




How to Measure the Researcher Impact with the Aid of its Impactable Area: A Concrete Approach Using Distance Geometry

Beniamino Cappelletti-Montano¹ · Gianmarco Cherchi¹ · Benedetto Manca¹  · Stefano Montaldo¹ · Monica Musio¹

Accepted: 26 July 2024
© The Author(s) 2024

Abstract

Assuming that the subject of each scientific publication can be identified by one or more classification entities, we address the problem of determining a similarity function (distance) between classification entities based on how often two classification entities are used in the same publication. This similarity function is then used to obtain a representation of the classification entities as points of an Euclidean space of a suitable dimension by means of optimization and dimensionality reduction algorithms. This procedure allows us also to represent the researchers as points in the same Euclidean space and to determine the distance between researchers according to their scientific production. As a case study, we consider as classification entities the codes of the American Mathematical Society Classification System.

Keywords Bibliometrics · Dimensionality reduction · Linear programming

1 Introduction

The use of bibliometric indicators is becoming more and more pervasive in academic and scientific life. Despite various standpoints from scientific associations and institutions — one over all, the DORA declaration (see Cagan, 2013) — nowadays bibliometrics plays a

Beniamino Cappelletti-Montano, Gianmarco Cherchi, Benedetto Manca, Stefano Montaldo, and Monica Musio contributed equally to this work.

✉ Benedetto Manca
bmanca@unica.it

Beniamino Cappelletti-Montano
b.cappellettimontano@unica.it

Gianmarco Cherchi
g.cherchi@unica.it

Stefano Montaldo
montaldo@unica.it

Monica Musio
mmusio@unica.it

¹ Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, Cagliari 09124, Italy

central role both for university rankings and for the careers of individual researchers. For instance, the Italian higher education system explicitly involves a rigid use of bibliometric indicators both for allocating funding among universities and for regulating promotions to associate/full professorships (see e.g., Abramo and D'Angelo, 2015; Baccini et al., 2019; Cappelletti-Montano et al., 2021).

In order to apply correctly and properly a bibliometric indicator one has to ensure that the object to which that bibliometric index is applied — a university, a department, or a single researcher — is compared within similar entities. For instance, it makes no sense to compare the h -index of a biologist with that of a physicist.

There have been several attempts at taking into account the appropriate context for “measuring” bibliometric indicators associated to an article (and hence to a researcher or to an aggregate of researchers) or at developing a new system of unbiased qualitative judgment of research publications (Murtagh et al., 2018). The simplest choice is to consider the subject categories associated with the journal where the article is published. In Ioannidis et al. (2019), using SCOPUS, a database of top 2% authors for each scientific field and subfield has been created (the database was recently updated in Ioannidis, 2022). Web of Sciences database (<http://webofscience.help.clarivate.com/en-us/Content/author-record.html>) recently implemented a new function, named *Author Impact Beamplot*, which shows, for each year, the best citation percentile of the author's articles with respect to all articles published in that year in all the Journal Citation Reports (JCR) categories where the author's articles are published. This is based on early ideas from Bornmann and Marx (2014) (see also Bornmann and Haunschild, 2018; Haunschild et al., 2019).

Although suggestive, these proposals have several limitations. The main issue is that the “subject categories” are often too wide and they do not allow to compare correctly articles and authors. For instance in the subject category “mathematics,” we can find disciplines which have very different behaviors from the bibliometric point of view: there are articles in abstract algebra, which usually gain few citations, and articles in applied/computational mathematics which receive many more citations.

Another interesting proposal is to weight each citation received according to the average number of bibliography lengths (actually, the number of active references) of the articles in the journal in which the article, from which the citation came, is published (see for instance Waltman and van Eck, 2013, and references therein).

In order to use bibliometric indicators for researchers' careers, the Italian higher education system required that each professor must belong to one among the 190 competition sectors in which all disciplines were rigidly subdivided by Italian legislation. For each of such competition sectors some thresholds for the number of articles, total citations, and h -index were required as necessary conditions for becoming an associate or full professor. However, also this practice did not prevent the occurrence of the same aforementioned problems. For instance, the competition sector “01/B1 - Computer Sciences” includes professors working on logic/theoretical computer science, who traditionally have low bibliometric indexes, and professors working on machine learning or bioinformatics, who have, on average, very high numbers of articles and citations (cf. Demetrescu et al., 2020). A common point in any of the above proposals is to consider “close” and hence “comparable” any two authors whose scientific activities fall in a pre-specified field. This intuitive idea of “closeness” can be mathematically treated by using the notion of distance. In other terms, one should define a distance function which associates to any couple of researchers A and B a positive number telling how far A and B are. This problem is related to the topics of the delineation of scientific fields and the classification of scholarly journals. We do not deal with these topics, and refer the reader to the Glänzel et al. (2019) and Baccini et al. (2022), and references therein.

In this article, we address the problem of defining the above distance function under the assumption that to each scientific article, it is attached a code (or possibly more than one) that classifies the article's topics in a very detailed way. This condition may appear too restrictive, however, there are several contexts where it holds: we can mention, for instance, the well-known Mathematics Subject Classification (MSC), Physics and Astronomy Classification Scheme (PACS), Computing Classification System (CCS), Journal of Economic Literature (JEL) classification codes for the fields, respectively, of mathematics, physics/astronomy, computer sciences, and economics. On the other hand, it is abstractly possible to define a classification of scientific articles in any field using the title or the keywords. A proposal in this direction, named SciVal Topic Prominence, was recently implemented in the SCOPUS database (https://service.elsevier.com/app/answers/detail/a_id/27947/supporthub/scopus/).

Then, by means of an optimization and geometric deep learning process, we are able to represent the codes as points of a 3-dimensional space endowed with a distance function. This mathematical construction, which takes into account the frequency of articles sharing the same code, allows us to give a quantitative measure of the distance of two codes. We extend this procedure to authors, and in this way, we are able to consider a neighborhood of a given author, i.e., the set of authors which are close (hence comparable) to him/her.

We test this construction for the field of mathematics, using the database "Zentralblatt Math" where the MSC classification is used for each article.

2 Representation of Classification Entities and Authors

In this section with classification code, we mean an alphanumeric string used in a scientific database to identify a subject in a given discipline. For instance, in the MSC a classification code is of type **12A26**, while in the PACS is typically written as **02.40.Hw**.

The aim of this paragraph is to find a good way to represent a set of classification codes as a set of points in an Euclidean space \mathbb{R}^n , for some suitable dimension n . With good, we indicate that similar codes, in some sense that we will explain soon, are represented as close points in \mathbb{R}^n .

2.1 The Similarity Function for Classification Entities

Let \mathcal{C} be a set containing N classification codes and let \mathcal{A} be the set of all scientific articles in a given database to which at least one of the codes in \mathcal{C} has been assigned. If $c \in \mathcal{C}$ and $a \in \mathcal{A}$, we shall write $c \in a$ meaning that the article a is assigned to the scientific code c .

We can then define a matrix $F = (F_{i,j}) \in \mathbb{R}^{N \times N}$ such that:

$$\begin{aligned} F_{i,j} &= F(c_i, c_j) = \#\{a \in \mathcal{A} : c_i, c_j \in a\} \quad i \neq j \\ F_{i,i} &= F(c_i, c_i) = \#\{a \in \mathcal{A} : c_i \in a\} \end{aligned} \quad (1)$$

Thus, for $i \neq j$, $F_{i,j}$ represents the number of articles that share at least the codes c_i and c_j , while $F_{i,i}$ the number of articles that have at least the code c_i .

Our goal is to establish a meaningful distance between two codes, c_i and c_j , capable of discerning the extent of their relationship. Specifically, when codes c_i and c_j exhibit a close association, meaning the presence of one in articles almost invariably coincides with the presence of the other and vice versa, the resulting distance should be zero. Conversely, when the two codes are unrelated, and almost never concurrently featured in articles, the distance should be maximal. To formalize this notion, we propose the adoption of the following

function where to ensure values fall within the $[0, 1]$ interval, we assign a distance of 1 when the codes are unrelated.

