CLUSTERING EDUCATIONAL DATA: A HIGH SCHOOL STUDENTS' PERFORMANCE ANALYSIS

Matteo Farnè and Gioia Taraborrelli

Department of **Statistical Sciences**,

University of **Bologna**

Measurement in STEM Education (MESE1)

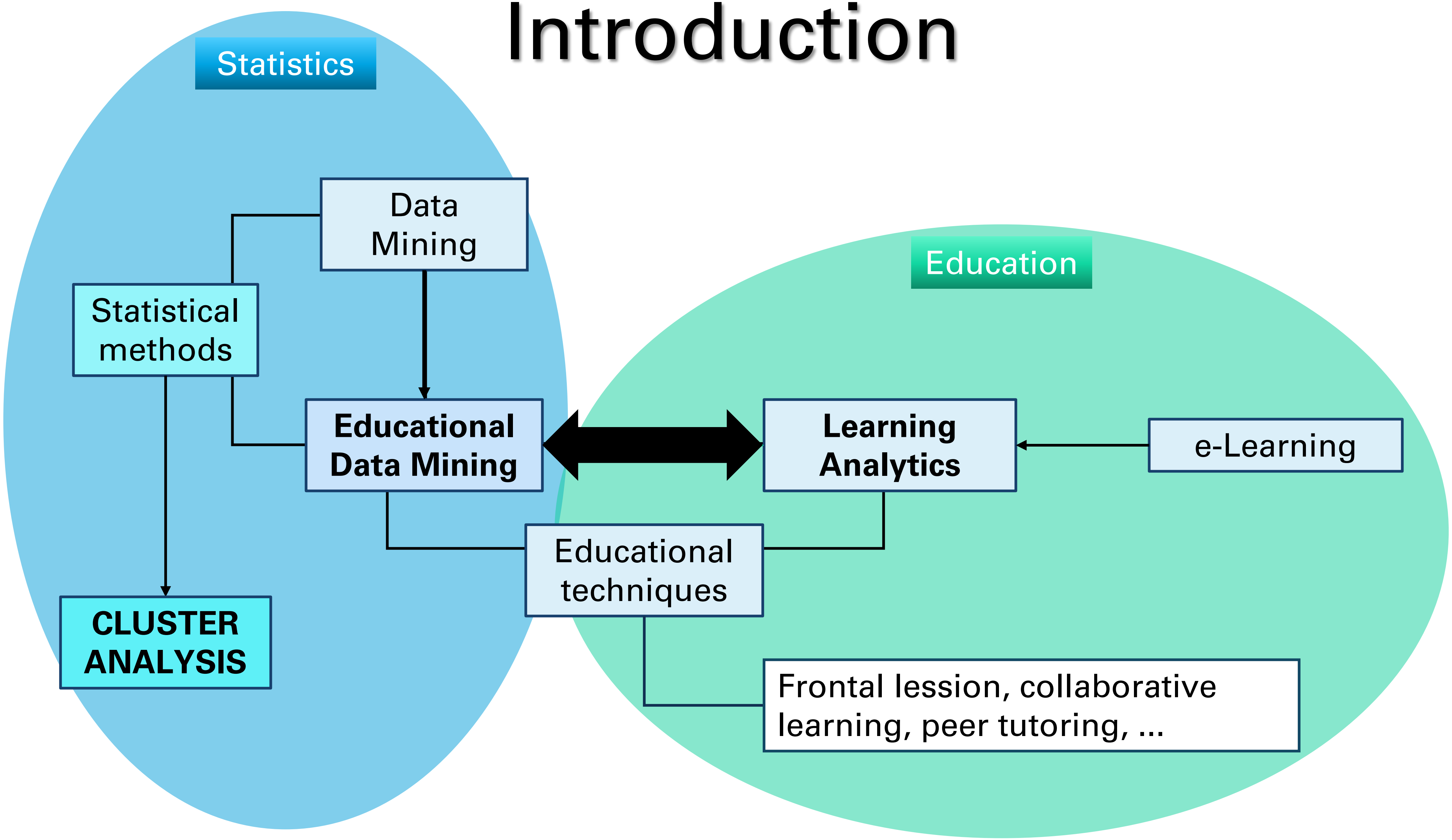
Napoli, 30-31 January 2023

Introduction

- ✓ <u>Aim of the research:</u> clustering **students' performance** through time using **statistical methods** for data mining.
- ✓Statistical methods for **educational data** may have multiple purposes, which may be grossly divided in **two macro-categories**: uncovering hidden patterns from educational data, or measuring how and how much students learn.
- ✓ The first purpose is described by the label **EDUCATIONAL DATA MINING**, the second by the label **LEARNING ANALYTICS.**

➢ *Farnè, M.; Taraborrelli, G. (2023). Come sfruttare gli Educational Data? Un inquadramento di usi e metodologie di analisi. Induzioni, 62/63, 1-2, forthcoming.*

Introduction



Educational Data Mining

«Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly **large-scale data** that come from educational settings and using those methods to **better understand students, and the settings** which they learn in» ([https://educationaldatamining.org/)](https://educationaldatamining.org/)

Some of the **Data Mining methods** applied in EDM (Villanueva et al., 2018):

- Markov Chains
- Classification algorithms
- **Cluster Analysis**
- Bayesian networks
- Association rules
- Linear regression

Main **stakeholders** in EDM (Romero, Ventura, 2010):

- Students
- Tutors and educators
- Researchers
- Schools and universities
- Organizations and administrators

➢ Villanueva, A.; Moreno, L. G.; Salinas, M. J. (2018). *Data mining techniques applied in educational environments: Literature review*. Digital Education Review - Number 33, June 2018.
