right]}$$

where the first term of the numerator is the average distance of the *i*-th student in cluster *k* to *l*-th student placed in a different cluster *p* (*p = 1,..., q*), minimized over clusters. The second term is the average distance between the *i*-th student and another student *j* placed in the same cluster *k*.

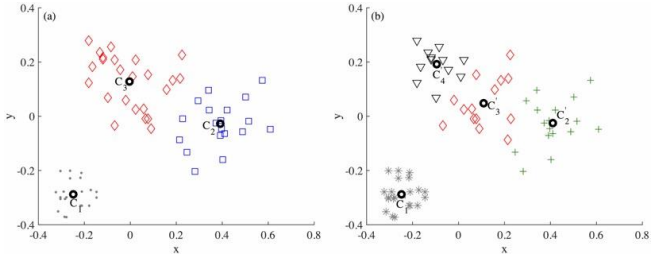
K-MEANS ALGORITHM



K-MEANS ALGORITHM



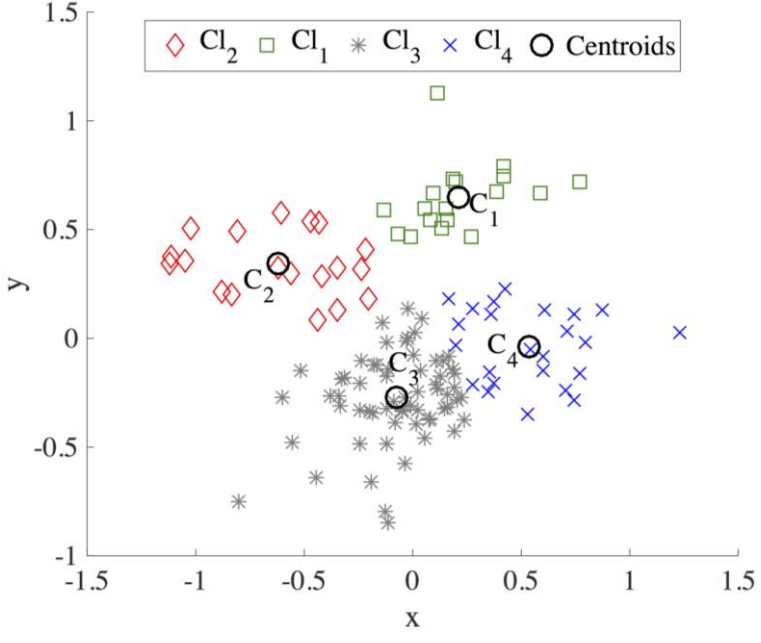
K-MEANS ALGORITHM

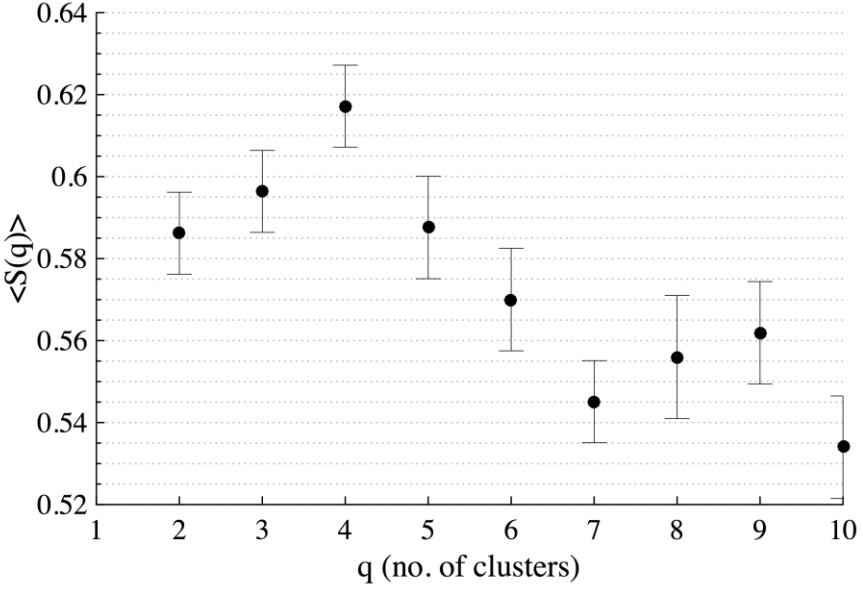


The Silhouette Function: an example of use to obtain the best clustering solution

Number of clusters (q)	**Silhouette Average value $\langle S(q)\rangle$ (CI)**	**Silhouette Average value for cluster $\langle S(q)\rangle_k$, k=1…q (CI)**			
3	0.795 (0.780 – 0.805)	k			
		1	2	3	
		0.953 (0.951 – 0.956))	0.79 (0.78 – 0.81)	0.66 (0.63 – 0.68)	
4	0.729 (0.711 – 0.744)	k			
		1	2	3	4
		0.953 (0.951 – 0.956)	0.67 (0.64 – 0.69)	0.77 (0.74 – 0.79)	0.44 (0.40 – 0.47)

K-MEANS ALGORITHM - THE SILHOUETTE FUNCTION





CHARACTERIZATION OF CLUSTERS

Once the appropriate partitioning of data has been found, the educational researcher is interested in characterizing each cluster to make sense of what the partition means in pedagogical terms.

A possible way to do this is to take into account the **most frequently used answering strategies** in each cluster.