In details, using the matrix F , we can then define a similarity function $d : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}^+$ as follows:

$$d(c_i, c_j) := 1 - \min \left\{ \frac{F_{ij}}{F_{ii}}, \frac{F_{ij}}{F_{jj}} \right\}. \quad (2)$$

We observe that the quantity $d_{ij} := d(c_i, c_j)$ is close to zero when the number of articles containing c_i and c_j is close to the number of articles containing only c_i or only c_j . On the other hand, d_{ij} is close to 1 when the number of articles containing c_i and c_j is small with respect to the number of articles containing c_i .

While the similarity function provides insights into the relationships between codes, it falls short in revealing the frequency of a single code, c_i . To address this limitation, we introduce a conceptual universal code, denoted by O , as described below. This universal code is postulated to be present in all articles under consideration. Then, the distance d , between a code c_i and the universal code O should gauge the frequency F_{ii} in the following manner: when $d(O, c_i)$ approaches zero, it indicates that c_i is present in almost all articles; conversely, if c_i is infrequently present in an article, the distance value should converge toward 1, signifying its rarity in the dataset.

We thus define the universal code O by the condition $F(O, O) = \sum F_{ii}$. The formula

$$d(O, c_j) = 1 - \min \left\{ \frac{F(O, c_j)}{F(O, O)}, \frac{F(O, c_j)}{F_{jj}} \right\} = 1 - \min \left\{ \frac{F_{jj}}{\sum_i F_{ii}}, \frac{F_{jj}}{F_{jj}} \right\} = 1 - \frac{F_{jj}}{\sum_i F_{ii}},$$

guarantees that the universal code O precisely fulfill the intended purpose within the scope we have introduced.

To simplify the notation, we denote with $\tilde{\mathcal{C}}$ the set $\mathcal{C} \cup \{O\}$ and with \tilde{D} the $(N+1) \times (N+1)$ matrix with entries $\tilde{D}_{ij} := d(c_i, c_j)$ when $i, j \in \{1, \dots, N\}$ and $\tilde{D}_{i(N+1)} = \tilde{D}_{(N+1)i} := d(O, c_i)$, $j \in \{1, \dots, N\}$.

2.2 Representation of Classification Entities as Vectors

Using the similarity matrix \tilde{D} defined in the previous section, we want to define a representation map $R : \tilde{\mathcal{C}} \rightarrow \mathbb{R}^k$ for some $k \in \mathbb{N}$, so that the Euclidean distance between any two points $R(c_i)$ and $R(c_j)$ is equal to the similarity value \tilde{D}_{ij} .

Finding the position of elements knowing their relative distances is often referred to as the distance geometry problem (DGP). The DGP arises in different contexts depending on the dimension k of the representation considered. Some of the most common applications are wireless sensor networks for $k = 1, 2$, or 3 (Singer, 2011), determination of protein structure from nuclear magnetic resonance experiments for $k = 3$ (Bahr et al., 2009; Tabaghi et al., 2019), controlling fleets of underwater autonomous vehicles for $k = 3$ (Wüthrich, 1989) (more details on how to solve the DGP are given in Appendix A).

In order to apply the distance geometry approaches, described so far in general, to our situation, we first need to define a simple undirected graph G . We let the vertices set V be equal to the classification entities set $\tilde{\mathcal{C}}$, the edges set $E = \{(i, j) \mid \tilde{D}_{ij} \neq 1 \text{ or } \tilde{D}_{ij} \neq 0, i, j \leq N+1\}$ and the weight function given by the values of the matrix \tilde{D} .

By factorizing the solution of the DGP, we obtain a representation of the classification entities $\tilde{\mathcal{C}}$ in \mathbb{R}^{N+1} which we will denote with $\tilde{\mathcal{C}}^{N+1}$. Since $N+1$ can be very large, we apply a dimensionality reduction algorithm to $\tilde{\mathcal{C}}^{N+1}$ as described in Appendix A to obtain a representation $\tilde{\mathcal{C}}^k$ of the classification entities in \mathbb{R}^k with $k \ll N+1$. In the following, we will

describe a first attempt to use the representation \tilde{C}^k to obtain a representation of researchers based on their scientific production so that researchers publishing in the same field are close to each other.

2.3 Representation of Researchers

Given a representation \tilde{C}^k of the considered classification entities in \mathbb{R}^k , for some dimension k , we consider a set \mathcal{R} of researchers and we assume that it is possible to associate to every element $r \in \mathcal{R}$ a subset $\tilde{C}_r^k \subseteq \tilde{C}^k$ according to the scientific production of the researcher; for example, if \tilde{C} is the set of the American Mathematical Society (AMS) classification codes, we associate to a researcher every code that has ever been used in all the articles published by the researcher.

For any $r \in \mathcal{R}$, we would like to obtain a representation \mathbf{r} of r as an element of \mathbb{R}^k , i.e., the same space where the classification entities are represented. Perhaps, one of the most intuitive approaches is to consider the following weighted sum

$$\mathbf{r} := \sum_{\mathbf{c} \in \tilde{C}_r^k} w_c \mathbf{c}, \quad (3)$$

where the weights w_c allow to give more importance to the classification entities that better represent the researcher. For example, w_c can be equal to the number of scientific articles published by the researcher under the classification entity c .

The idea is that given two researchers r_1 and r_2 , the corresponding points \mathbf{r}_1 and \mathbf{r}_2 in \mathbb{R}^k are as close to each other as the scientific production of r_1 and r_2 are similar according to the classification system given by \tilde{C} . Therefore, it seems reasonable to evaluate the publications of a researcher r by comparing r with only the researchers whose representation in \mathbb{R}^k is closest (in some sense) to the representation of r .

For example, one could find the smallest ball around \mathbf{r} which contains an appropriate number of classification entities, and compare the researcher r only with the researchers whose representation in \mathbb{R}^k lies inside this ball. We test this approach in the case study of Section 3 proving that in some fields it provides good results.

3 Case Study: The American Mathematical Society Mathematics Subject Classification

In order to validate our theoretical approach, we considered the American Mathematical Society mathematics subject classification system as a classification dictionary for research articles.

The MSC consists in a set of alphanumerical codes representing several (almost all) mathematical fields. Each code is divided into three hierarchical layers:

- (i) the first level is represented by a 2-digit number, one for each mathematical discipline, e.g., **53** for differential geometry, **23** for real function analysis;
- (ii) the second level is represented by a Latin letter which indicates a specific area covered by the first level discipline, e.g., **53A** for classical differential geometry;
- (iii) the third level is represented by a 2-digit number, one for each specific kind of mathematical object, well-known problem, or research area in the field identified by the first two levels, e.g., **53A05** is the code representing surfaces in Euclidean or related space.

Usually, every article submitted to a journal indexed by the AMS is classified by one or more MSC codes (either chosen by the authors or by the editors and validated by the AMS reviewers). This classification system is available in both the AMS (<https://mathscinet.ams.org/mathscinet/>) and Zentralblatt (<https://zbmath.org>) repositories.

Since the latter offers open-API (Petraera et al., 2021) to collect information on the articles contained in the repository, we used a small Python script to retrieve the number of articles classified by one or two MSC elements. In order to maintain the computational cost of the DGP instance acceptable, we considered the 1863 codes representing mathematical analysis or geometry subjects. This choice is also justified by the consideration that researchers dealing with mathematical analysis and geometry represent the vast majority of those dealing with pure mathematics. Moreover, there are several researchers with a research production between these two areas, a fact that will be important to test our analysis.

In accordance with the notation used in Section 2, we denote with \mathcal{C} the set of classification codes under consideration, with F the $N \times N$ matrix such that $F_{ij} = \#\{\text{articles with codes } c_i, c_j\}$ (with $N=1863$) and with D the matrix containing the similarity values between codes as defined by Eq. 2.

As in Section 2, we denote with $\tilde{\mathcal{C}}$ the set obtained by adding the universal code O to \mathcal{C} and \tilde{D} the matrix obtained by adding a row and a column to D to include the similarity values between O and the other elements in \mathcal{C} .

