Analysis of gender equality in US college sports' funding via spatial clustering

Computational Statistics (2026) 41:75
https://doi.org/10.1007/s00180-026-01758-y

ORIGINAL PAPER

Analysis of gender equality in US college sports' funding via spatial clustering

Pierpaolo D'Urso[1] · **Livia De Giovanni**[2,3] · **Lorenzo Federico**[2]

Received: 21 November 2025 / Accepted: 14 April 2026

Abstract
In this paper, we analyze the geographic differences in the level of equality of resource distribution between women's and men's college sports in the United States. We compute the distribution of various equality measures across universities in each of the 52 top-level administrative divisions (the 50 states plus the District of Columbia and Puerto Rico) and use a spatially corrected fuzzy clustering algorithm to explore and partition the dataset. We see the emergence of clear spatial patterns that correspond to the usual divide in American politics, which are further strengthened and smoothed using a modularity-based spatial correction term.

1 Introduction and literature review

Gender inequality in sport, in terms of funding for sports events and infrastructure, and athletes' salaries, has been a long-standing issue that has recently garnered attention, both within communities of sport fans and insiders and through the general public. Even in sports like tennis, where the top female athletes have been professional for decades, and the tournaments of the WTA Tour receive ample media coverage, studies (Flake et al. 2013; Mercer and Edwards 2020) reveal that the top female play-

✉ Lorenzo Federico
lfederico@luiss.it

Pierpaolo D'Urso
pierpaolo.durso@uniroma1.it

Livia De Giovanni
ldegiovanni@luiss.it

[1] Department of Social Science and Economics, Sapienza University, Piazzale Aldo Moro 5, 00185 Rome, Lazio, Italy

[2] Department of AI, Data and Decision Sciences, Luiss University, Viale Romania 32, 00197 Rome, Lazio, Italy

[3] Data Lab, Luiss University, Viale Pola 12, 00198 Rome, Lazio, Italy

Published online: 28 April 2026

ers earn around 20% less than their male counterparts. The situation is much bleaker in sports that are considered typically masculine Koivula (2001); Yiapanas (2025). In association football, in most European countries, male players are recognized as fully professional athletes, not only in the top flight of each country, but also in the lower levels of the "football pyramid". As Ribeiro and Lima (2019) showed, even in the Portuguese league, which is not one of the richest (Liu et al. (2016) classified it as a "farm league", from which teams usually sell their best players to richer leagues), players at the top level easily earn hundreds of thousands if not millions of euros per year, while players down in the third or fourth level are still able to earn a living off their sport activity. In comparison, the first fully professional female league in all of Europe was founded only in 2018 in England.

The same disparity extends beyond the top levels of professional sport to the funding of local sports associations, as analyzed by Devine (2018) in Great Britain, and by Yenilmez (2021) in Turkey, which in turn affects the accessibility of sports facilities for girls and women.

Between amateur and professional sports, especially in North America, lies the college sports system. While college athletes are non-professional and can only be compensated with scholarships that cover their college taxes and expenses, in many sports, college programs are the main route towards a professional career for North American athletes. Consequently, availability and adequate funding for college sport facilities are crucial to provide young athletes with opportunities for a professional career. As shown by Norman et al. (2021) and O'Connor (2021), funding for women's college sport programs is consistently inferior compared to men's.

In this paper, we investigate how the magnitude of gender inequality in funding for college sports changes across the United States using a spatial K-medoids clustering method developed in Cangemi et al. (2025).

We measure the distribution of several indices of gender inequality in sport funding for the universities of each state and use a suitable distance between distributions to build a fuzzy entropic clustering algorithm (Miyamoto and Mukaidono 1997). We apply it both in its baseline version and with a spatial regularization term based on fuzzy modularity (Nepusz et al. 2008).

Fuzzy clustering with spatial regularization has been extensively applied across many fields since the seminal paper by Pham (2001) in computer vision. Other algorithms have been developed to handle different types of data (Krishnapuram et al. 1999; Coppi et al. 2010; D'Urso et al. 2023, 2024).

Fuzzy clustering methods are popular in sports data science, being used to understand team sports tactics Narizuka and Yamazaki (2019), individual player performances (D'Urso et al. 2023; Carpita et al. 2023), and to process sports video data (Lu and Tan 2003). A specific application of fuzzy clustering with spatial correction is D'Urso et al. (2025), where the attributes are performance metrics of tennis players and tournaments, while the network structure does not represent an physical spatial structure, but rather the participation of players in tournaments.

The paper is structured as follows. In Sect. 2, the data and the model used are presented. Section 3 reports the results of the application of the models to spatial clustering of US administrative divisions, based on their level of gender equality. Section 4 concludes the paper and provides directions for future work.

2 The model

In this section, we present the clustering model we are going to use. In Sect. 2.1, we present the dataset we use to analyse inequality in college sport funding and define an appropriate distance between units, based on the attributes we selected from the dataset. In Sect. 2.2, we recall the notion of fuzzy modularity and define the clustering model, detailing both the objective function to optimize and the algorithmic procedure.

2.1 The data

We use data about gender equality in the funding of US college sports taken from the Equity in Athletics Data Analysis website Office of Postsecondary Education (2025). The database from the year 2023 contains information about the funding for men's and women's sports for 2040 in different colleges.

We consider 5 different attributes that describe different expenses the college sustained to support their teams:

- Average Full-time Head Coach Salary,
- Average Full-time Assistant Coach Salary,
- Recruitment Expenses,
- Athletic student aid,
- Operating cost per team.

For each attribute, the database registers the value for women's and men's sports for each college where it applies. We choose these attributes because we want to focus on the economic side of gender inequality between male and female student college athletes, both in terms of remuneration of the staff (athletes themselves are not paid a salary in college sport) and resources given to the teams to function properly.