In the case of K-means clustering, the **most frequently used answering strategies** in each cluster coincide with the components of the array associated with the centroids in that cluster.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

Our research sample consists of 36 freshmen attending the Undergraduate Program in Chemical Engineering.

We administered an open-ended questionnaire made of 6 questions.

We want to analyse the **lines of reasoning** applied by undergraduate students when asked to make sense of situations related to **thermally activated phenomena**.

We want to investigate the explanation and generalization skills in undergraduate Chemical Engineering students.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

The Questionnaire

1. In modern oil mills olive oil flows inside metallic pipes. These pipes are often enclosed in bigger, coaxial pipes in which hot water flows. Explain the possible reason of this, pointing out what are the quantities needed for a description of the proposed situation and for the construction of an explicative model.

2. In chemistry it is well known from Eyring's absolute rate theory that the viscosity of a fluid follows the following law:

$$\eta = Ae^{E_{vis}/kT}$$

   Describe each listed quantity, clarifying its physical meaning and the relations with the other quantities.

3. In petroleum industry additives are often added to gas oil to work as catalysts. What do you think can the role of these additives be in the flowing of gas oil in a pipe?

4. Can you give a microscopic interpretation of the law seen in question 2)?

5. Can you think of other natural phenomena which can be explained by a similar model?

6. Which similarities can be identified in the previous phenomena? Is it possible to find a common physical quantity which characterizes all the systems you discussed in the previous questions?

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

In this case it is as if we had 6 possible values or properties that characterize each student.

So The MDS allow us to move from a space with six dimensions to a space with two dimensions (two cartesian coordinates for each student).

At the end of the coding procedure, we obtained one shared list of M = 55 typical answers given by the students when tackling the questions.

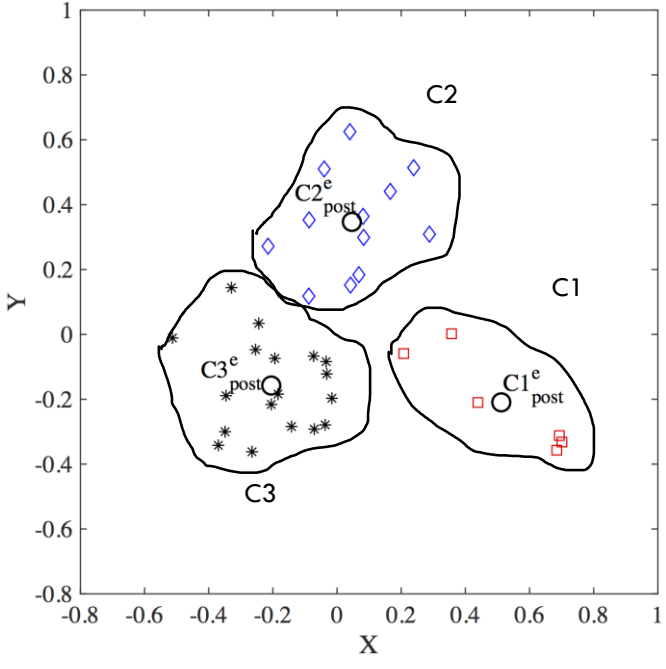
AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

k-means graphs. Each point in this Cartesian plane represents a student.

We can plot the student sample as reported in the figure and then we can apply the k-means algorithm to find possible clusters.

Points labelled $C1^{e}_{post}$, $C2^{e}_{post}$, $C3^{e}_{post}$ are the cluster centroids

<S>= 0.72 (C.I. = 0.63 … 0.78)



AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

An overview of results obtained by applying the k-means algorithm.

Cluster centroid	$C1^{e}_{post}$	$C2^{e}_{post}$	$C3^{e}_{post}$
More frequently given answers	1K, 2F, 3L-3M, 4I, 5I, 6K	1K, 2F, 3I, 4H, 5D, 6G	1H, 2E, 3J, 4F, 5E, 6H
Number of students	6	12	18
$\langle S_k(3) \rangle$	0.64	0.69	0.77

The codes used for the most frequently given answers refer to the answering strategies for the questionnaire items

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

Characterisation of student sample

Cluster C1

Students are clearly able to explain the situations and problems proposed in the questionnaire relating them to a functioning mechanisms based on the idea of thermal activation (1K-3L/M-4I-5I-6K).

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION – THERMAL PHENOMENA

Characterisation of student sample

Cluster C2 and C3

- Students are still anchored to memories of past studies (3I/J, 5D).
- Students show to be able to explain the flow process in mathematical terms (1H, students in C3) or by citing a functioning mechanism (1K, 2F, students in C2).
- They (both C2 and C3) discuss the role of an additive by considering the energy gap concept but frequently do not relate it to interaction between molecules (3I/J). However, in some cases, the Arrhenius-like expression for viscosity is interpreted in terms of interaction between molecules.
- They seem to possess generalization skills, even if in some cases limited to familiar contexts (5D/E-6G/H).

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

In the hierarchical clustering algorithm (*H-ClA*), each student is initially considered as a separate cluster.

Then, the two “closest” students are linked as a cluster and this process is continued (in a stepwise manner) to join

- one student with another one;
- a student with a cluster;
- a cluster with another cluster,

until all the students are combined into one single cluster as one moves up the hierarchy.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

The results of hierarchical clustering are graphically displayed as a tree, referred to as the *hierarchical tree* or *dendrogram*.