We then solve the DGP instance with input data N and \tilde{D} using the linear reformulation (9) to obtain a representation of the MSC codes Y in \mathbb{R}^N . Since N is quite large, we have applied several dimensionality reduction algorithms to obtain a representation of the codes in \mathbb{R}^3 . Since we are also interested in visualizing the representations of the AMS codes and the researchers, we tested our approach with $k = 2, 3$. However, the quality of the results obtained with $k = 2$ was significantly poorer than the one obtained using $k = 3$ in terms of the clustering of the researchers, while the AMS codes representations for $k = 2$ and $k = 3$ are not particularly different. We have reported the results obtained for $k = 2$ in Appendix 2. Therefore, in the rest of the article, we will only consider representations in \mathbb{R}^3 .

In our analysis, we have considered the following dimensionality reduction algorithms (we refer to Appendix A and the references therein for more details): principal component analysis (PCA), random projections (RP), linear discriminant analysis (LDA), isometric feature mapping (ISOMAP), uniform manifold approximation and projection (UMAP), multi-dimensional scaling (MDS) having as input the solution Y of the DGP problem (MDS¹) and the dissimilarity matrix \tilde{D} (MDS²).

Since we solve the DGP in an approximate way (see Appendix A for more details), it is important to evaluate the solution obtained with respect to the distances encoded in \tilde{D} . Two scores to evaluate the quality of the DGP solution are given by the maximum deviation error (MDE) and largest deviation error (LDE) defined by

$$\begin{aligned} \text{MDE} &:= \frac{1}{|E|} \sum_{\{i,j\} \in E} \frac{|\|y_i - y_j\|_2 - \tilde{D}_{ij}|}{\tilde{D}_{ij}} \\ \text{LDE} &:= \max_{\{i,j\} \in E} \frac{|\|y_i - y_j\|_2 - \tilde{D}_{ij}|}{\tilde{D}_{ij}}, \end{aligned} \quad (4)$$

where E is the set of edges of the graph G used as input for the DGP instance.

Table 1 shows the values of MDE and LDE for the representation of $\tilde{\mathcal{C}}$ in \mathbb{R}^N and the representations in \mathbb{R}^3 obtained using different dimensionality reduction algorithms.

Table 1 The table shows the values of MDE and LDE computed on the original representation in \mathbb{R}^N using the DGP approach and the representations in \mathbb{R}^3 obtained using dimensionality reduction algorithms

Representation	MDE	LDE
Original	0.0079	5.39×10^{-6}
PCA	0.0065	3.39×10^{-6}
RP	0.0019	5.19×10^{-6}
ISOMAP	0.0061	8.03×10^{-6}
MDS ¹	0.0018	3.23×10^{-6}
MDS ²	0.0017	3.22×10^{-6}
UMAP	0.0051	3.37×10^{-6}
LDA	0.0060	3.4×10^{-6}

We observe that the dimensionality reduction algorithms help to improve the quality of the representation due to the fact that the high number of dimensions of the original representation has a lot of misleading information. We remark that better values of MDE and LDE do not necessarily imply that the representation is the best one for every purpose. In fact, looking at the results, we obtained, the MDS algorithm is the best one, but by looking at the configuration obtained by MDS (see Fig. 1d), we observe that if one is interested in understanding the clusters of mathematical fields this configuration gives very poor information with respect to, for example, the configuration obtained by the UMAP algorithm (see Fig. 1f) which obtained the worst score in terms of MDE and LDE.

Figure 1 shows the configurations of the classification codes we considered in \mathbb{R}^3 obtained with different dimensionality reduction algorithms.

In Table 2, we reported three different clustering indicators to evaluate the AMS codes classification representations in \mathbb{R}^3 obtained using different dimensionality reduction algorithms. The indicators we considered are the following (for more details, we refer to Davies and Bouldin, 1979; Rousseeuw, 1987):

- a) the within-cluster sum of squares (WCSS) which measures, for each cluster C , the average distance between a point in C and its centroid;
- b) the mean silhouette score, with values in the interval $[-1, +1]$, which evaluates, on average, the similarity of a point to its cluster compared to other clusters;
- c) the Davies-Bouldin index which evaluates both the separation between different clusters and the variation within the same cluster (lower value the better).

While WCSS highlights the quality of each cluster individually, silhouette and DB index point out the quality of the whole partition obtained using different dimensionality reduction algorithms. For this reason, in Table 2, we have reported the first metric for each cluster and the other two at a global clustering level.

The results reported in the table above show that the Umap algorithm seems to have the better performances on average on the different metrics we have considered. This behavior is also noticeable by looking at the graphic representations of the clustering in Fig. 1.

Once we obtained the representation of MCS codes as element in \mathbb{R}^3 , we made a first attempt in representing researchers using the approach described in Section 2.

We considered the researchers which are members of the Italian association “Istituto Nazionale di Alta Matematica” (INDAM) belonging to the mathematical analysis and geometry subgroups which are named GNAMPA and GNSAGA respectively. For each researchers considered, we collected the MCS codes used in every published articles and the number of times the same code as been used by the researcher. We denote with r the researcher, with

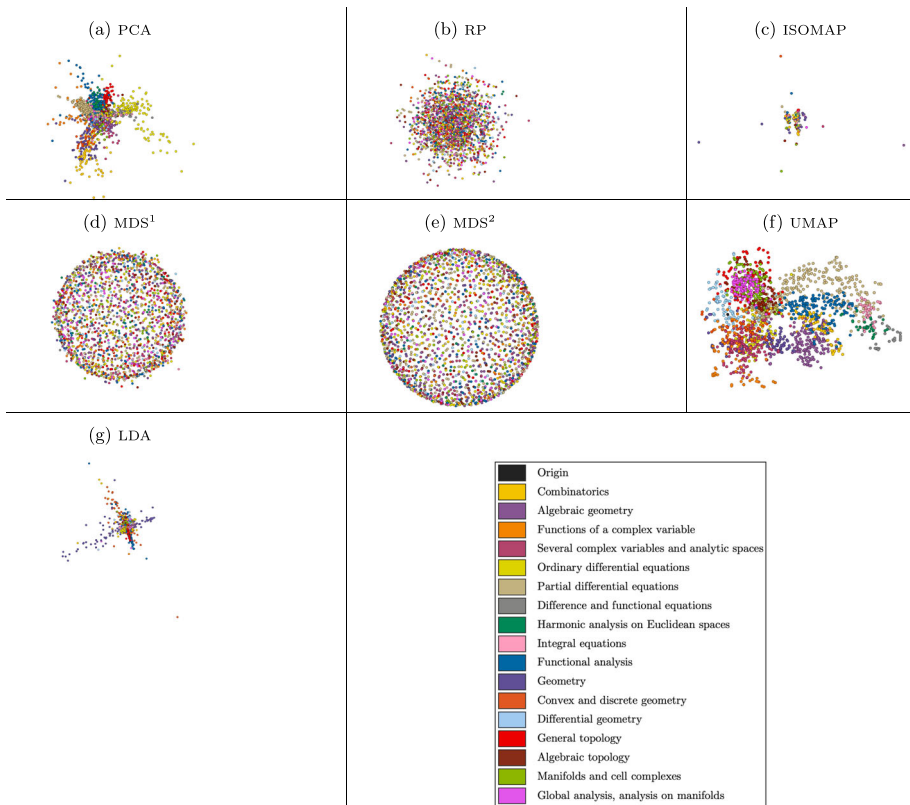


Fig. 1 The figure shows the representation of the AMS classification codes as point in \mathbb{R}^3 obtained using different dimensionality algorithms, colored according to their first hierarchical level. The pictures in this figure are produced with Cinolib (Livesu, 2019) and Py3DViewer (Cherchi et al., 2019) libraries

$\{c_1^r, \dots, c_{n_r}^r\}$ the codes used by r and with t_i^r the number of times the code c_i^r has been used by the researcher. Then, a representation in \mathbb{R}^3 of r is obtained as the weighted centroid of the codes used by the researcher:

$$\mathbf{r} = \sum_{i=1}^{n_r} \mathbf{c}_i^r t_i^r, \quad (5)$$

where \mathbf{c}_i^r is the representation in \mathbb{R}^3 of the code c_i^r .