The goal of the present paper is to identify geographic areas with a higher or lower disparity in the funding male and female college teams receive. In our final clustering, the units are the US top-level administrative divisions with more than 1 university considered in the census, that is, the 50 states plus the District of Columbia and Puerto Rico, and they are classified based on the distribution of the inequality indicators across the colleges in each division. For each attribute, we compute the ratio of its values for women's and men's sports, for every college where both data are available.

We thus build an attribute matrix $\mathrm{X} = \{x_{i,j} : i \leq 2040. j \leq 5\}$, where

$$x_{ij} = \frac{\text{attribute } j \text{ in college } i \text{ for women's sport}}{\text{attribute } j\text{in college } i\text{for men's sport}}, \tag{1}$$

if such information is available, and is left empty otherwise.

We thus obtain the 5 attributes for each college:

- Avg. Full-time Head Coach Salary Ratio,

- Avg. Full-time Assistant Coach Salary Ratio,
- Recruitment Expenses Ratio,
- Athletic student aid Ratio,
- Operating cost per team Ratio.

For each administrative division, we thus build the empirical distribution of each attribute over the colleges in that division, that is, for every division *n* and every attribute *j* we have

$$S_{n,j} = \{x_{ij} \ \forall \text{ college } i \text{ in division } n\}. \tag{2}$$

We thus build for all the administrative divisions and attributes the empirical cumulative distribution functions

$$\hat{F}_{n,j}(t) = \frac{1}{|S_{n,j}|} \sum_{x \in S_{n,j}} 1_{x \leq t}, \tag{3}$$

and use the Chebichev (or $\mathbb{L}^\infty$) distance between the two empirical cumulative distribution functions of an attribute in two different administrative divisions as a dissimilarity measure:

$$\|\hat{F}_{n,j} - \hat{F}_{m,j}\|_\infty = \sup_{t \in \mathbb{R}} |\hat{F}_{n,j}(t) - \hat{F}_{m,j}(t)|, \tag{4}$$

this distance is mostly used as the statistic of a 2-sample Kolmogorov-Smirnov test (Hodges 1958). It has the desirable property of allowing us to compare samples of different sizes without requiring any assumption about the shape of the distributions. It is worth pointing out that for every two samples $A = \{a_1, \ldots a_k\}$ and $B = \{b_1, \ldots, b_l\}$, $\|\hat{F}_A - \hat{F}_B\|_\infty \in [0.1]$, and $\|\hat{F}_A - \hat{F}_B\|_\infty = 0$ if and only if all relative frequencies are identical, and $\|\hat{F}_A - \hat{F}_B\|_\infty = 1$ if and only if the two ranges do not overlap, that is, $a_i > b_j \ \forall i, j$ or $b_j > a_i \ \forall i, j$.

We thus identify the position of every division in the attribute space as the vector of its 5 empirical cumulative distribution functions for the five attributes

$\mathrm{z}_n = (\hat{F}_{n,1}, \ldots, \hat{F}_{n,5})$.

From this, we define the attribute-based distance between two administrative divisions *n* and *m* as the $\mathbb{L}^2$ distance over 5-fold product of the $\mathbb{L}^\infty$ space of real-valued functions, that is

$$d_{2,\infty}(\mathrm{z}_n, \mathrm{z}_m) = \Big(\sum_{j=1}^{5} (\|\hat{F}_{n,j} - \hat{F}_{m,j}\|_\infty)^2 \Big)^{1/2}. \tag{5}$$

The resulting 52×52 distance matrix is shown in Fig. 1 and is then used as input in the algorithm.

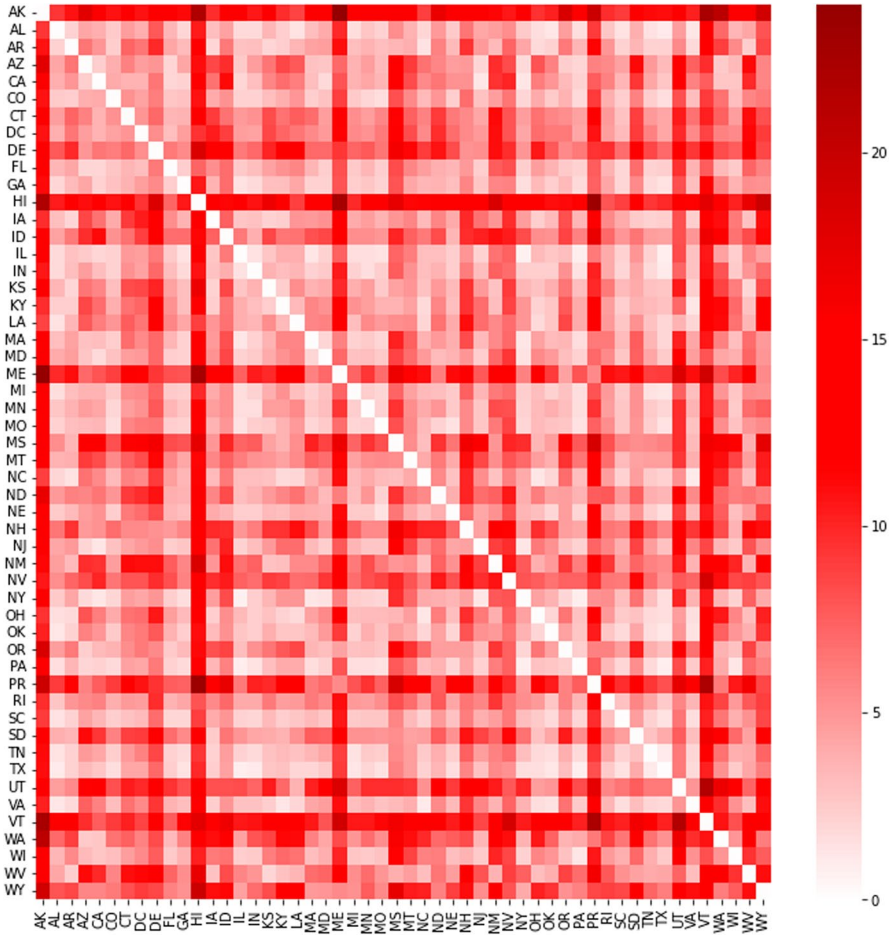

Fig. 1 Matrix of squared distances $d_{2,\infty}$ between the attributes of different administrative divisions