The term ‘closest’ is identified by a specific rule coincident with a so called *linkage algorithm*.

Hence, for different linkage algorithm the corresponding distance between a student and a cluster or a cluster and another cluster is differently computed.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Linkage algorithms

The choice of a linkage algorithms is one of the most relevant aspects of H_ClA, because different algorithms may generate different dendrograms and, so, different results.

Among the many linkage algorithms described in the literature, the following have been taken into account in education studies:

- **Single**
- **Complete**
- **Average**
- **Weighted average.**

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Linkage algorithms

Each **linkage** defines the distance between two clusters by defining a new metric (called "ultrametric") and influences the interpretation of the word "closest".

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Linkage algorithms

Single linkage, also called *nearest neighbor linkage*, links two clusters *r* and s by using the smallest distance between the students in *r* and those in *s*.

Complete linkage, also called *farthest neighbor linkage*, uses the largest distance between the students in *r* and the ones in *s*.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Linkage algorithms

Average linkage links two clusters *r* and s by using the average distance between the students in *r* and those in *s*.

Weighted average linkage uses a recursive definition for the distance between two clusters.

If cluster *r* was created by combining clusters *p* and *q*, the distance between *r* and another cluster *s* is defined as the average of the distance between *p* and *s* and the distance between *q* and *s.*

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Linkage algorithms

When the source data are in binary form (as in our case), the ***single*** and ***complete*** linkage algorithms do not give a smooth progression of the distances.

For this reason, when the source data are in binary form, the workable linkage algorithms actually reduce to the ***average*** or ***weighted average*** ones.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Cophenetic correlation coefficient

The cophenetic coefficient is a measure of how faithfully a dendrogram preserves the pairwise distances between the original un-modeled data points.

$$c_{coph} = \frac{\sum_{i<j}\left(d_{ij} - \langle D \rangle\right) \cdot \left(\delta_{ij} - \langle \Delta \rangle\right)}{\sqrt{\sum_{i<y}\left(d_{ij} - \langle D \rangle\right)^2 \cdot \sum_{i<j}\left(\delta_{ij} - \langle \Delta \rangle\right)^2}}$$

where:

- d_{ij} is the distance between elements *i* and *j* in *D*.
- δ_{ij} is the ultrametric distance between elements *i* and *j* in Δ, i.e., the height of the link at which the two elements *i* and *j* are first joined together.
- <*D*> and <Δ> are the average values of *D* and Δ, respectively.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Cophenetic correlation coefficient

In fact, as a Pearson-like correlation coefficient, it tries to quantify the “goodness” of a possible linear relationship between D and Δ under the hypothesis that these two matrices are statistically independent.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Cophenetic correlation coefficient

However

This hypothesis of linear relationship between D and Δ is not generally verified, and in many cases the relationship between D and Δ may not be monotonic.

Moreover

even in the case of a linear relationship between the corresponding values of the two matrices (and therefore a high value of the cophenetic coefficient), the difference between these may not be small.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Distance coefficient

Merigot et al. (2010) discuss a method based on measuring the distance between the two matrices *D* and Δ.

The metric proposed by the authors is, in many cases, not effective because it returns the same distance values for different types of linkage, thus failing to discriminate between them.

So, we proposed the following definition of distance between two corresponding elements of *D* and Δ

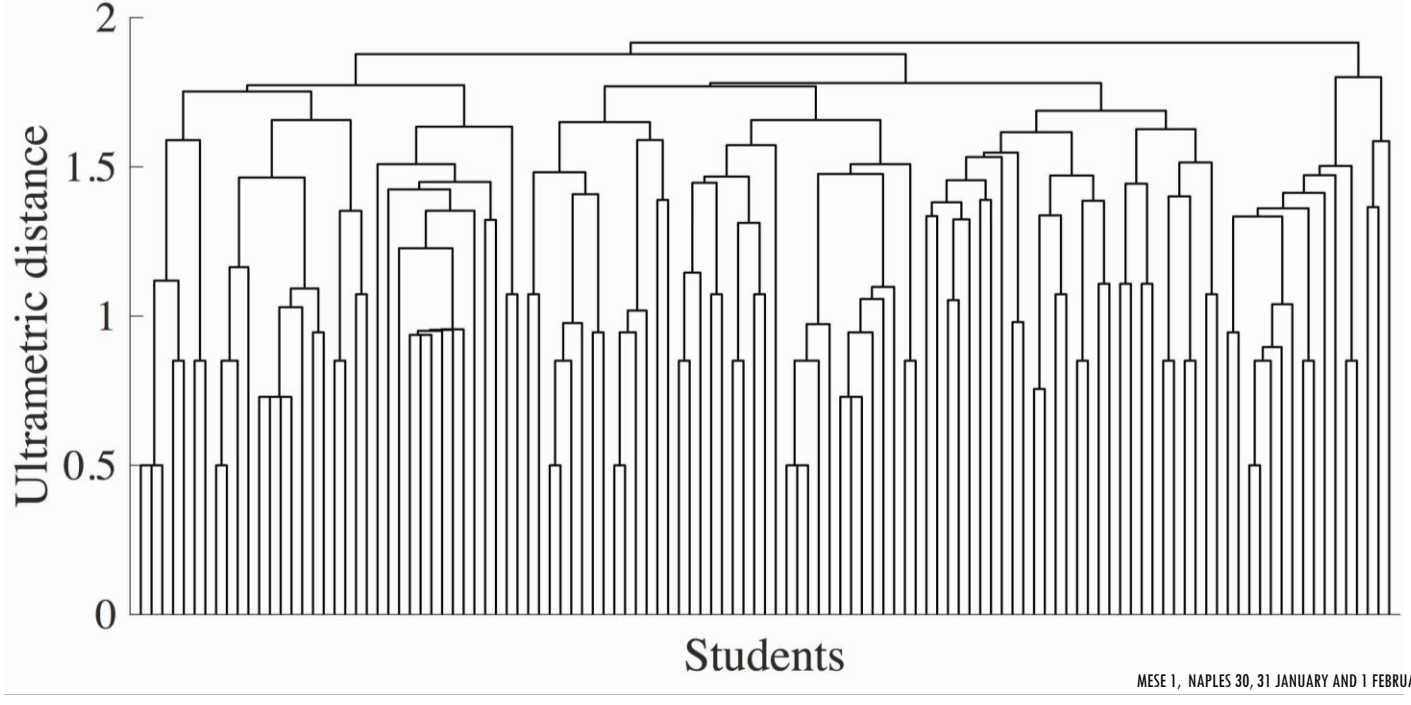
$$\sqrt{\sum_i \sum_j \left(d_{ij} - \delta_{ij}\right)^2}$$