In Fig. 2, we have inserted the distribution of the researchers according to the representation obtained using different dimensionality reduction algorithms. We are aware that a static figure of points in \mathbb{R}^3 can be hard to analyze. Therefore, we have provided two animations as additional material to this article, the first shows the representation of the AMS codes while the second the representation of researchers. Both animations are obtained by changing the point of view.

In order to evaluate the quality of the representation of the researchers, we computed appropriate balls around each of their representation in \mathbb{R}^3 and we checked how many researchers of the same area were lying inside the ball. In fact, both the subgroup GNAMPA and GNSAGA are divided into clusters according to different research areas in mathematical analysis and

Table 2 Details on the clustering obtained for the AMS classification codes using different dimensionality reduction algorithm

Dim. red.	Category	Cluster	WCSS	Silhouette	DB index
PCA	Geometry	Combinatorics	0.15	-0.07	2.67
		Algebraic geometry	0.02		
		Geometry	0.04		
		Convex and discrete geometry	0.02		
		Differential geometry	0.02		
		General topology	0.07		
		Algebraic topology	0.04		
		Manifolds and cell complexes	0.03		
		Global analysis, analysis on manifolds	0.02		
	Analysis	Functions of complex variables	0.06		
		Several complex variables	0.03		
		Ordinary differential equations	0.10		
		Partial differential equations	0.03		
		Difference and functional equations	0.13		
		Harmonic analysis	0.02		
		Integral equations	0.04		
		Functional analysis	0.04		
		Geometry	Combinatorics		
	Algebraic geometry		51.69		
Geometry	52.43				
Convex and discrete geometry	46.90				
Differential geometry	47.88				
General topology	54.71				
Algebraic topology	50.71				
Manifolds and cell complexes	49.72				
Global analysis, analysis on manifolds	47.77				
RP	Analysis	Functions of complex variables	45.46	-0.18	36.99
		Several complex variables	50.33		
		Ordinary differential equations	44.44		
		Partial differential equations	50.09		
		Difference and functional equations	45.42		
		Harmonic analysis	43.26		
		Integral equations	34.27		
		Functional analysis	48.33		

Table 2 continued

Dim. red.	Category	Cluster	WCSS	Silhouette	DB index
ISOMAP	Geometry	Combinatorics	1.79	-0.66	13.58
		Algebraic geometry	7.42		
		Geometry	10.71		
		Convex and discrete geometry	4.52		
		Differential geometry	0.98		
		General topology	1.48		
		Algebraic topology	2.07		
		Manifolds and cell complexes	3.17		
		Global analysis, analysis on manifolds	1.29		
	Analysis	Functions of complex variables	2.10		
		Several complex variables	1.79		
		Ordinary differential equations	1.49		
		Partial differential equations	0.74		
		Difference and functional equations	2.96		
		Harmonic analysis	1.28		
		Integral equations	1.41		
		Functional analysis	1.47		
MDS ¹	Geometry	Combinatorics	44.50		
		Algebraic geometry	44.34		
		Geometry	41.39		
		Convex and discrete geometry	41.10		
		Differential geometry	44.48		
		General topology	43.33		
		Algebraic topology	45.14		
		Manifolds and cell complexes	42.28		
		Global analysis, analysis on manifolds	41.51		
	Analysis	Functions of complex variables	45.68		
		Several complex variables	45.46		
		Ordinary differential equations	43.77		
		Partial differential equations	45.35		
		Difference and functional equations	39.56		
Harmonic analysis	41.46				
Integral equations	43.82				
Functional analysis	44.75				

Table 2 continued

Dim. red.	Category	Cluster	WCSS	Silhouette	DB index
MDS ²	Geometry	Combinatorics	44.15	- 0.20	34.12
		Algebraic geometry	43.21		
		Geometry	43.95		
		Convex and discrete geometry	44.54		
		Differential geometry	42.91		
		General topology	43.61		
		Algebraic topology	43.66		
		Manifolds and cell complexes	44.07		
		Global analysis, analysis on manifolds	44.28		
	Analysis	Functions of complex variables	44.50		
		Several complex variables	44.24		
		Ordinary differential equations	43.99		
		Partial differential equations	44.34		
		Difference and functional equations	44.10		
		Harmonic analysis	44.21		
		Integral equations	41.27		
		Functional analysis	44.23		
		Geometry	Combinatorics		
Algebraic geometry	0.43				
Geometry	0.43				
Convex and discrete geometry	0.25				
Differential geometry	0.31				
General topology	0.23				
Algebraic topology	0.30				
Manifolds and cell complexes	0.27				
Global analysis, analysis on manifolds	0.33				
UMAP	Analysis	Functions of complex variables	1.52	0.13	2.14
		Several complex variables	0.60		
		Ordinary differential equations	0.48		
		Partial differential equations	0.50		
		Difference and functional equations	0.95		
		Harmonic analysis	0.17		
		Integral equations	0.47		
		Functional analysis	0.58		

Table 2 continued

Dim. red.	Category	Cluster	WCSS	Silhouette	DB index
LDA	Geometry	Combinatorics	4.37		
		Algebraic geometry	0.40		
		Geometry	18.63		
		Convex and discrete geometry	14.37		
		Differential geometry	1.51		
		General topology	3.15		
		Algebraic topology	0.77		
		Manifolds and cell complexes	1.08		
	Analysis	Global analysis, analysis on manifolds	1.12		
		Functions of complex variables	1.13	-0.32	12.78
		Several complex variables	0.60		
		Ordinary differential equations	6.05		
		Partial differential equations	0.32		
		Difference and functional equations	2.61		
		Harmonic analysis	0.29		
		Integral equations	2.67		
		Functional analysis	4.14		

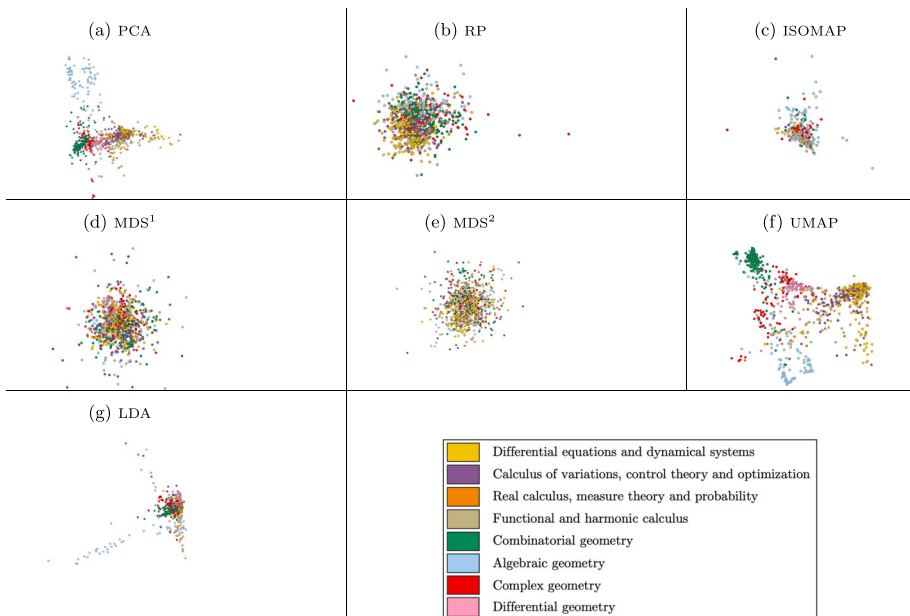


Fig. 2 The figure shows the representation of the researchers of the GNAMPA and GNSAGA clusters of INDAM as point in \mathbb{R}^3 obtained using different dimensionality algorithms colored according to their sub-cluster. The pictures in this figure are produced with Cinolib (Livesu, 2019) and Py3DViewer (Cherchi et al., 2019) libraries

geometry. To validate our approach, we considered “similar” two researchers belonging to the same cluster.