The information about geographic contiguity is represented as a binary adjacency matrix $\mathrm{A} = \{a_{n,m}, n, m \leq 52\}$ where $a_{n,m} = 1$ if administrative divisions n and m share a land border and 0 otherwise. Note that this way, Hawaii, Puerto Rico, and Alaska have no contiguous units. The first two indeed are composed only of islands, and the latter is separated from the rest of the continental US, sharing a land border only with the Canadian administrative divisions of British Columbia and Yukon, which are not considered in this study. The resulting adjacency matrix is shown in Fig. 2

2.2 The clustering algorithm

We use the Fuzzy C-Medoids with modularity spatial correction (FCMd-MSC), introduced in Cangemi et al. (2025). This algorithm requires the units to be defined in a metric space, in which the distance between two units is based on the respective

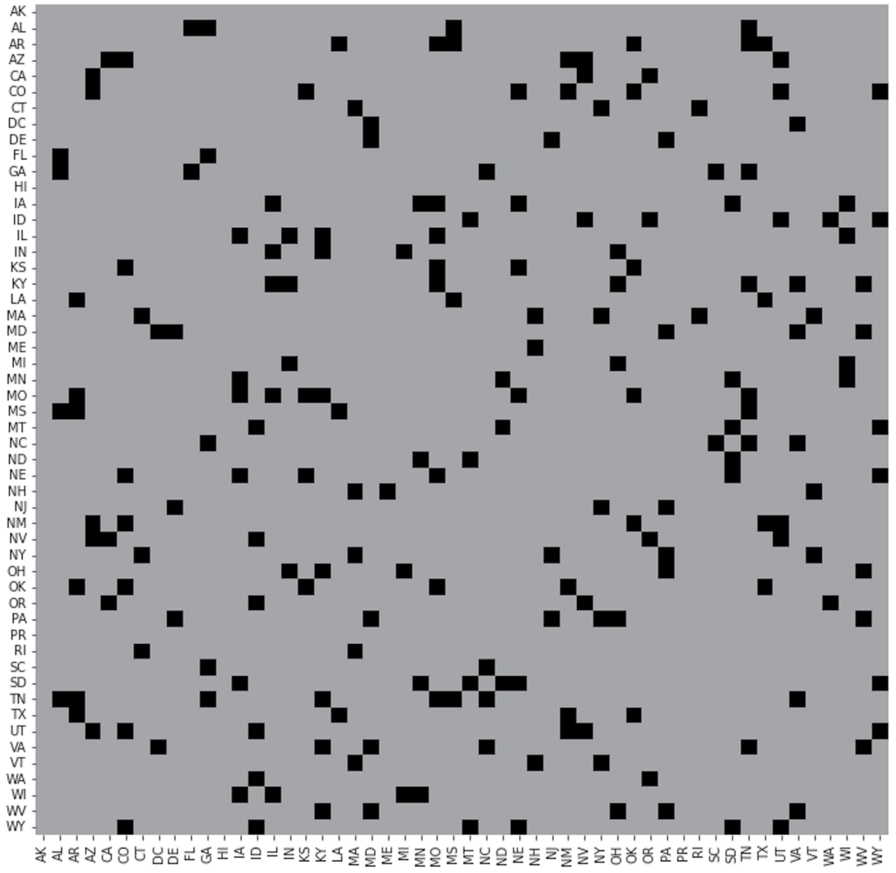

Fig. 2 Adjacency matrix between different administrative divisions. Black dots indicate that the row and column divisions share a land border

attributes. In practice, we optimize a convex combination of the fuzzy modularity of the network defined by the adjacency matrix A and of a fuzzy entropic clustering over the units using the aforementioned $d_{2,\infty}$, defined in (5). The algorithm, as expected of a Partition around Medoids algorithm, outputs a fuzzy partition matrix $\mathrm{U} = \{u_{n,c} : n \leq 52, c \leq C\}$ of the set of units, where each $u_{n,c} \in [0.1]$ represents the level of membership of unit *n* to cluster *c*, together with a set of prototype units $(\hat{z}_1, \ldots, \hat{z}_c, \ldots, \hat{z}_C)$, called medoids, one for each cluster, as introduced in Krishnapuram et al. (1999). The use of medoids instead of means in this case is necessary, because in the space of cumulative distribution functions with $\mathbb{L}^\infty$ distance it is not possible to define a centroid of a cluster in the same way as we could in an Euclidean space, that is, as the point that minimizes the sum of the squared distances of all the members of the cluster, weighted by their membership. Indeed, such a centroid distribution function might be non-unique, making it doubtful whether the algorithm can even converge. The medoids approach instead guarantees the possibility to find a prototype unit in the cluster, helping both with the convergence of the algorithm and

the interpretation of the output. We define the fuzzy modularity as in Nepusz et al. (2008). From the adjacency matrix A, we compute for every unit $n = 1, \ldots, 52$ its *degree* that is, the number of connections it has with other units, as

$$w_n = \sum_{m=1}^{N} a_{n,m}, \tag{6}$$

and the total degree of the network, which represents twice the total number of connections in the entire network, as the sum of all degrees

$$L = \sum_{n=1}^{N} w_n. \tag{7}$$

We define the modularity matrix $\mathrm{B} := \{b_{n,m} : n, m \leq N\}$ as

$$b_{n,m} = a_{n,m} - \frac{w_n w_n}{L}. \tag{8}$$

This represents the difference between the values in the actual adjacency matrix and their expected values under a suitable null model, such as the configuration model (Molloy and Reed 1998) or the inhomogeneous rank-1 random graph (Norros and Reittu 2006).