which is inspired by the well-known Frobenius norm and is a matrix 2-norm.

It is worth noting that the use of a matrix norm does not need any hypothesis on the relationship between the distance and the ultrametric distance.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Reading a dendrogram and finding clusters in it can be a rather complex and arbitrary process.



AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria

There is not a widely accepted criterion that can be applied to determine the distance values to be chosen for identifying the clusters.

Different criteria, named *stopping criteria*, aimed at finding the optimal clustering solution are discussed in the literature.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria

"*Inconsistency Coefficient*" (I_k)

"*Variation Ratio Criterion*" (*VRC*)

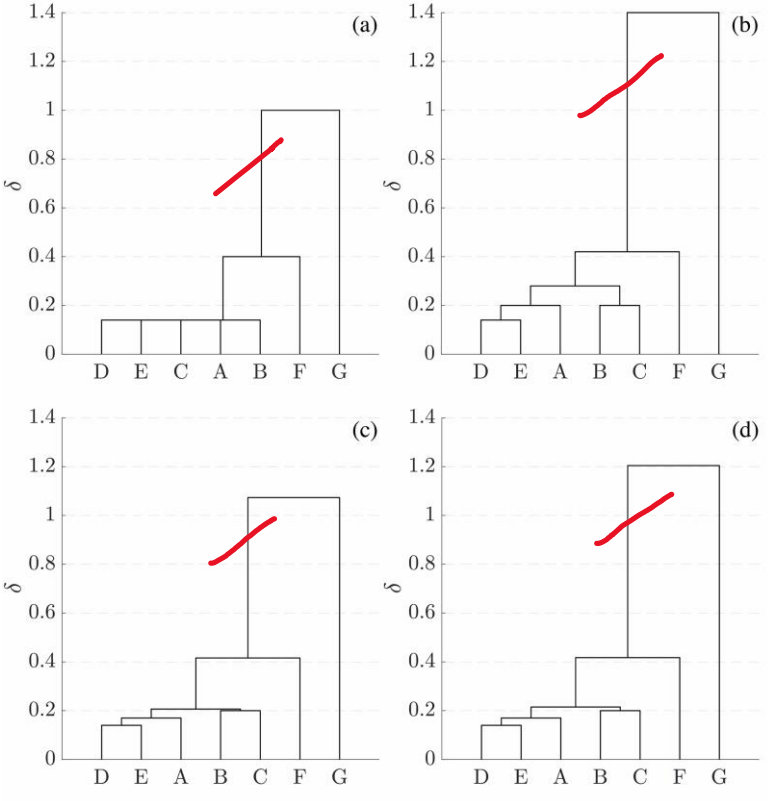
"*Cluster Differentiation Coefficient*" (*CDC*)

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria - Inconsistency Coefficient

To find clusters that can be considered distinct from each other.

If a link of two clusters is “appreciably” higher than the links below it, the link is inconsistent and two clusters can be considered disconnected.



AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria - Inconsistency Coefficient

We consider two clusters, *s* and *t*, whose distance value is reported in matrix Δ and that converge in a new link, *k,* (with *k* = 1, 2, … *N*-1).

If we indicate by *δ(k)* the height in the dendrogram of such a link, its *inconsistency coefficient* is calculated as follows

$$I_k = \frac{\delta(k) - \langle \delta(k) \rangle_n}{\sigma_n\big(\delta(k)\big)}$$

by considering a number of link below the link *k* equal to *n.*

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria – Variation Ratio Criterion

Cowgill et al. (1999) obtain the best clustering solution to a given problem by using the so called Variation Ratio Criterion (*VRC)*.

This criterion gives a coefficient that relates the best clustering solution to two factors

- high cluster separation
- high cluster compactness.