➢ Romero, C.; Ventura, S. (2010). *Educational Data Mining: A Review of the State-of-the-Art*. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(6), 601-618.

Educational Data Mining

Student model: how does a typical student learn?

Domain model: how is a specific subject typically learnt?

✓ Develop **tailor-based** student **learning**.

✓ Identify different **student types**.

✓ **Stimulate students** in a **better** way.

✓ **Improve** study **materials**.

✓ **Personalize** and monitor the **learning** experience.

✓ **Improve** efficiency and effectiveness.

- Villanueva, A.; Moreno, L. G.; Salinas, M. J. (2018). *Data mining techniques applied in educational environments: Literature review*. Digital Education Review - Number 33, June 2018.
- Romero, C.; Ventura, S. (2010). *Educational Data Mining: A Review of the State-of-the-Art*. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. 40(6), 601-618.

Educational Data Mining
(Villanueva et al., 2018)

Technique \ Domain	Dropping out or Retention Analysis	VLO or VLE Analysis	Performance and students evaluation Analysis	Generation of Educational Recommendations	Learning pattern Identification	Students patterns Identification	Students related Prediction
Correlation Analysis		1				1	
Decision Trees	5	3	8	2	2	6	2
Regression Trees			1				
Markov Chains				1			
Classification	4	2	4		3	1	3
Clustering		7	3	5	3	9	2
Differential Sequence Mining						1	
Sequential Patterns		4	1	1	7	3	2
Bayesian Networks		2	1		1	1	6
Neural Networks	1	2	2		1		5
Association rules		8	1	7	14	9	1
Linear regression					1		1

- Villanueva, A.; Moreno, L. G.; Salinas, M. J. (2018). *Data mining techniques applied in educational environments: Literature review*. Digital Education Review - Number 33, June 2018.