In details, given a researcher r and its representation \mathbf{r} , we considered the m_r closest codes around it, where m_r is equal to 0.75 times the number of codes used by the researcher r . The factor 0.75 has been chosen in order to flatten the distribution of the number of codes used by different researchers. Then, we considered the smallest ball containing the closest codes and we used this neighborhood of \mathbf{r} to check how many similar researchers were close to it. In order to compute an accuracy score for the researchers’ representation, we considered the neighborhood U of \mathbf{r} given by the smallest ball around it containing the m_r closest codes to it and then we considered the number of other researchers contained in U belonging to the same INDAM subgroup as r .

In Table 3, we reported the average accuracy for each INDAM subgroup, computed as described above, together with the three clustering indicators WCSS, silhouette, and DB index, we introduced before in order to evaluate the clustering obtained by the different dimensionality reduction algorithms.

As for Table 2, we have reported the first two metrics for each cluster and the other two at a global clustering level.

Similarly to Table 2, the results reported in the table show that the Umap algorithm seems to have the better performances on average on the different metrics we have considered. This behavior is also noticeable by looking at the graphic representations of the clustering in Fig. 2.

We remark that we could have computed the researchers representation using the non projected codes data Y . However, we have observed that the high-dimensional representation produces poorer results than the representations obtained by first applying a dimensionality reduction algorithm.

4 Discussion

As shown in Fig. 1 the borders between scientific disciplines are often blurred and their boundaries extend along different geometric shapes according to the scientific field. In this way, it can happen that a researcher working in geometry is closer to a colleague working in mathematical analysis than another one working in geometry. We have found several examples of Italian researchers for whom this situation happens. The same conclusions drawn from this empirical application to some important sectors of mathematics can be extended a fortiori to even bigger, rough classifications, such as ASJC or that used for evaluating the highly cited researchers. Indeed in these classifications, any two articles/researchers belonging to the same class (e.g., mathematics) are assumed to have distance 0 each other. However, in this paper, we found a formula for deriving the true distance (or at least a distance closer to the real distance) between articles/researchers and proved that, at least for mathematical disciplines, distances between articles/researchers may rather be variable.

On the other hand, our approach can be useful for avoiding uncorrected comparisons which might led to disastrous consequences even at the individual career level. This is the case of the “National Scientific Qualification”, i.e., the procedure for obtaining the full or associate professorship in Italy. We recall that, according to the Italian legislation, each professor belongs to one among the 190 scientific fields (competition sectors) in which all disciplines are rigidly classified. For each competition sector, ANVUR (the Italian research evaluation agency) calculates some thresholds for the total citations and the h -index to be overcome as necessary condition for being admitted to the calls for the full professorship (calculations are done for the last 15 years) and associated professorship (calculations are done for the last 10

Table 3 Details on the clustering obtained for the researchers in the GNAMPA and GNSAGA subgroups represented as points in \mathbb{R}^3

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
PCA	GNSAGA	Algebraic geometry	0.87	0.17		
		Combinatorial geometry	0.88	0.02		
		Complex geometry	0.55	0.04		
		Differential geometry	0.54	0.01	0.04	3.07
GNAMPA	GNAMPA	Functional and harmonic analysis	0.47	0.04		
		Real analysis, measure theory, and probability	0.16	0.05		
		Variations calculus, control theory, and optimization	0.40	0.01		
		Differential equations and dynamical system	0.60	0.02		
GNSAGA	GNSAGA	Algebraic geometry	0.16	6.13		
		Combinatorial geometry	0.24	5.96		
		Complex geometry	0.12	5.65		
		Differential geometry	0.13	3.79	-0.06	7.00
GNAMPA	GNAMPA	Functional and harmonic analysis	0.17	4.68		
		Real analysis, measure theory, and probability	0.08	6.57		
		Variations calculus, control theory, and optimization	0.25	4.56		
		Differential equations and dynamical system	0.42	3.64		

Table 3 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
ISOMAP	GNSAGA	Algebraic geometry	0.22	0.55		
		Combinatorial geometry	0.21	0.33		
		Complex geometry	0.07	0.08		
		Differential geometry	0.09	0.17	-0.11	10.66
GNAMPA	GNAMPA	Functional and harmonic analysis	0.15	0.21		
		Real analysis, measure theory, and probability	0.01	0.19		
		Variations calculus, control theory, and optimization	0.24	0.08		
		Differential equations and dynamical system	0.39	0.11		
MDS ¹	GNSAGA	Algebraic geometry	0.08	5.64		
		Combinatorial geometry	0.16	4.95		
		Complex geometry	0.07	3.33		
		Differential geometry	0.08	3.04	-0.07	7.67
GNAMPA	GNAMPA	Functional and harmonic analysis	0.11	5.47		
		Real analysis, measure theory, and probability	0.03	7.95		
		Variations calculus, control theory, and optimization	0.21	4.57		
		Differential equations and dynamical system	0.33	3.12		

Table 3 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
MDS ²	GNSAGA	Algebraic geometry	0.14	6.58		
		Combinatorial geometry	0.16	5.18		
		Complex geometry	0.08	3.45		
		Differential geometry	0.09	3.64	-0.06	10.53
GNAMPA	GNSAGA	Functional and harmonic analysis	0.09	4.52		
		Real analysis, measure theory, and probability	0.01	8.11		
		Variations calculus, control theory, and optimization	0.20	3.86		
		Differential equations and dynamical system	0.33	2.99		
UMAP	GNAMPA	Algebraic geometry	0.47	0.66		
		Combinatorial geometry	0.84	0.20		
		Complex geometry	0.52	1.20		
		Differential geometry	0.65	0.47	0.08	2.76
GNAMPA	GNSAGA	Functional and harmonic analysis	0.41	0.55		
		Real analysis, measure theory, and probability	0.09	0.84		
		Variations calculus, control theory, and optimization	0.35	0.62		
		Differential equations and dynamical system	0.59	0.35		

Table 3 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
LDA	GNSAGA	Algebraic geometry	0.75	3.23		
		Combinatorial geometry	0.77	0.11		
		Complex geometry	0.41	0.26		
		Differential geometry	0.61	0.18	0.01	4.64
GNAMPA		Functional and harmonic analysis	0.41	0.63		
		Real analysis, measure theory, and probability	0.09	1.25		
		Variations calculus, control theory, and optimization	0.35	0.25		
		Differential equations and dynamical system	0.59	0.13		

years). Such thresholds are computed as the median value of all scientific papers published by researchers belonging to that competition sector, irrespective of the scientific subfields

Table 4 Citations received by the top 10% and the top 25% articles published in the macro-code **46A** (upper table) and **35A** (bottom table), and calculations of bibliometric indicators for full professorship (15 years) and associated professorship (10 years) according the Italian National Scientific Qualification rules

Year	top 10%	top 25%
2023	0	0
2022	1	0
2021	3	1
2020	4	2
2019	6	2
2018	6	4
2017	8	4
2016	7	3
2015	10	5
2014	10	4
2013	9	4
2012	10	5
2011	14	4
2010	13	5
2009	13	4
Citations last 15 years	114	47
<i>h</i> -index last 15 years	7	4
Citations last 10 years	64	29
<i>h</i> -index last 10 years	6	4
2023	0	0
2022	1	0
2021	4	2
2020	7	4
2019	12	6
2018	16	6
2017	21	9
2016	21	11
2015	27	12
2014	25	12
2013	27	11
2012	27	12
2011	31	13
2010	28	12
2009	29	13
Citations last 15 years	276	123
<i>h</i> -index last 15 years	11	9
Citations last 10 years	134	73
<i>h</i> -index last 10 years	7	6

Table 5 Thresholds stated in the Italian National Scientific Qualification procedure for the competition sector “01/A3 –Mathematical Analysis, Probability, and Mathematical Statistics”

	Incl. into the Committee	Full Prof.	Associate Prof.
Total citations last 10 years	–	–	56
<i>h</i> -index last 10 years	–	–	5
Total citations last 15 years	167	84	–
<i>h</i> -index last 15 years	8	6	–

where they work. Further, higher thresholds are to be overcome by full professors who want to stay in the evaluation committees (for more details see Cappelletti-Montano et al., 2021; Marzolla, 2016).