For a given configuration of *N* elements in *q* clusters, this value is defined as

$$VRC = \frac{BGSS}{q-1} \Big/ \frac{WGSS}{N-q}$$

with WGSS (Within Group Squared Sum) , BGSS (Between Group Squared Sum)

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria – Cluster Differentiation Coefficient

In order to quantify the information in a clustering, let's take into account the product between

- the number of clusters
- their distinctness

By taking into account the product between the number of clusters and the cluster distinctness we have a non-monotonic behavior with respect to the number of clusters.

Its maximum value give us the maximum information about the sample.

AGGLOMERATIVE HIERARCHICAL CLUSTERING ALGORITHM

Stopping criteria – Cluster Differentiation Coefficient

We define the *Cluster Differentiation Coefficient* (CDC) as follows

$$CDC = \frac{4 \cdot q}{N^2 \cdot l \cdot \binom{q}{2}} \cdot \sum_{i=1 \ldots q} \sum_{j=1 \ldots q} \left(n_i \cdot n_j \cdot \Theta_{ij} \right)$$

where n_i and n_j are the number of elements in clusters i and j, respectively, Θ is the "distinctness" of clusters i and j, defined as the number of components of cluster i and j centroids that are different each other, l is the total number of centroid components and $\binom{q}{2}$ is the number of combinations of q elements taken two at a time.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

The sample consists of 117 Italian students (aged 18–19) attending the last year of their 5-year secondary school course. They have completed a questionnaire made up of six open-ended questions on the concept of models and modelling.

A list of 43 typical students' answering strategies has been prepared according to the coding procedure already described.

We analyse a binary matrix composed of 43 rows and 117 columns.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

The Questionnaire

1. Models are widely used in the sciences, but what is, in your opinion, a model in physics?

2. What is a mathematical model?

3. Are models human creations or do they already exist in nature?

4. What are the main characteristics of a model?

5. Can any natural phenomena be described or explained by a model? Explain your answer.

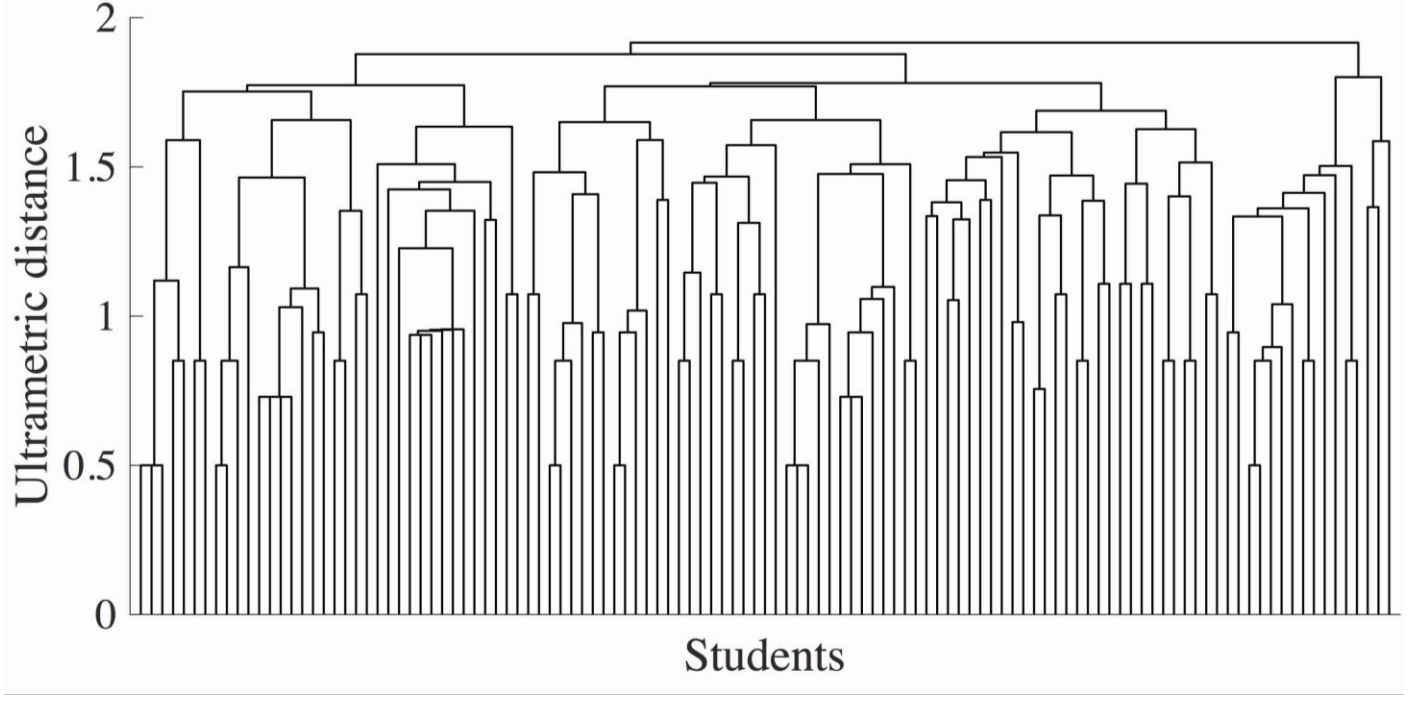
6. Can a natural phenomenon always be expressed by mathematical formulas? Explain your answer.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