Learning Analytics

«Learning Analytics (LA) is defined as the process of **analyzing educational data** which includes the measurement, collection, analysis and reporting of data on students and the school context, to **understand** and **optimize learning** and environment in which they learn» (Lang et al., 2017)

Learning Analytics methods **fit** (Reimann, 2016):

- **High volume** data
- **Longitudinal** data
- Data from **different sources**
- Data from **different levels of learning** (during time)

Most used methods in LA (Avella et al., 2016):

- Data visualization techniques
- Social Network Analysis
- Predictive Models
- **Cluster Analysis**
- Relationship Mining
- Discovery with Models

- Lang, C., Siemens, G., Wise, A., & Gasevic, D. (2017). *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research. New York, NY: SOLAR
  In Romero, C.; Ventura, S. (2020). *Educational data mining and learning analytics: An updated survey*. WIREs Data Mining Knowl Discov. 2020;10:e1355.
- Reimann, P. (2016). *Connecting learning analytics with learning research: the role of design-based research*. Learning: Research and Practice, 2(2), 130–142.
- Avella, J. T.; Kebritchi, M.; Nunn, S. G.; Kanai, T. (2016). *Learning Analytics Methods, Benefits, and Challenges in Higher Education: A Systematic Literature Review*. Online Learning, Volume 20 Issue 2, June 2016.

Educational Data Mining
vs.
Learning Analytics

Points **in common**:
✓Education as main application
✓Data-intensive approaches to education research
✓Goal of enhancing educational practice

Differences: association VS prediction

Educational Data Mining	Learning Analytics
Automated methods to analyse data	Central role of human judgement
Reductionist focus	Holistic focus
Automated adaptation	Support human intervention
Learning as a research topic	Aspects of education as a research topic

➢ Baker, R. S.; Inventado, P. S. (2014). *Educational Data Mining and Learning Analytics*. Learning Analytics (pp.61-75).