Now, let us consider a researcher belonging to the competition sector “01/A3-Mathematical Analysis, Probability, and Mathematical Statistics” and working on functional analysis. We may then assume that his/her articles belongs to the codes **46A-XX**. Let us assume that he/she published one article per year, which is customary in pure mathematics. In Table 4 (upper table), we have reported, for each year, the number of citations received among the top 10% and the top 25% articles in the fields belonging to **46A-XX**. It follows that even if the researcher in consideration published each of his/her articles in the top 25% of **46A-XX**, he/she still does not surpass the thresholds (listed in Table 5) and consequently can not obtain the national scientific qualification for becoming full professor or even associate

Table 6 Citations received by the top 10% and the top 25% articles published in the macro-code **54A**, and calculations of bibliometric indicators for full professorship (15 years) and associated professorship (10 years) according the Italian National Scientific Qualification rules

Year	top 10%	top 25%
2023	0	0
2022	0	0
2021	2	0
2020	0	1
2019	3	2
2018	5	2
2017	6	2
2016	5	2
2015	3	2
2014	7	3
2013	7	2
2012	8	2
2011	9	3
2010	8	2
2009	7	2
Citations last 15 years	70	25
<i>h</i> -index last 15 years	6	2
Citations last 10 years	38	16
<i>h</i> -index last 10 years	4	2

Table 7 Thresholds stated in the Italian National Scientific Qualification procedure for the competition sector “01/A2 – Geometry and Algebra”

	Inc. into the Committee	Full Prof.	Associate Prof.
Total citations last 10 years	–	–	20
<i>h</i> -index last 10 years	–	–	3
Total citations last 15 years	93	35	–
<i>h</i> -index last 15 years	6	4	–

professor. Furthermore, senior professors dealing with the topics stated in **46A-XX**, who published in the top 10% of their field, still can not be member of the evaluation committee.

On the other hand, as shown in Table 4 (bottom table) if one publishes on topics of partial differential equations inside **35R-XX**, which still pertains mathematical analysis, he/she will receive much more citations, so surpassing easily the above thresholds.

Similar situations occur also for other sectors. For instance, let us consider a researcher belonging to the competition sector “01/A2 – Geometry and Algebra” and working on general topology. We may then assume that his/her articles belongs to the codes **54A-XX**. Also in this case, let us assume that he/she published one article per year. In Table 6, we have reported, for each year, the number of citations received among the top 10% and the top 25% articles in the fields belonging to **54A-XX**. It follows that even if a researcher published each of his/her articles in the top 25% of **54A-XX**, he/she still does not surpass the thresholds (listed in Table 7) and consequently his/her career as a full professor or even as associate professor will always be precluded. Furthermore, also in this case senior professors, who published in the top 10% in the field **54A-XX**, still can not be member of the evaluation committee.

These considerations point out the fallacy of starting with predefined, poorly constructed, wide-meshed classifications used in current bibliometric analyses. In this paper, we stated the necessity of changing stand-point, and center all bibliometric analyses on the scientific interests of the individual researcher, who should be compared only with those belonging to a neighborhood of him/her, i.e., all the points at distance less than ε from him/her, where ε is a predefined positive number. The point is then to calculate such distance. The procedure illustrated in Section 2 is a first proposal in this direction.

Appendix A: Algorithmic Approach

In this section, we will illustrate in more details the mathematical approach used to obtain a representation of classification entities based on a similarity function defined on the set of entities considered.

Let $G = (V, E)$ be a graph with vertices V and edges E . A graph G is undirected if it does not contain duplicate edges and loops and it is weighted if it is defined a function $d : E \rightarrow \mathbb{R}$.

In order to understand the structure of a given undirected weighted graph $G = (V, E, d)$ a crucial step is to define a representation map $R : V \rightarrow \mathbb{R}^k$ for some $k \in \mathbb{N}$, so that the Euclidean distance between $R(i)$ and $R(j)$ is equal to $d(i, j)$. Application of this approach are given by wireless sensor networks for $k = 1, 2$ or 3 (Singer, 2011), determination of proteins structure from nuclear magnetic resonance experiments for $k = 3$ (Bahr et al.,

2009; Tabaghi et al., 2019), controlling fleets of underwater autonomous vehicles for $k = 3$ (Wüthrich, 1989).

Often, the above problem is called the distance geometry problem, which can be stated as follows:

Distance Geometry Problem *Given an integer $k > 0$ and a simple undirected graph $G = (V, E)$ with an edge weight function $d : E \rightarrow \mathbb{R}^+$, determine whether there exists a realization $R : V \rightarrow \mathbb{R}^k$ such that*

$$\|R(i) - R(j)\| = d(i, j) \quad \forall \{i, j\} \in E. \tag{6}$$

Clearly, not all graphs G admit a realization in \mathbb{R}^k which satisfies (6), in fact it is sufficient that the distances relative to three vertices do not satisfies the triangular inequality. Therefore, it is necessary to consider approximate solution to the DGP.

One of the standard tools used to obtain such approximate solution is mathematical programming (for more details, we refer to Liberti, 2020). Consider the following semi-definite optimization problem:

$$\begin{cases} \min_{X \geq 0} & \sum_{(i,j) \in E} X_{ii} + X_{jj} - 2X_{ij} \\ \text{s.t.} & X_{ii} + X_{jj} - 2X_{ij} \geq d_{ij}^2 \quad \forall (i, j) \in E \end{cases} \tag{7}$$

More precisely, we are looking for the positive semi-definite symmetric matrix X which minimizes the quantity $\sum_{(i,j) \in E} X_{ii} + X_{jj} - 2X_{ij}$ and satisfies the condition $X_{ii} + X_{jj} - 2X_{ij} \geq d_{ij}^2$ for every $(i, j) \in E$.

A solution X^* of Eq. 7 is a symmetric positive semi-definite matrix, which can be factorized as $X^* = YY^T$. If we denote with y_i the i -th row of Y^T , we have that

$$\|y_i - y_j\|^2 = \langle y_i, y_i \rangle + \langle y_j, y_j \rangle - 2\langle y_i, y_j \rangle = X_{ii} + X_{jj} - 2X_{ij} = d_{ij}^2. \tag{8}$$

Therefore, we can consider the rows of Y^T as a representation of the vertices of the graph G . This gives a representation of G in \mathbb{R}^n , where n is the order of the matrix X^* and thus coincide with the number of vertices of G .

It is well-known that solving a semi-definite program like Eq. 7 in a reasonable amount of time can be difficult when the number of vertices and edges of the graph G are large. To avoid this issue, we consider the following linear approximating reformulation, i.e., a mathematical program whose solution are also solution to the program (7) up to some error:

$$\begin{cases} \min_{X, T} & \sum_{(i,j) \in E} X_{ii} + X_{jj} - 2X_{ij} \\ \text{s.t.} & X_{ii} + X_{jj} - 2X_{ij} \geq d_{ij}^2 \quad \forall (i, j) \in E \\ & \sum_{\substack{j \neq i \\ j \in V}} T_{ij} \leq X_{ii} \quad \forall i \leq n \\ & -T \leq X \leq T \end{cases} \tag{9}$$

We observe that, in program (9), the variables are two matrices X and T , but we do not require them to be positive semi-definite, which in practice reduce the complexity of the program and allows to obtain a solution in a much faster time.

A solution \tilde{X} of Eq. 9 is not necessarily a positive semi-definite matrix, hence it cannot be factorized as $\tilde{X} = \tilde{Y}\tilde{Y}^T$. However, we can consider the closest positive semi-definite matrix \hat{X} to \tilde{X} , obtained by zeroing the negative eigenvalues of \tilde{X} and its factorization $\hat{X} = \hat{Y}\hat{Y}^T$ which approximately satisfies the corresponding condition (8).