We can thus define, up to a constant, the fuzzy modularity of the partition U with respect to the adjacency matrix A as

$$Q(\mathrm{U}, \mathrm{A}) = \sum_{n=1}^{N} \sum_{m=1}^{N} b_{n,m} \sum_{c=1}^{C} u_{n,c} u_{m,c} (1 - \delta_{n,m}). \tag{9}$$

Note that $\sum_{c=1}^{C} u_{n,c} u_{m,c}$ is high when units *n*, *m* have high membership to the same cluster. This means that if $b_{n,m}$ is positive (usually, when *n* and *m* are connected) there is a bonus for classifying them in the same cluster, while if $b_{n,m}$ is negative (when *n* and *m* are not connected) there is a penalty for classifying them in the same cluster. Moreover, for units with low degree (that is, with very few adjacent units) the presence of a connection is weighted more, while for units with high degree the absence of a connection is more important.

We can now define explicitly the optimization problem, which is further regulated by the parameters γ, which controls the relative importance of the adjacency matrix and the attributes, *C*, which corresponds to the number of clusters and *p*, which tunes the fuzziness of the partition. We thus attempt to solve the following minimization problem,

$$\min_{\mathrm{U},\hat{\mathrm{z}}_c} J_{p,C,\gamma}(\mathrm{U},\hat{\mathrm{z}}_c) := (1-\gamma)\sum_{n=1}^{N}\sum_{c=1}^{C} u_{n,c} d^2_{2,\infty}(\mathrm{z}_n,\hat{\mathrm{z}}_c) + p\sum_{n=1}^{N}\sum_{c=1}^{C} u_{n,c}\log(u_{n,c}) - \frac{\gamma}{2}\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{m=1}^{N} u_{n,c} b_{n,m} u_{m,c}(1-\delta_{n,m}). \tag{10}$$

under the following constraints:

$$u_{n,c} \geq 0. \quad \sum_{c=1}^{C} u_{n,c} = 1. \tag{11}$$

In practice, we subtract the fuzzy modularity from the objective function of an entropic fuzzy clustering. This makes sense because, in fuzzy clustering, the objective function is to be minimized, whereas in modularity-based algorithms, higher modularity is favored. The algorithm thus seeks an optimal balance between good performance with respect to the attribute and network structures. The parameter $\gamma \in [0.1]$ tunes the relative importance of the attributes and the spatial structure, with $\gamma = 0$ corresponding to a Fuzzy C-medoids clustering without spatial correction and $\gamma = 1$ corresponding to fuzzy modularity maximization with no attributes term. Higher values of p correspond to a greater fuzziness of the partition.

Given that an exact solution of the optimization problem is not feasible, we take an iterative approach, minimizing the objective function with respect to the membership matrix for fixed values of the medoids and then with respect to the medoids for fixed values of the membership matrices, until the algorithm converges at least to a local minimum. We find that the minima with respect to the memberships $u_{n,c}$ of $J_{p,C,\gamma}(\mathrm{U},\hat{\mathrm{z}}_c)$ using the Lagrangian method, and we obtain the update rule for U at each step as

$$u_{n,c} = \frac{\exp\left\{-\frac{1}{p}\left((1-\gamma)d^2_{2,\infty}(\mathrm{z}_n,\hat{\mathrm{z}}_c) - \gamma\sum_{m=1}^{N} b_{n,m}u_{m,c}(1-\delta_{n,m})\right)\right\}}{\sum_{c'=1}^{C}\exp\left\{-\frac{1}{p}\left((1-\gamma)d^2_{2,\infty}(\mathrm{z}_n,\hat{\mathrm{z}}_{c'}) - \gamma\sum_{m=1}^{N} b_{n,m}u_{m,c'}(1-\delta_{n,m})\right)\right\}}. \tag{12}$$

The optimization over all the possible choices of $(\hat{\mathrm{z}}_1,\ldots,\hat{\mathrm{z}}_c,\ldots,\hat{\mathrm{z}}_C)$ for fixed values of U is done by searching for the medoids of each cluster over all the units with the corresponding maximal membership. This restriction carries two important advantages: (1) it reduces computational complexity; (2) it makes sure that a unit from cluster c cannot be chosen as medoid of cluster c'. This could happen, as the network structure does not influence the medoids' selection, but is used in the update of the memberships.

The procedure is described in pseudocode in Algorithm 1, for further information see (Cangemi et al. (2025), Sect. 2.2).

Fix C, $max.iter$, $conv = 1 \times 10^{-9}$ and initialize randomly the membership degree matrix $\mathbf{U}$;
Set $iter = 0$;
Set $medoids := (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_c)$, arbitrarily;
repeat
  Set $\mathbf{U_{old}} = \mathbf{U}$;
  Update $medoids$ as follows:
  for $c = 1$ to C **do**
    Define $members = \{n \leqslant N : c = \arg\max_{1 \leqslant k \leqslant C} u_{n,k}\}$
    if $members$ is not empty **then**
      $q = \arg\min_{q \in members} \sum_{n=1}^{N} u_{n,c} d^2(\mathbf{x}_n, \mathbf{x}_q)$
      Set $\Rightarrow \tilde{\mathbf{x}}_c = \mathbf{x}_q$
    end if
  end for
  Update $\mathbf{U}$ using (12);
  $iter \leftarrow iter + 1$;
until $\|\mathbf{U_{old}} - \mathbf{U}\|_1 < conv$ or $iter = max.iter$
return $\mathbf{U}$, $(\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_c)$.