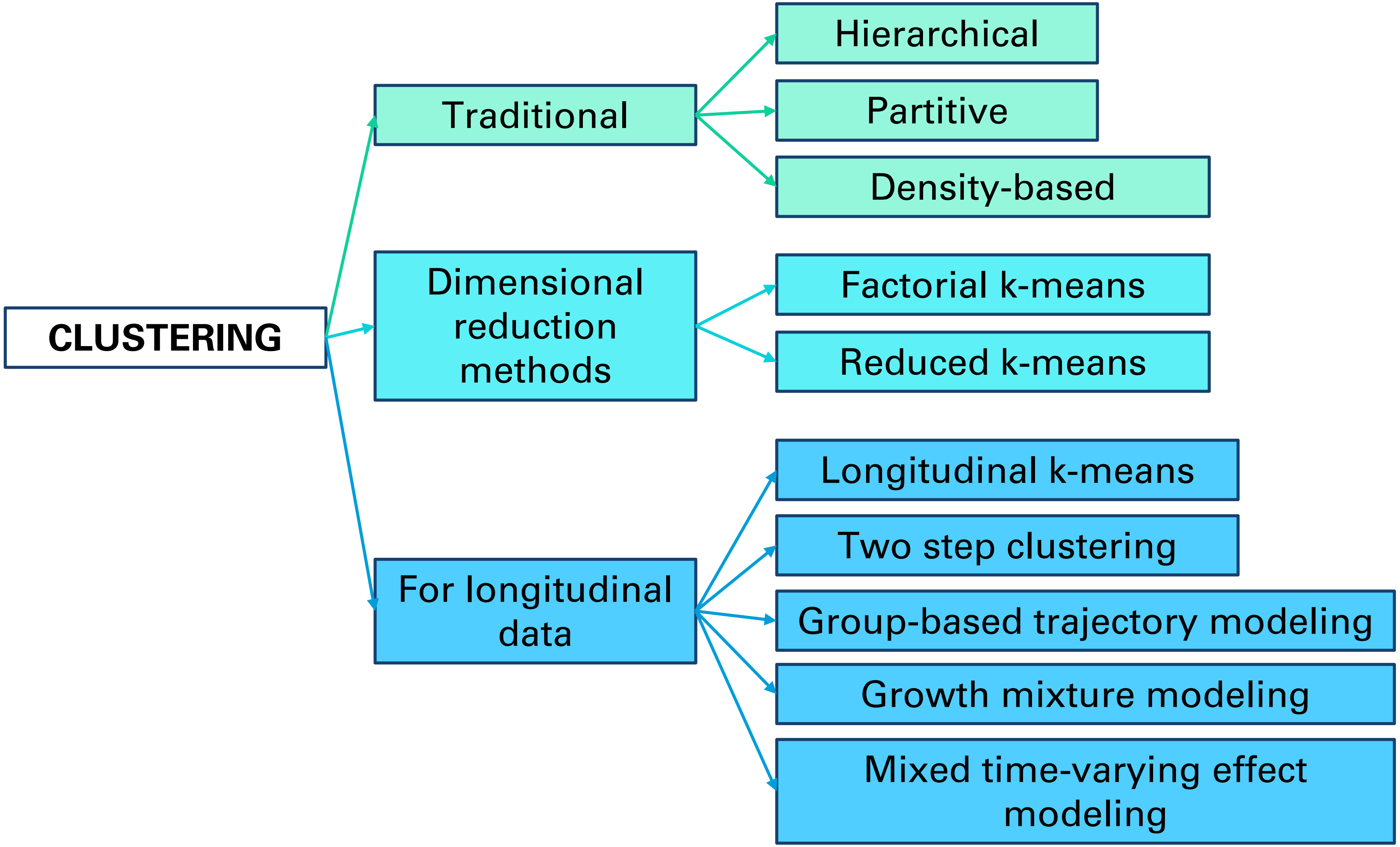
Cluster Analysis

Method used for data analysis: **Cluster Analysis.**

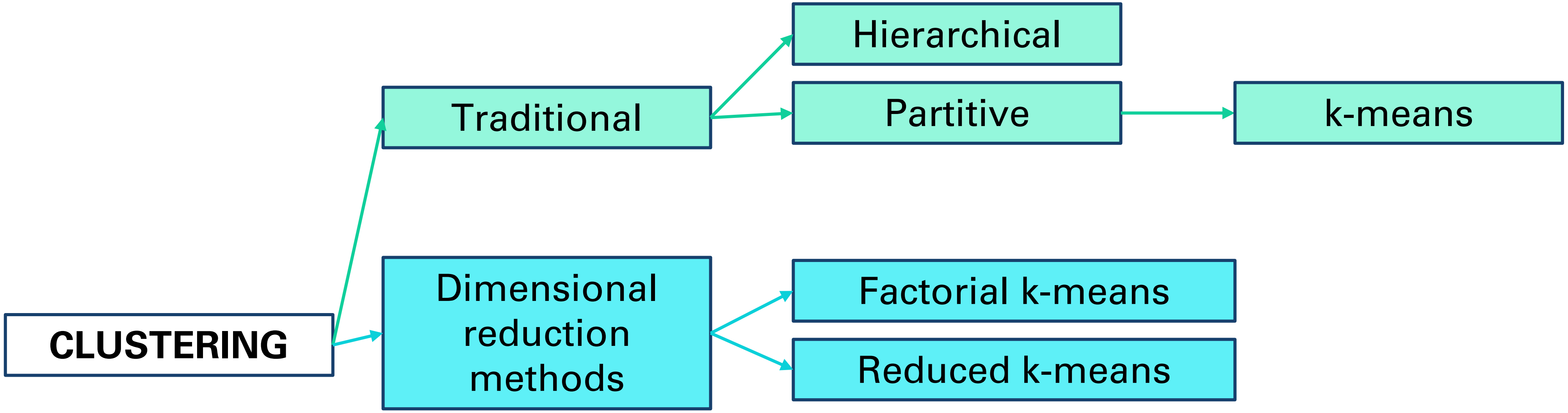
«Cluster Analysis is an unsupervised learning technique aimed to group n statistical units respect to p variables in a certain number of groups, such that units which belong to the same group are similar to each other and are different to units which belong to other groups»

- Cerioli, A.; Zani, S. (2007). *Analisi dei dati e data mining per le decisioni aziendali*. Giuffrè Editore.