We observe that both programs Eqs. 7 and 9 give a realization of the graph G in \mathbb{R}^n , where n is the number of vertices of G , which can be very large. If one is interested in a realization of G in a certain dimension $k < n$, one possible solution is to apply a dimensionality reduction algorithm to the points in \mathbb{R}^n . Clearly, such algorithm should preserve the mutual distances between points, so that points in \mathbb{R}^n which are close (resp. far) to each other are mapped to points in \mathbb{R}^k close (resp. far) to each other. There exist several standard dimensionality reduction algorithms attempting to satisfy this condition, which can be subdivided into linear and non-linear algorithms: linear algorithms perform the projection onto a lower-dimensional space using a linear mapping, which has the advantage of being computationally efficient, but can be less accurate. Non-linear algorithms, on the other hand, are computationally more expensive but try to infer non-linear relations between points (e.g., the fact that the points lie on a differentiable manifold) and use them to project them onto a lower-dimensional space with more accuracy.

In the following, we will briefly describe some of the most used dimensionality reduction algorithms of both linear and non-linear case (for more details, we will refer to Hotelling, 1933; Jackson, 2005; Jolliffe, 2002; Xie et al., 2017).

Linear Dimensionality Reduction Algorithms Principal component analysis (Hotelling, 1933) is one of the most famous linear dimensionality reduction algorithms and it performs a linear mapping of the data to a lower-dimensional space so that the variance of the data in the low-dimensional representation is maximized. The algorithm is based on the search of orthogonal directions explaining as much variance of the data as possible. PCA has been used with good results in several fields, e.g., face recognition (Turk & Pentland, 1991) and reconstruction of 3-D objects from their appearance (Murase & Nayar, 1995).

Another linear algorithm to reduce the dimensionality of data which performs well in different situations is given by random projections. Given a matrix $A \in \mathbb{R}^{m \times n}$ representing m vectors in \mathbb{R}^n , a random projection consists in a randomly sampled matrix $T \in \mathbb{R}^{n \times k}$ which satisfies the following property (known as the Johnson-Lindenstrauss Lemma Johnson, 1984): with arbitrarily high probability the following holds

$$(1 - \varepsilon) \|A_i - A_j\|_2^2 \leq \|AT_i - AT_j\|_2^2 \leq (1 + \varepsilon) \|A_i - A_j\|_2^2 \quad (10)$$

where A_i denotes the i -th row of the matrix A , $\varepsilon \in (0, 1)$ is a constant. Therefore, the rows of the matrix $AT \in \mathbb{R}^{m \times k}$ represent m vectors in \mathbb{R}^k whose mutual distances are approximately the same as the distances of the points in \mathbb{R}^n .

Random projections have been used with good results in different dimensionality reduction tasks such as pre-processing of text data (Kurimo, 1999), nearest-neighbor search (Indyk & Motwani, 1998), learning high-dimensional Gaussian mixture models (Dasgupta, 1999).

Non-Linear Dimensionality Reduction Algorithms As stated above, non-linear dimensionality reduction algorithms try to understand relations between the data points considered in order to obtain a more accurate projection. One of the strategies to obtain such information is the so-called kernel trick, which consists in mapping the points into a higher-dimensional space, called the feature space, using a non-linear map, so that it is possible to apply a linear algorithm to the points in the feature space and obtain a more accurate projection. An example of this algorithm is given by the kernel PCA, where after mapping the points into a higher-dimensional space a standard PCA is applied.

More sophisticated algorithms, like MDS, ISOMAP, and UMAP try to understand the structure of Riemannian manifold embedded in the ambient space which best approximates the data points considered. This structure allows to define a distance between data points

considered which is more accurate than the standard Euclidean one, especially if the Riemannian manifold considered is far from being linear (flat) and it is curved (e.g., a paraboloid or a swiss roll surface). We are not going to describe in details how these algorithms work, we just say that MDS, ISOMAP, and UMAP try to approximate the Riemannian distance between points with different methods, so that it is possible to use this distance to perform a dimensionality reduction of the data points in a lower-dimensional space. For more details, we refer to Cox and Cox (2000); McInnes et al. (2018); Tenenbaum et al. (2000).

A.1 Implementation

We have ran the numerical experiments on an Intel i5-4460 3.00 GHz 4-core with 132 GB of RAM under a i686 GNU/Linux operating system. For the implementation, we have used the Python language with the following packages and modules:

- PCA: [sklearn.decomposition.PCA](#);
- RP: [sklearn.random_projection.GaussianRandomProjection](#);
- Isomap: [sklearn.manifold.Isomap](#);
- MDS: [sklearn.manifold.MDS](#);
- Umap: [umap-learn.umap](#)
- LDA: [sklearn.discriminant_analysis.LinearDiscriminantAnalysis](#).

For the DGP, we have implemented the optimization problem using the Python package [gurobipy](#) which solves the problem via the Gurobi solver (Gurobi Optimization, LLC, 2024) for linear programs. The problem is solved using a barrier method without crossover.

The figures in this article are produced with Cinolib (Livesu, 2019) and Py3DViewer (Cherchi et al., 2019) libraries.

Appendix B: Detailed Results for $k = 2$

In this section, we report the results we have obtained by projecting the AMS codes and the researchers in \mathbb{R}^2 . As stated in Section 3, the quality of the representations of AMS codes for $k = 2$ is similar to the one obtained for $k = 3$, but significantly poorer for the clustering of the researchers. Tables 8 and 9 show, respectively, the values for the MDE and LDE

Table 8 The table shows the values of MDE and LDE computed on the original representation in \mathbb{R}^N using the DGP approach and the representations in \mathbb{R}^2 obtained using dimensionality reduction algorithms

Representation	MDE	LDE
Original	0.0079	$5.39 \cdot 10^{-6}$
PCA $_{\mathbb{R}^2}$	0.0063	$3.40 \cdot 10^{-6}$
RP $_{\mathbb{R}^2}$	0.0023	$4.60 \cdot 10^{-6}$
Isomap $_{\mathbb{R}^2}$	0.0059	$8.03 \cdot 10^{-6}$
MDS $^1_{\mathbb{R}^2}$	0.0022	$3.37 \cdot 10^{-6}$
MDS $^2_{\mathbb{R}^2}$	0.0023	$3.32 \cdot 10^{-6}$
Umap $_{\mathbb{R}^2}$	0.0048	$3.39 \cdot 10^{-6}$
LDA $_{\mathbb{R}^2}$	0.0060	$3.41 \cdot 10^{-6}$

Table 9 Details on the clustering obtained for the researchers in the GNAMPA and GNSAGA subgroups represented as points in \mathbb{R}^2

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
PCA	GNSAGA	Algebraic geometry	0.58	0.07		
		Combinatorial geometry	0.86	0.00		
		Complex geometry	0.54	0.02		
		Differential geometry	0.55	0.00	0.02	4.24
GNAMPA	GNAMPA	Functional and harmonic analysis	0.27	0.03		
		Real analysis, measure theory, and probability	0.14	0.03		
		Variations calculus, control theory, and optimization	0.37	0.01		
		Differential equations and dynamical system	0.60	0.01		
RP	GNSAGA	Algebraic geometry	0.14	6.28		
		Combinatorial geometry	0.23	4.80		
		Complex geometry	0.12	5.41		
		Differential geometry	0.16	3.59	-0.08	8.63
GNAMPA	GNAMPA	Functional and harmonic analysis	0.18	5.15		
		Real analysis, measure theory, and probability	0.11	6.39		
		Variations calculus, control theory, and optimization	0.24	4.77		
		Differential equations and dynamical system	0.40	3.79		

Table 9 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
ISOMAP	GNSAGA	Algebraic geometry	0.19	0.41		
		Combinatorial geometry	0.21	0.32		
		Complex geometry	0.07	0.07		
		Differential geometry	0.09	0.16	-0.12	10.98
GNAMPA	GNAMPA	Functional and harmonic analysis	0.15	0.20		
		Real analysis, measure, theory, and probability	0.01	0.18		
		Variations calculus, control theory, and optimization	0.24	0.08		
		Differential equations and dynamical system	0.39	0.10		
MDS ¹	GNSAGA	Algebraic geometry	0.13	5.50		
		Combinatorial geometry	0.19	5.03		
		Complex geometry	0.10	3.60		
		Differential geometry	0.11	4.09	-0.07	9.38
GNAMPA	GNAMPA	Functional and harmonic analysis	0.11	4.51		
		Real analysis, measure theory, and probability	0.02	5.60		
		Variations calculus, control theory, and optimization	0.22	3.55		
		Differential equations and dynamical system	0.37	2.36		