Algorithm 1 Fuzzy C-medoids with modularity spatial correction (FCMd-MSC) (Cangemi et al. 2025)

3 Results

We run the optimization algorithm at 3 different spatial regularization levels with $\gamma = 0.0.3, 0.5$. In practice, we want to consider three separate cases where the spatial structure is completely ignored, considered a secondary feature, and considered on equal ground to the attributes, respectively. Even if the algorithm as defined in Cangemi et al. (2025) includes the possibility of choosing the optimal γ using a validity measure, we prefer not to apply it, as it would result in an optimal value $\gamma > 0.5$ which for the analysis we are interested is of no use, as, in the problem considered, the spatial structure is the well-known geography of the United States, which is not an interesting topic of analysis on its own. Thus, it is not insightful to consider it as the main feature around which the algorithm optimizes the partition, and treats the attributes as a secondary correction term.

We optimize the parameters p and C for each selected value of γ based on the values that maximize the *fuzzy silhouette* (Campello and Hruschka 2006). The fuzzy silhouette is the most common validity index present in the literature for the optimization of C and given that it does not depend explicitly on the objective function it is robust when applied to entropic clustering and to the joint selection of C and p. It is interesting to note (see Table 1) that for all the 3 values of γ considered, the optimal choice is $C = 2$ and $p = 0.6$, with the medoids of the two clusters being Texas and New York, whose values of the attributes are described in Table 2. This shows that the two structures we are jointly analyzing (the spatial relations and the attributes) broadly agree, and thus the presence of a spatial regularization term and its weighting causes the reassignment of a few units, but does not change the large-scale nature of the clusters. Moreover, we note that the fuzzy silhouette values decrease very rapidly

Table 1 Value of the fuzzy silhouette for different values of p and C for $\gamma = 0, 0.3, 0.5$

$\gamma = 0$	p	0.2	0.4	**0.6**	0.8	1.0
	2	0.282	0.301	**0.310**	0.256	0.256
C	3	0.193	0.170	0.180	0.186	0.190
	4	0.134	0.096	0.096	0.094	0.091
$\gamma = 0.3$	p	0.2	0.4	**0.6**	0.8	1.0
	2	0.289	0.307	**0.321**	0.297	0.300
C	3	0.083	0.094	0.107	0.115	0.126
	4	0.088	0.086	0.075	0.063	0.063
$\gamma = 0.5$	p	0.2	0.4	**0.6**	0.8	1.0
	2	0.245	0.257	**0.277**	0.092	0.204
C	3	0.044	0.047	0.054	0.087	0.104
	4	0.038	−0.004	−0.012	0.063	0.028

Selected values are in bold

Table 2 Descriptive statistics of the samples in the two medoid states

NY	Count	25%	50%	75%	Mean	St. dev
Avg. full-time head coach salary ratio	135	0.833	0.976	1.013	0.923	0.204
Avg. full-time assistant coach salary ratio	139	0.792	0.946	1.002	0.903	0.231
Operating cost per team ratio	140	0.606	0.761	0.926	0.796	0.392
Recruitment expenses ratio	112	0.545	0.922	1.169	1.124	1.846
Athletic student aid ratio	48	0.759	0.963	1.156	0.934	0.311
TX	Count	25%	50%	75%	Mean	St. dev
Avg. full-time head coach salary ratio	101	0.701	0.910	1.018	0.878	0.364
Avg. full-time assistant coach salary ratio	109	0.705	0.878	1.000	0.809	0.281
Operating cost per team Ratio	109	0.573	0.763	0.945	0.791	0.316
Recruitment expenses ratio	97	0.392	0.600	1.000	0.815	0.940
Athletic student aid ratio	87	0.689	0.876	1.077	0.913	0.328

Count represents the number of universities considered in the study, $25\%, 50\%$ and 75% are the quartiles

for $C > 2$, in particular in the presence of spatial regularization. This shows how the division into only two clusters is strongly favored by the attribute distribution.

Indeed, already for $\gamma = 0$, that is, without including explicitly the spatial structure, clear geographic patterns are visible, as shown in the maps in Fig. 3, with cluster 2 (medoid New York) consisting mostly of the Pacific Coast and New England, with Wyoming being the only geographically isolated state among its members in mainland US. Cluster 1 (medoid Texas) is instead mostly composed of states in the Great Plains, Rocky Mountains, and the south, with only Rhode Island being spatially disconnected from the rest of the cluster, other than Hawaii and Alaska. From Table 2, which summarizes the distribution of the different indices in the medoid states, we see that the typical values of the equality indicators are higher in cluster 2. Only in the case of *Operating cost per team Ratio* we see that the distributions are similar. It has to be noted that operating costs are not something that a university can decide freely, but also depend on more practical reasons related to the local geography. For example, most coastal states are more densely populated, making travel costs lower for local teams. It is still worth noting that, in both medoid states, the median of the