Cluster Analysis



Cluster Analysis



Dataset

- n = 17 students of a high school (classical studies curriculum)
- t = 3 quarters → dataset is divided into 3 distinct sub-datasets, one for each period of time
  - 1st quarter of the 1st year of high school (t=1)
  - 2nd quarter of the 1st year of high school (t=2)
  - 1st quarter of the 2nd year of high school (t=3)
- p = 5 variables observed in each period (15 variables in total):
  - Grade in Maths
  - Grade in Italian
  - Grade in Greek
  - Grade in Latin
  - Hours of absence
- Analysis were carried out via R and the algorithms were applied to each sub-dataset individually.

Dataset

- n = 17 students of a high school (classical studies curriculum)
- However there was an outlier, with higher hours of absence.
  - Through trimmed k-means algorithm results were compared in each period with and without outlier, in terms of average silhouette width (ASW), which is a reliability measure of the partition of the groups obtained.
  - The outlier was removed because values of ASW were better, or at least similar, without it.
- In conclusion, n = 16 students.

Cluster Analysis techniques

1) Average Linkage Method (see Hastie et al., 2009)
2) Partitive Cluster Analysis → k-means (MacQueen, 1967)
3) Reduced k-means (De Soete and Carroll, 1994)
4) Factorial k-means (Vichi and Kiers, 2001)

Number of groups (k)?

- ✓ k=3 groups in t=1 e t=2, k=2 in t=3 →hierarchical clustering, k-means, reduced k-means.
- ✓ k=2 groups in each period→ factorial k-means.

➢ These methods were compared in terms of average silhouette width.

➢ Hastie, T.; Tibshirani, R., Friedman, J.. (2009). *The Elements of Statistical Learning.* New York: Springer.
➢ Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49-64.
➢ De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In *New approaches in classification and data analysis* (pp. 212-219). Springer Berlin Heidelberg.
➢ J. B. MacQueen (1967): *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

Results

- Membership were identical in t=1 and t=2 using average linkage, k-means and reduced k-means;
- Membership were identical using k-means e reduced k-means (which represents a generalization of the first method);
- Average linkage generates groups with an unbalanced numerosity in t=3 (2nd year);
- Using factorial k-means the structure of data is not understood→ forced interpretation of membership of the groups. This was confirmed by the low values of ASW.

Results

- Average linkage method → partition in t=3 has an unbalanced number of members per group.

- Factorial k-means is a method too complex to be applied to the sub-datasets, which were composed of few observations (n=16) and few variables (x=5) for each sub-dataset.

Average linkage method - partitions

Membership in t=1 and t=2 are the same:
Group 1 (n=9)→ students with fairly good grades overall (no failures) and many hours of absence
Group 2 (n=4) → students with high grades overall and few hours of absence
Group 3 (n=3) → students with low grades overall and many hours of absence

Membership in t=3:
Group 1 (n=14) → students with few hours of absence, higher grades in Maths and Italian, lower grades in Greek and Latin
Group 2 (n=2) → students with many hours of absence, lower grades in Maths and Italian, higher grades in Greek and Latin

ID	Membership t=1	Membership t=2	Membership t=3
ERSA	1	1	1
LIRI	2	2	1
MABE	3	3	1
TIRA	1	1	1
TITE	1	1	1
DOCI	2	2	1
CERI	1	1	1
PITA	3	3	1
AMSA	2	2	1
NORA	2	2	1
MANA	1	1	2
BENE	1	1	1
RIRE	1	1	1
RECI	1	1	2
ETNI	3	3	1
DEAV	1	1	1