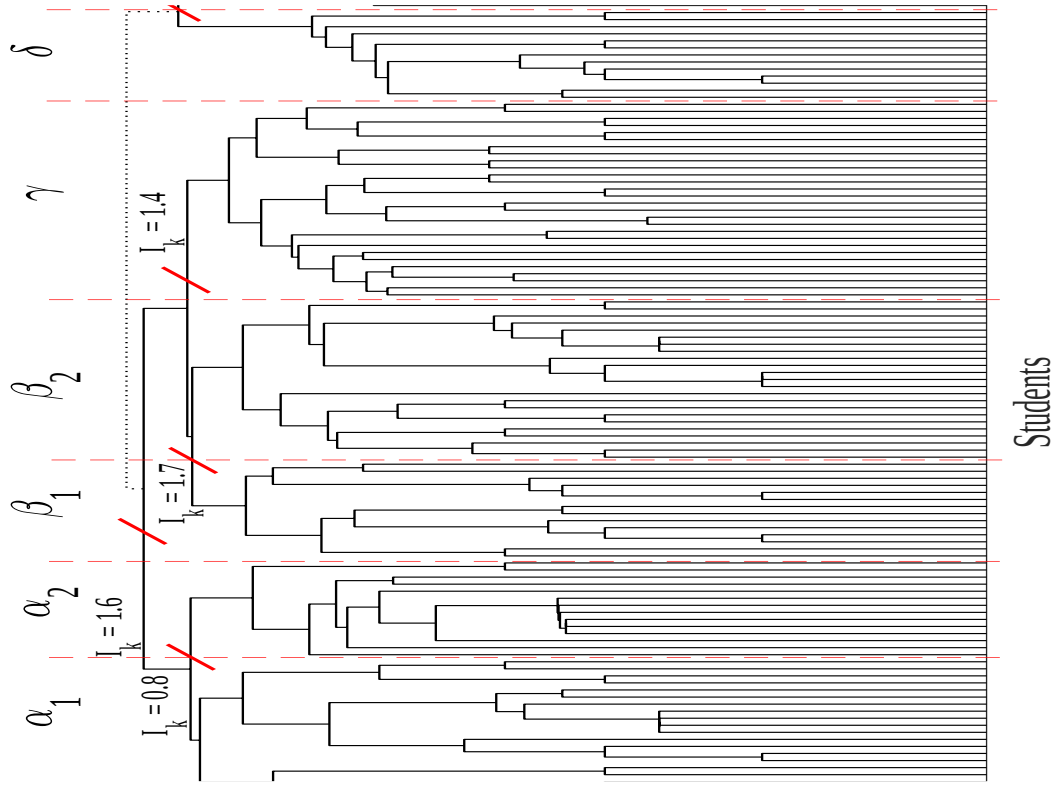
Cophenetic and 2-norm distance values for different linkage methods.

Linkage/Criterion	Cophenetic	2-norm
Single	0.76	5603
Complete	0.69	3528
Average	0.83	1793
Weighted Average	0.81	1889

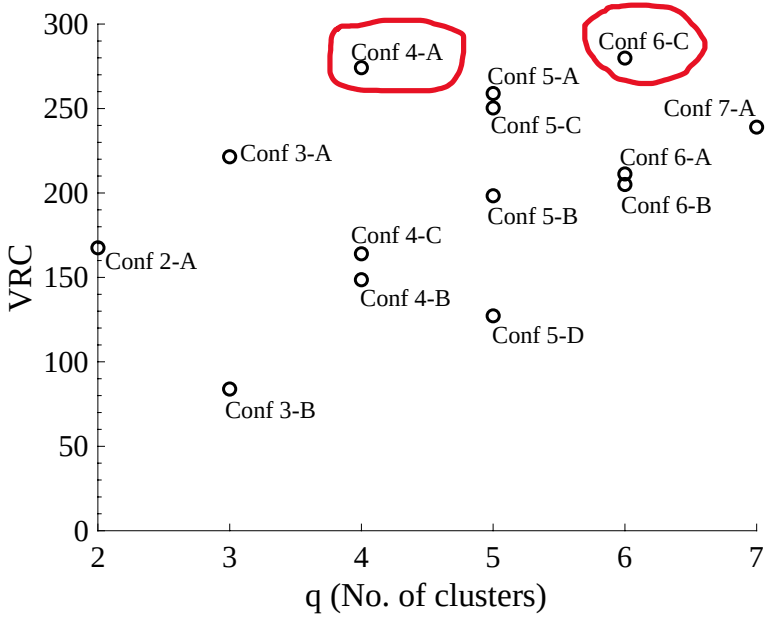
AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