Table 9 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
MDS ²	GNSAGA	Algebraic geometry	0.16	5.25		
		Combinatorial geometry	0.19	4.32		
		Complex geometry	0.12	3.20		
		Differential geometry	0.09	4.18	-0.09	31.40
GNAMPA	GNAMPA	Functional and harmonic analysis	0.10	4.63		
		Real analysis, measure theory, and probability	0.12	3.39		
		Variations calculus, control theory, and optimization	0.22	3.68		
		Differential equations and dynamical system	0.38	2.70		
UMAP	GNSAGA	Algebraic geometry	0.56	0.52		
		Combinatorial geometry	0.80	0.25		
		Complex geometry	0.59	0.79		
		Differential geometry	0.47	0.35	0.01	5.54
GNAMPA	GNAMPA	Functional and harmonic analysis	0.37	0.62		
		Real analysis, measure theory, and probability	0.21	0.94		
		Variations calculus, control theory, and optimization	0.35	0.54		
		Differential equations and dynamical system	0.57	0.43		

Table 9 continued

Dim. red.	INDAM groups	Cluster	Accuracy	WCSS	Silhouette	DB index
LDA	GNSAGA	Algebraic geometry	0.71	3.01		
		Combinatorial geometry	0.70	0.08		
		Complex geometry	0.37	0.16		
		Differential geometry	0.38	0.15	-0.01	4.14
GNAMPA	GNAMPA	Functional and harmonic analysis	0.39	0.34		
		Real analysis, measure theory, and probability	0.17	0.96		
		Variations calculus, control theory, and optimization	0.32	0.15		
		Differential equations and dynamical system	0.57	0.02		

computed on the representations in \mathbb{R}^N and \mathbb{R}^2 , and the details on the clustering obtained for the researchers in the GNAMPA and GNSAGA subgroups in \mathbb{R}^2 .

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00357-024-09490-2>.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement. B.M. is funded by PON “Ricerca e Innovazione” 2014-2020 (PON R&I, CUP F25F21002210003). G.C. is partially funded by PRIN 2020 project financed by MUR (CUP F73C22000430001). M.M. is partially funded by Fondazione di Sardegna. S.M. is partially funded by PNRR e.INS Ecosystem of Innovation for Next Generation Sardinia (CUP F53C22000430001, codice MUR ECS00000038). B.C.M., B.M., and S.M. are members of INDAM - Istituto Nazionale di Alta Matematica.

Data Availability Statements The datasets generated by the survey research during and/or analyzed during the current study are available in the Zentralblatt MATH repository, <https://zbmath.org>.

Declarations

Ethical Approval This research does not contain any studies with human participations or animals performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abramo, G., & D’Angelo, C. A. (2015). An assessment of the first “scientific habilitation” for university appointments in Italy. *Economia Politica*, 32(3), 329–357.
- Baccini, A., De Nicolao, G., Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS ONE*, 14(9).
- Baccini, F., Barabesi, L., Baccini, A., Khelifaoui, M., & Gingras, Y. (2022). Similarity network fusion for scholarly journals. *Journal of Informetrics*, 16(1), 101226.
- Bahr, A., Leonard, J. J., & Fallon, M. F. (2009). Cooperative localization for autonomous underwater vehicles. *International Journal of Robotics Research*, 28(6), 714–728.
- Bornmann, L., & Haunschild, R. (2018). Plots for visualizing paper impact and journal impact of single researchers in a single graph. *Scientometrics*, 115(1), 385–394.
- Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 98(1), 487–509.
- Cagan, R. (2013). The San Francisco declaration on research assessment. *DMM Disease Models and Mechanisms*, 6(4), 869–870.
- Cappelletti-Montano, B., Columbu, S., Montaldo, S., Musio, M. (2021). New perspectives in bibliometric indicators: Moving from citations to citing authors. *Journal of Informetrics*, 15(3).
- Cherchi, G., Pitzalis, L., Frongia, G.L., Scateni, R. (2019). The py3dviewer project: A python library for fast prototyping in geometry processing. *Italian chapter conference 2019 - smart tools and apps in computer graphics, stag 2019* (pp. 121–128).
- Cox, T., & Cox, M. (2000). *Multidimensional scaling* (2nd ed.). CRC Press.

- Dasgupta, S. (1999). Learning mixtures of Gaussians. *Annual symposium on foundations of computer science - proceedings* (pp. 634–644).
- Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, 1*(2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>
- Demetrescu, C., Finocchi, I., Ribichini, A., & Schaerf, M. (2020). On bibliometrics in academic promotions: A case study in computer science and engineering in Italy. *Scientometrics, 124*(3), 2207–2228.
- Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (2019). *Springer handbook of science and technology indicators*. Springer.
- Gurobi optimization, LLC (2024). Gurobi optimizer reference manual. Retrieved from <https://www.gurobi.com>
- Haunschild, R., Bornmann, L., & Adams, J. (2019). R package for producing beamplots as a preferred alternative to the h index when assessing single researchers (based on downloads from web of science). *Scientometrics, 120*(2), 925–927.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology, 24*(6), 417.
- Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Conference Proceedings of the Annual ACM Symposium on Theory of Computing* (p. 604–613).
- Ioannidis, J.P. (2022). September 2022 data-update for “updated science-wide author databases of standardized citation indicators”. *Mendeley Data, 4*.
- Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. *PLoS biology, 17*(8), e3000384.
- Jackson, J. E. (2005). *A user's guide to principal components*. John Wiley & Sons.
- Johnson, W. B. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemp Math, 26*, 189–206.
- Jolliffe, I. T. (2002). *Principal component analysis for special types of data*. Springer.
- Kurimo, M. (1999). Indexing audio documents by using latent semantic analysis and som. Kohonen maps (pp. 363–374). Elsevier.
- Liberti, L. (2020). Distance geometry and data science. *TOP, 28*(2), 271–339.
- Livesu, M. (2019). Cinolib: A generic programming header only c++ library for processing polygonal and polyhedral meshes (Vol. 11820 LNCS).
- Marzolla, M. (2016). Assessing evaluation procedures for individual researchers: The case of the Italian national scientific qualification. *Journal of Informetrics, 10*(2), 408–438.
- McInnes, L., Healy, J., Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint [arXiv:1802.03426](https://arxiv.org/abs/1802.03426)
- Murase, H., & Nayar, S. K. (1995). Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision, 14*(1), 5–24.
- Murtagh, F., Orlov, M., & Mirkin, B. (2018). Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research. *Journal of Classification, 35*(1), 5–28.
- Petrera, M., Trautwein, D., Beckenbach, I., Ehsani, D., Müller, F., Teschke, O., . . . Schubotz, M. (2021). zbMATH Open: API solutions and research challenges. arXiv preprint [arXiv:2106.04664](https://arxiv.org/abs/2106.04664)
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.
- Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis, 30*(1), 20–36.
- Tabaghi, P., Dokmanic, I., Vetterli, M. (2019). On the move: Localization with kinetic Euclidean distance matrices. ICASSP, IEEE international conference on acoustics, speech and signal processing - proceedings (vol. 2019-May, pp. 4893–4897).
- Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.
- Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience, 3*(1), 71–86.
- Waltman, L., & van Eck, N. J. (2013). Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison [Article]. *Scientometrics, 96*(3), 699–716.
- Wüthrich, K. (1989). Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science, 243*(4887), 45–50.
- Xie, H., Li, J., Xue, H. (2017). A survey of dimensionality reduction techniques based on random projection. arXiv preprint [arXiv:1706.04371](https://arxiv.org/abs/1706.04371)