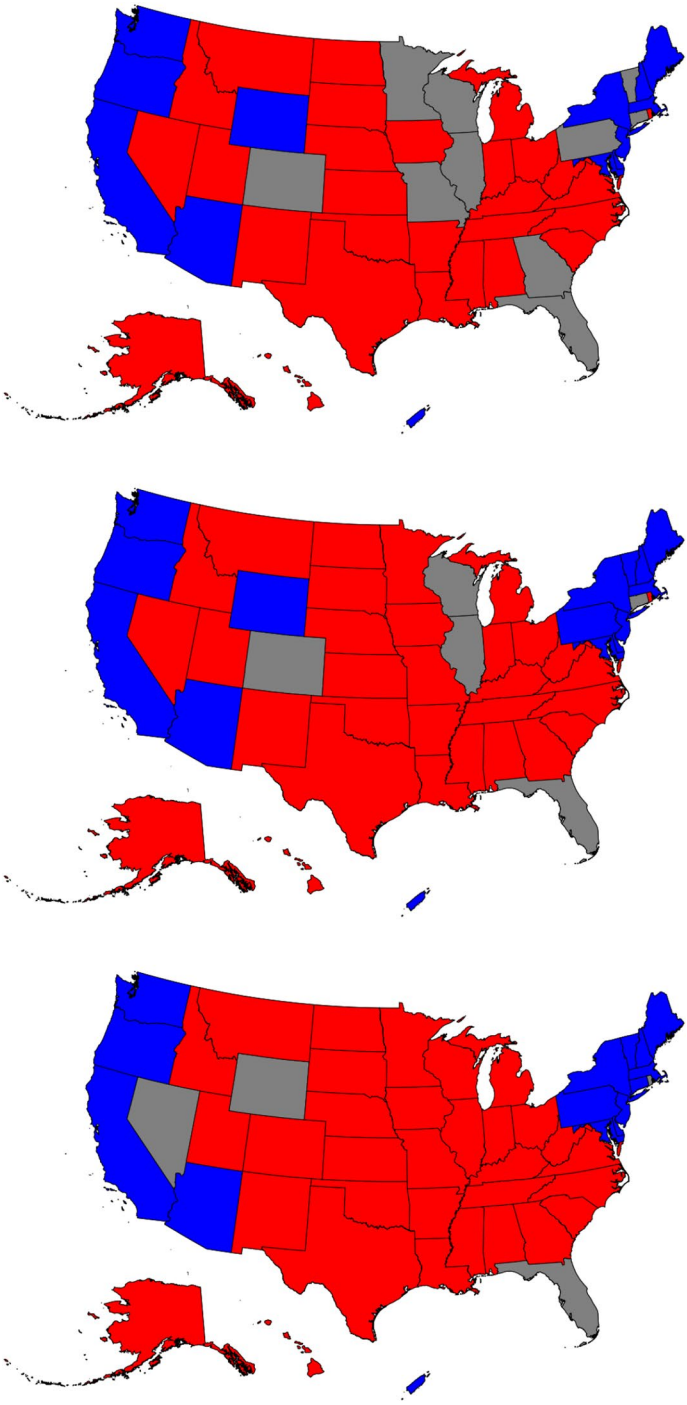

Fig. 3 Map of the cluster memberships with $\gamma = 0$, $\gamma = 0.3$ (middle) and $\gamma = 0.5$ from top to bottom. Red color indicates cluster 1, blue cluster 2, and grey fuzzy membership (no membership above 0.7)

ratio between each expense for the female and male team is below 1, showing that, even in more progressive states, the universities that have achieved full equality of funding are still a minority.

We see how the spatial term introduced by the fuzzy modularity enhances the geographic cohesion of the clusters without changing the nature and the meaning of the clusters themselves. We observe that the partition obtained for $\gamma = 0.3$ has an even higher fuzzy silhouette than the one for $\gamma = 0$. This is a sign of the fact that introducing spatial correction also improves the ability to cluster correctly along attribute lines.

This is linked to the fact that the United States already shows a very strong and further increasing geographic polarization of politics (Johnston et al. 2020). The Pacific Coast and New England are characterized by a largely urbanized and progressive population, which is more sensitive towards social issues such as gender equality, while the Southeast and the central states of the Great Plains and Rocky Mountains have a mostly rural population that holds more conservative positions and values, which consider most sports a masculine endeavor. This inevitably reflects in the policies of local governments and university administrations, creating a clear geographical divide in the distribution of resources for men's and women's sports activities, which is enhanced, but not created from scratch, by the spatial correction term in our model.

At $\gamma = 0.5$ we see instead that the fuzzy silhouette decreases, a sign that to improve even further the spatial cohesion of the cluster, it is necessary to sacrifice something in terms of attribute-based cohesion.

We see that most of the states that were fuzzy in the clustering without spatial regularization ($\gamma = 0$), like Georgia, Vermont, and Minnesota are assigned to the same cluster as their neighbors at $\gamma = 0.3, 0.5$. On the other hand, local outliers like Wyoming (in Cluster 2 but surrounded by Cluster 1 states in the partition found for $\gamma = 0$) and Rhode Island (the other way around), turn into fuzzy units at $\gamma = 0.5$, when the spatial information is regarded as equal in relevance to the attribute distributions. The full description of memberships for all the 3 partitions is given in Table 3.

We further observe that the addition of a spatial term somewhat slows down the convergence of the algorithm, which, as shown in Fig. 4, takes more steps with higher values of γ. This is a consequence of the fact that modularity is very far from a convex function and is difficult to optimize in most cases. Still, even with $\gamma = 0.5$, the convergence of the algorithm is very fast on such a small dataset, taking on average less than 0.06 seconds.