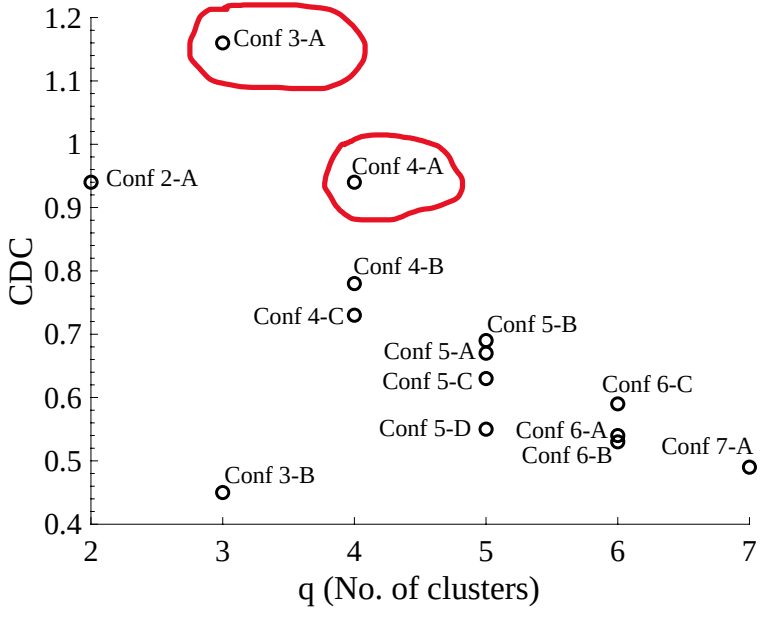


AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING



AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

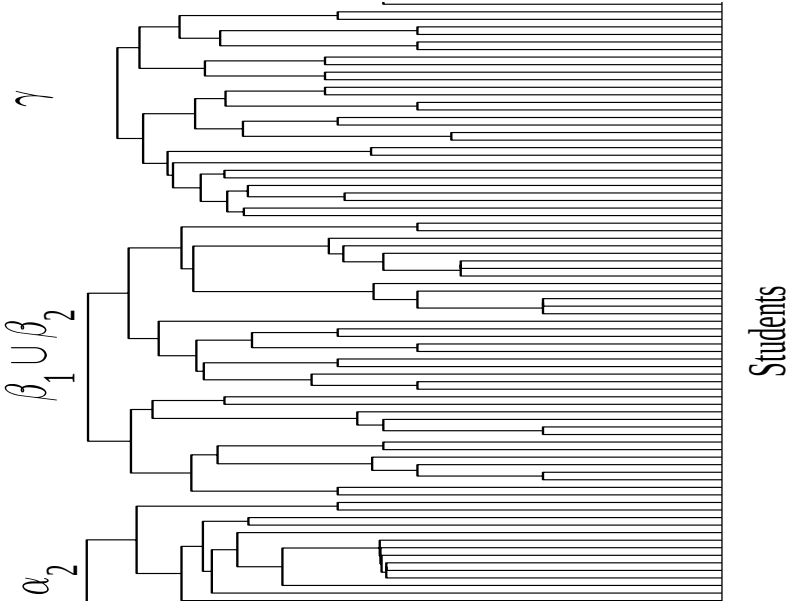




AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

Dendrogram plot of our sample in which four clusters (solution 4-A) are clearly highlighted.

- $\alpha_1 \cup \alpha_2$
- $\beta_1 \cup \beta_2$
- γ
- $\delta \cup \varepsilon$



AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - MODELS AND MODELLING

An overview of results obtained by *H-ClA* method: four cluster solution.

Cluster	$\alpha = \alpha_1 \cup \alpha_2$	$\beta = \beta_1 \cup \beta_2$	γ	$\delta \cup \varepsilon$
Most frequently given answers	1A, 2C, 3D, 4A, 5A, 6B	1C, 2D, 3B, 4E, 5B, 6E	1C, 2E-G, 3D-E, 4F, 5E, 6G	1E, 2H, 3F, 4H, 5G, 6H
Number of students	36	37	28	16

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - COMPARING CLUSTERING METHODS

I want now to discuss how is possible to **compare hierarchical** and **non-hierarchical** analysis methods to study their differences and possible coherence aspects.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - COMPARING CLUSTERING METHODS

As Meila et al. (2007) point out:

"*Just as one cannot define a best clustering method out of context, one cannot define a criterion for comparing clusters that fits every problem*."

A criterion called *Variation of Information* (*VI*) can be applied.

It measures the difference in information shared between two particular partitions of data and the total information content of the two partitions.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - COMPARING CLUSTERING METHODS

The smaller the distance between the two clustering solutions, the more these are coherent with each other, and vice versa.

VI values can be normalized to the 0 – 1 range.

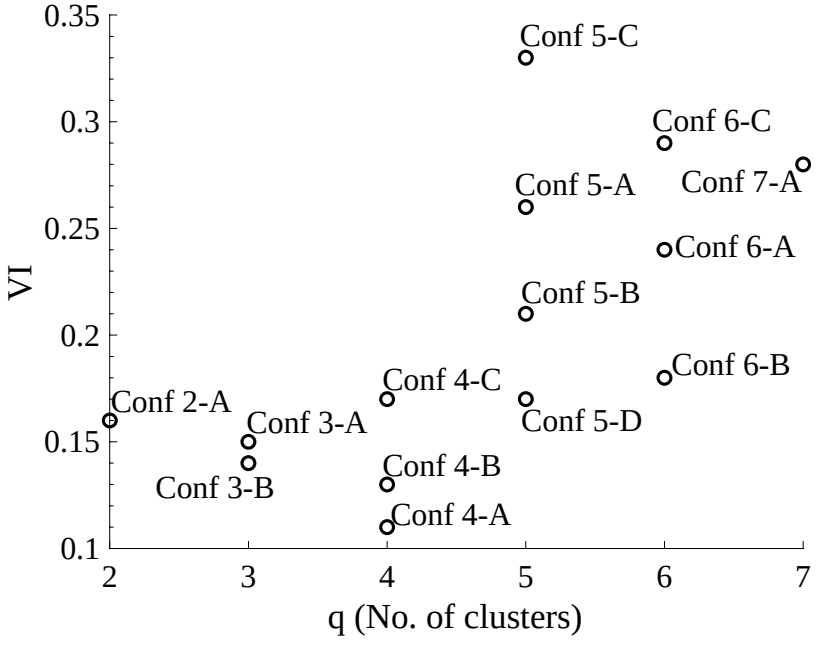
A value equal to 0 indicates very similar clustering results, and a value equal to 1 corresponds to very different ones.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - COMPARING CLUSTERING METHODS

Models and Modelling

This graph reports the values of VI for the comparison between the 3-cluster solution with k-means algorithm and many NH-ClA cluster solutions.