K-means clustering - partitions

Membership in t=1 and t=2 are the same:
Group 1 (n=9)→ students with fairly good grades overall (no failures) and many hours of absence
Group 2 (n=4) → students with high grades overall and few hours of absence
Group 3 (n=3) → students with low grades overall and many hours of absence

Membership in t=3:
Group 1 (n=7) → students with many hours of absence, lower grades in Maths and Italian, higher grades in Greek and Latin
Group 2 (n=9) → students with few hours of absence, higher grades in Maths and Italian and lower grades in Greek and Latin

ID	Membership t=1	Membership t=2	Membership t=3
ERSA	1	1	2
LIRI	2	2	2
MABE	3	3	1
TIRA	1	1	2
TITE	1	1	2
DOCI	2	2	2
CERI	1	1	1
PITA	3	3	1
AMSA	2	2	2
NORA	2	2	2
MANA	1	1	1
BENE	1	1	1
RIRE	1	1	1
RECI	1	1	1
ETNI	3	3	2
DEAV	1	1	2

Reduced k-means – partitions

Membership in t=1 and t=2 are the same **(dimension = grade point average)**:
Group 1 (n=9)→ students with fairly high grades overall
Group 2 (n=4) → students with high grades overall
Group 3 (n=3) → students with loe grades

Membership in t=3 (dimension = level of ability in basic subjects, i.e. Maths and Italian):
Group 1 (n=9) → students with higher grades in basic subjects and lower grades in Greek and Latin
Group 2 (n=7) → students with lower grades in basic subjects and higher grades in Greek and Latin

Notice that the results are the same of k-means algorithm (in t=3 groups are inverted).

ID	Membership t=1	Membership t=2	Membership t=3
ERSA	1	1	1
LIRI	2	2	1
MABE	3	3	2
TIRA	1	1	1
TITE	1	1	1
DOCI	2	2	1
CERI	1	1	2
PITA	3	3	2
AMSA	2	2	1
NORA	2	2	1
MANA	1	1	2
BENE	1	1	2
RIRE	1	1	2
RECI	1	1	2
ETNI	3	3	1
DEAV	1	1	1

Factorial k-means – partitions

Membership in t=1 (dimension = level of ability in basic subjects, i.e. Maths and Italian):
Group 1 (n=11)→ students with heterogeneous performance
Group 2 (n=5) → students with higher grades in Maths and Italian

Membership in t=2 (dimension = grade point average):
Group 1 (n=8)→ students with medium-low grades
Group 2 (n=8) → students with higher grades (no failures

Membership in t=3 (dimension = difficulty in Latin):
Group 1 (n=10) → students with very low grades in Latin (with two exceptions however)
Group 2 (n=6) → students with failures only in Latin

ID	Membership t=1	Membership t=2	Membership t=3
ERSA	1	1	2
LIRI	1	2	1
MABE	1	1	1
TIRA	2	1	1
TITE	2	2	1
DOCI	1	2	1
CERI	1	1	2
PITA	1	1	2
AMSA	1	2	2
NORA	1	2	1
MANA	2	1	1
BENE	1	2	1
RIRE	2	2	1
RECI	2	1	2
ETNI	1	1	1
DEAV	1	2	2

Conclusions

- The most acceptable and understandable results were obtained through reduced k-means and k-means in terms of ASW because these methods understood better data structure (few observations and few variables).
  - ➢ Groups seem to compact from the 1st year to the 2nd year of high school → from k=3 to k=2.
  - ➢ Moreover, RKM highlights the presence of a latent dimension which justifies the obtained partition, namely, the **ability in basic subjects**.
- Ideas for future:
  - ❑ Use other (**dynamic**) statistical methods to analyse educational dataset;
  - ❑ Measure the impact of teaching methodologies alternative to frontal lession, like **machine-learning** based ones;
  - ❑ Repeat the analysis considering the **whole year**, and not quarters, as period of time, and gather **more variables**.