4 Conclusions

We have analyzed the geographic divide in the way resources are distributed between male and female teams in US college sports. To do so, we have designed a new version of the Fuzzy C-Medoids with Modularity Regularization algorithm from Cangemi et al. (2025), which is capable of processing data aggregated at the state level. Using a non-parametric distance between distributions, we can compare the empirical distributions of various equality indices in different states, even in the

Table 3 The units' membership values in the application of the FCMd-MSC with $p = 0.6$, $C = 2$, and different values of γ

	$\gamma = 0$		$\gamma = 0.3$		$\gamma = 0.5$	
	C1	C2	C1	C2	C1	C2
AK	0.99	0.01	0.97	0.03	0.93	0.07
AL	0.93	0.07	0.91	0.09	0.93	0.07
AR	0.96	0.04	0.98	0.02	0.99	0.01
AZ	0.08	0.92	0.17	0.83	0.17	0.83
CA	0.07	0.93	0.07	0.93	0.03	0.97
CO	*0.65*	*0.35*	*0.68*	*0.32*	0.82	0.18
CT	*0.67*	*0.33*	*0.35*	*0.65*	0.10	0.90
DC	*0.54*	*0.46*	*0.45*	*0.55*	*0.32*	*0.68*
DE	0.04	0.96	0.02	0.98	0.01	0.99
FL	*0.49*	*0.51*	*0.56*	*0.44*	*0.66*	*0.34*
GA	*0.55*	*0.45*	0.73	0.27	0.85	0.15
HI	1.00	0.00	0.99	0.01	0.96	0.04
IA	0.98	0.02	0.96	0.04	0.98	0.02
ID	1.00	0.00	0.98	0.02	0.84	0.16
IL	*0.45*	*0.55*	*0.67*	*0.33*	0.87	0.13
IN	0.89	0.11	0.90	0.10	0.93	0.07
KS	0.77	0.23	0.83	0.17	0.91	0.09
KY	0.97	0.03	0.98	0.02	0.99	0.01
LA	0.99	0.01	0.99	0.01	0.98	0.02
MA	0.22	0.78	0.05	0.95	0.01	0.99
MD	0.16	0.84	0.11	0.89	0.04	0.96
ME	0.00	1.00	0.01	0.99	0.02	0.98
MI	0.87	0.13	0.84	0.16	0.86	0.14
MN	*0.64*	*0.36*	*0.70*	*0.30*	0.83	0.17
MO	*0.66*	*0.34*	0.91	0.09	0.98	0.02
MS	1.00	0.00	1.00	0.00	0.99	0.01
MT	0.98	0.02	0.96	0.04	0.95	0.05
NC	0.96	0.04	0.95	0.05	0.95	0.05
ND	0.71	0.29	0.77	0.23	0.83	0.17
NE	0.96	0.04	0.94	0.06	0.97	0.03
NH	0.10	0.90	0.03	0.97	0.01	0.99
NJ	0.06	0.94	0.02	0.98	0.01	0.99
NM	0.97	0.03	0.95	0.05	0.93	0.07
NV	0.99	0.01	0.90	0.10	*0.49*	*0.51*
NY	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
OH	0.99	0.01	0.97	0.03	0.94	0.06
OK	0.98	0.02	0.98	0.02	0.99	0.01
OR	0.14	0.86	0.12	0.88	0.04	0.96
PA	*0.51*	*0.49*	0.12	0.88	0.02	0.98
PR	0.01	0.99	0.03	0.97	0.08	0.92
RI	0.99	0.01	0.89	0.11	*0.52*	*0.48*
SC	0.94	0.06	0.90	0.10	0.89	0.11
SD	0.99	0.01	0.97	0.03	0.98	0.02
TN	0.98	0.02	0.99	0.01	0.99	0.01
TX	**1.00**	0.00	**1.00**	0.00	**1.00**	0.00

Table 3 (continued)

	$\gamma = 0$		$\gamma = 0.3$		$\gamma = 0.5$	
	C1	C2	C1	C2	C1	C2
UT	1.00	0.00	0.99	0.01	0.94	0.06
VA	0.96	0.04	0.93	0.07	0.84	0.16
VT	*0.60*	*0.40*	0.16	0.84	0.04	0.96
WA	0.06	0.94	0.10	0.90	0.10	0.90
WI	*0.37*	*0.63*	*0.51*	*0.49*	0.74	0.26
WV	1.00	0.00	0.97	0.03	0.85	0.15
WY	0.03	0.97	0.27	0.73	*0.64*	*0.36*

The membership values of cluster medoids are in bold, those of fuzzy units (maximal membership< 0.7) in italics

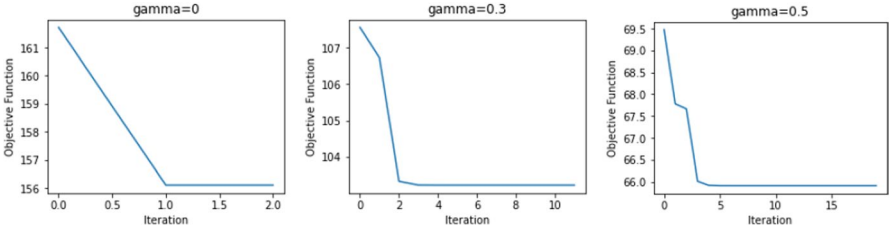

Fig. 4 Evolution of the objective function in each iteration of Algorithm 1 in the optimal cases for $\gamma = 0; 0.3; 0.5$

presence of incomplete data and without making any assumptions about the specific shape of each distribution.

Our algorithm reconstructs the well-documented divide between the more conservative states, represented by Texas as their medoid, and the more progressive ones, represented by New York. This geographic distinction is already noticeable without an explicit spatial correction. We observe that the units in Cluster 1. with medoid Texas, our indicators show a greater level of resource inequality between women's and men's sports, while in units in Cluster 2, with medoid New York, the differences are still present but less pronounced. The spatial correction allows us to obtain more geographically coherent clusters, while maintaining the same large-scale division. Indeed, the model still outputs two clusters with the same two medoids at the three different levels of the weight given to the spatial correction, but reassigns, or turns into fuzzy units, some states bordering states in the opposite cluster.

This analysis highlights several strengths of our proposed model. We see how using actual units as medoids instead of abstract points in the space of distributions avoids the possible complications linked to the non-uniqueness of the centroid, and makes the interpretation of the output much more natural. Indeed, the centroid of a cluster of empirical cumulative distribution functions with different numbers of observations would not be easy to understand as a representative of the cluster and, further, it might not even be unique.