It is possible to conclude that the best agreement can be found for the 4-A clustering results of the H-CLA method.



It is worth noting that this result supports our previous decision to consider solution 4-A as the best H-CLA one.

AN EXAMPLE OF RESEARCH IN PHYSICS EDUCATION - COMPARING CLUSTERING METHODS

Models and Modelling

An overview of results obtained by k-means method: 3-cluster solution

Cluster	Cl_1	Cl_2	Cl_3
Most frequently given answers	1C, 2B, 3B, 4F, 5E, 6G	1A, 2C, 3B-C-D, 4A, 5A, 6B	1E, 2H, 3F, 4H, 5G, 6H
Number of students	67	37	13

Redistribution of students placed in *k-means* clusters into *H-ClA* clusters

Cluster	$\alpha = \alpha_1 \cup \alpha_2$	$\beta = \beta_1 \cup \beta_2$	γ	$\delta \cup \varepsilon$
Students in k-means cluster	$(31)Cl_2+(5)Cl_1$	$(33)Cl_1+(4)Cl_2$	$(28)Cl_1$	$(13)Cl_3+(2)Cl_2+(1)Cl_1$
Most frequently given answers	1A, 2C, 3D, 4A, 5A, 6B	1C, 2D, 3B, 4E, 5B, 6E	1C, 2E-G, 3D-E, 4F, 5E, 6G	1E, 2H, 3F, 4H, 5G, 6H

CONCLUSION - COMPARING CLUSTERING METHODS

It is possible to conclude that the *NH-ClA* method we discussed here allows the researcher to easily obtain and visualize in a 2-D graph a global view of student behavior with respect to the answers to a questionnaire and to obtain a first characterization of student behavior in terms of their most frequently used answering strategies.

The *H-ClA* method, on the other hand, although producing a graph that is not as easy to read as the one produced with the other method, allows the researcher to obtain results coherent with the *NH-ClA* ones and that may offer a finer grain detail of student behavior.

MESE 1

Measurement in STEM Education (MESE1)

Naples, 30-31 January, 1 February

DISP Dipartimento di Scienze Politiche Università degli studi di Napoli Federico II

Università degli Studi di Palermo

DiFC Dipartimento di Fisica e Chimica - Emilio Segrè

THANK YOU

Onofrio Rosario Battaglia

Dipartimento di Fisica e Chimica – Emilio Segrè, University of Palermo, Palermo, Italy

BIBLIOGRAPHIC REFERENCES

- Battaglia, O. R., Di Paola, B. Persano Adorno, D., Pizzolato N. and Fazio C. 2019 Evaluating the effectiveness of modelling-oriented workshops for engineering undergraduates in the field of thermally activated phenomena. Res. Sci. Ed. 49, 1395–1413.
- Borg I. and Groenen P. 1997 Modern multidimensional scaling. New York: Springer Verlag.
- Di Paola, B., & Collura, D. (2020). Collaborative Teaching in the Italian" Liceo Matematico": A Case Study of Co-Planning and Co-Teaching. In ICMI Study 25-Study Conference-Teachers of Mathematics Working and Learning in Collaborative Groups (Vol. 25, pp. 278-285). INTERNATIONAL COMMISSION ON MATHEMATICAL INSTRUCTION EDITORS.
- Haslina, N., Huria, D. and Karpudewan, M. 2019 Evaluating the effectiveness of Integrated STEM-lab activities in improving secondary school students' understanding of electrolysis. Chem. Educ. Res. Pract. 20, 1395–1413.
- Li, Y., Wang, K., & Xiao, Y. (2019). Exploring the status and development trends of STEM education research: A review of research articles in selected journals published between 2000 and 2018. **数学教育学**报(Journal of Mathematics Education), 28(3), 45–52.
- Li, Y., Wang, K., Xiao, Y. et al. Research and trends in STEM education: a systematic review of journal publications. IJ STEM Ed 7, 11 (2020). https://doi.org/10.1186/s40594-020-00207-6
- Lin L., Lee T. and Anderson Snyder L. 2018 Math Self-Efficacy and STEM Intentions: A Person-Centered Approach. Front. Psychol. 23.

BIBLIOGRAPHIC REFERENCES

- Manduca, C. A., Iverson, E. R., Luxenberg, M., Macdonald, R. H., McConnell, D. A., Mogk, D. W. and Tewksbury, B. J. 2017. Improving undergraduate STEM education: The efficacy of discipline-based professional development. Sci. Adv. 15.

- National Science Foundation (1998). Information technology: Its impact on undergraduate education in science, mathematics, engineering, and technology. (NSF 98–82), April 18–20, 1996. http://www.nsf.gov/cgi-bin/ getpub?nsf9882 Accessed 16 Jan 2018.