Moreover, the fuzzy nature of the clustering allows us to obtain a more nuanced understanding of what, in this case, appears to be a binary division of the units.

Using a fuzzy algorithm, we can identify both states with intermediate characteristics between those of the medoids and local outliers, that is, states that have different attributes from those of their neighbours, as fuzzy units. To further enhance outlier detection, it would be interesting to explore robust versions of the algorithm that would be more directly tailored to the detection and handling of local and/or global outliers.

Finally, this study expands on the work in Cangemi et al. (2025), which established the Fuzzy C-Medoids with modularity spatial correction algorithm in a distance-agnostic manner, but tested it on data belonging to an Euclidean space. We showcased with this application that the algorithm can be generalized to different settings in which the data structure requires using a different type of distance. It will be of interest in future studies to test how to apply the model data types and structures.

Funding Open access funding provided by Luiss University within the CRUI-CARE Agreement. No specific funding was received for this research.

Data availability Code and data used in the analysis can be found at https://github.com/LopiJ90/Analysis-of-Gender-Equality-in-US-college-sports-funding-via-spatial-clustering/tree/main.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

Ethical approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

References

Campello RJ, Hruschka ER (2006) A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets Syst 157(21):2858–2875

Cangemi D, D'Urso P, De Giovanni L, Federico L, Vitale V (2025) Fuzzy clustering and community detection: an integrated approach. J Classif. https://doi.org/10.1007/s00357-025-09520-7

Carpita M, Pasca P, Arima S, Ciavolino E (2023) Clustering of variables methods and measurement models for soccer players' performances. Ann Oper Res 325(1):37–56

Coppi R, D'Urso P, Giordani P (2010) A fuzzy clustering model for multivariate spatial time series. J Classif 27:54–88

Devine C (2018) Sex, sport and money: voice, choice and distributive justice in England, Scotland and wales. Sport Educ Soc 23(9):824–839

D'Urso P, De Giovanni L, Vitale V (2023) A robust method for clustering football players with mixed attributes. Ann Oper Res 325(1):9–36
D'Urso P, De Giovanni L, Federico L, Vitale V (2023) Fuzzy clustering of spatial interval-valued data. Spatial Stat 57:100764
D'Urso P, De Giovanni L, Federico L, Vitale V (2024) Fuzzy clustering with Barber modularity regularization. Stat Comput 34(6):214
D'Urso P, De Giovanni L, Federico L, Vitale V (2025) Network and attribute-based clustering of tennis players and tournaments. Comput Stat 40(4):1689–1712
Flake CR, Dufur MJ, Moore EL (2013) Advantage men: the sex pay gap in professional tennis. Int Rev Sociol Sport 48(3):366–376
Hodges J Jr (1958) The significance probability of the Smirnov two-sample test. Ark Mat 3(5):469–486
Johnston R, Manley D, Jones K, Rohla R (2020) The geographical polarization of the American electorate: a country of increasing electoral landslides? GeoJournal 85(1):187–204
Koivula N (2001) Perceives characteristics of sports categorized as gender-neutral, feminine and masculine. J Sport Behav 24(4):377–393
Krishnapuram R, Joshi A, Yi L (1999). A fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. FUZZ-IEEE'99. 1999 IEEE international fuzzy systems. conference proceedings (cat. no. 99ch36315) (Vol. 3) pp. 1281–1286.
Liu XF, Liu YL, Lu XH, Wang QX, Wang TX (2016) The anatomy of the global football player transfer network: club functionalities versus network properties. PloS One 11(6):e0156504
Lu H, Tan Y (2003) Unsupervised clustering of dominant scenes in sports video. Pattern Recogn Lett 24(15):2651–2662
Mercer HC, Edwards PS (2020). An analysis of gender inequality in professional tennis: a study of the cozening sport. Applied econometric analysis: emerging research and opportunities. IGI Global pp. 121–140.
Miyamoto S, Mukaidono M (1997). Fuzzy c-means as a regularization and maximum entropy approach. Proc of 7th international fuzzy systems association world congress (ifsa'97), ii (pp. 86–92).
Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Comb Probab Comput 7(3):295–305
Narizuka T, Yamazaki Y (2019) Clustering algorithm for formations in football games. Sci Rep 9(1):13172
Nepusz T, Petróczi A, Négyessy L, Bazsó F (2008) Fuzzy communities and the concept of bridgeness in complex networks. Phys Rev E-Stat Nonlinear Soft Matter Phys 77(1):016107
Norman M, Donnelly P, Kidd B (2021) Gender inequality in Canadian interuniversity sport: participation opportunities and leadership positions from 2010–11 to 2016–17. Int J Sport Policy Polit 13(2):207–223
Norros I, Reittu H (2006) On a conditionally poissonian graph process on a conditionally poissonian graph process. Adv Appl Probab 38(1):59–75
Office of Postsecondary Education of the U.S. Department of Education (2025). Equity in athletics data analysis cutting tool. https://ope.ed.gov/athletics/#/datafile/list. Accessed: (10/02/2025)
O'Connor JJ (2021) The means to an end: an examination of gender inequality in athletic aid distribution and graduation rates. Sport Soc 24(4):534–550
Pham DL (2001) Spatial models for fuzzy clustering spatial models for fuzzy clustering. Comput Vis Image Underst 84(2):285–297
Ribeiro AS, Lima F (2019) Football players' career and wage profiles football players' career and wage profiles. Appl Econ 51(1):76–87
Yenilmez MI (2021) Gender inequality and female sports participation in Turkey gender inequality and female sports participation in turkey. Central Eur J Sport Sci Med 33(1):27–41
Yiapanas G (2025) Addressing gender inequalities in European football: key dimensions and strategies insight. Sports Sci 7(1):711