Research Article

Nathan Carter*, Andrew Harrison, Amar Iyengar, Matthew Lanham, Scott Nestler, Dave Schrader and Amir Zadeh

Clustering algorithms to increase fairness in collegiate wrestling

https://doi.org/10.1515/jqas-2020-0101
Received September 4, 2020; accepted May 4, 2022;
published online June 28, 2022

Abstract: In NCAA Division III Wrestling, the question arose how to assign schools to regions in a way that optimizes fairness for individual wrestlers aspiring to the national tournament. The problem fell within cluster analysis but no known clustering algorithms supported its complex and interrelated set of needs. We created several bespoke clustering algorithms based on various heuristics (balanced optimization, weighted spatial clustering, and weighted optimization rectangles) for finding an optimal assignment, and tested each against the generic technique of genetic algorithms. While each of our algorithms had different strengths, the genetic algorithm achieved the highest value on our objective function, including when comparing it to the region assignments that preceded our work. This paper therefore demonstrates a technique that can be used to solve a broad category of clustering problems that arise in athletics, particularly any sport in which athletes compete individually but are assigned to regions as a team.

Keywords: cluster analysis; collegiate athletics; genetic algorithms; sports analytics; wrestling.

***Corresponding author**: **Nathan Carter**, Bentley University, Waltham, USA, E-mail: ncarter@bentley.edu. https://orcid.org/0000-0002-9293-5879
Andrew Harrison, University of Cincinnati, Cincinnati, USA
Amar Iyengar and Matthew Lanham, Purdue University, West Lafayette, USA, E-mail: amiyengar@gmail.com (A. Iyengar), lanhamm@purdue.edu (M. Lanham)
Scott Nestler, Notre Dame University, South Bend, IN, E-mail: snestler@nd.edu
Dave Schrader, Teradata University Network for Academics, San Diego, CA, E-mail: drdaveschrader@gmail.com.
Amir Zadeh, Wright State University, Dayton, USA, E-mail: amir.zadeh@wright.edu. https://orcid.org/0000-0002-3171-5629

1 A clustering problem from the NCAA

1.1 Background

While this paper demonstrates a technique applicable to a variety of problems, it was motivated by a particular problem posed by the then-president of the Division III (DIII) Wrestling Coaches Association (NCAA, 2020) of the National Collegiate Athletics Association (NCAA). The NCAA partitions schools involved in DIII Wrestling into a set of regions for the purposes of regional tournaments. (The number of schools changes from year to year, typically just over 100). High-ranking wrestlers from each region represent that region at an annual national tournament. The problem was that most coaches did not find the selection process fair when we surveyed them.

Fairness is a central theme of NCAA sports, but the creation of competitive regional tournaments is constrained by other factors as well, including travel distances and expenses. Regional organization plays a significant role in collegiate wrestling and affects the results of national tournament performance (Bigsby and Ohlmann, 2017).

In NCAA head-to-head dual competitions, wrestling teams send ten wrestlers to a tournament, with each wrestler competing in a different weight class. Individual matches occur between wrestlers in the same weight class, and each match results in a win or loss. Individual wrestlers must compete at the same location as their team members, and their match performances are aggregated to determine team victories in dual meet and tournament settings. Historically, the teams have been divided mainly by geography and tradition into six regions, and region assignments are updated every three years. In each of ten weight classes, the three highest placing wrestlers at each of the six regional tournaments are invited to compete at the national tournament, for a total of 180 wrestlers invited.

A small example can demonstrate the unfairness that motivated the coaches' concerns. Imagine for a moment

Table 1: Hypothetical data of wrestlers from six regions, ranked from the most skilled downward, with equally skilled wrestlers aligned horizontally.

Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
					1
1		1		1	
2		2	1	2	
		3		3	
3		4	2		2
4	1	5	3	4	
5	2	6	4	5	3
6	3	7	5	6	

that we knew the true skill of each NCAA DIII wrestler in a particular weight class and arranged all of them in columns of descending skill, one column per region. Hypothetical data of this type is shown in Table 1. Wrestlers of equal skill are in the same row, and those of higher skill are in higher rows than those of lower skill. Only the top few rows of the table are shown; one can imagine it continuing downward with additional wrestlers not relevant to this example.

Clearly from the table, some regions have a higher concentration of skilled wrestlers than others. For instance, Region 3 has several wrestlers that are better than each wrestler in Region 2. If we chose the top three wrestlers from each region to attend the national tournament, it would not include the best set of DIII wrestlers nationwide. Specifically, there exist several pairs of wrestlers with one being the higher skilled wrestler in the pair and yet the other attending the tournament while the first does not. One such example is wrestler 4 from Region 3 (higher skilled, not attending) and wrestler 1 from Region 2 (lower skilled, attending). In each such pair, the higher skilled wrestler would have an understandable complaint of unfairness.

Although the hypothetical data in Table 1 assumed that we knew the "true" skill of each wrestler, even if we had no such measurement, it would still be possible to perceive instances of unfairness because wrestlers and coaches are good at assessing relative skill. Furthermore, such unfairness is an unintended consequence of how the NCAA partitioned the schools into regions in the first place and is thus not within the control of the student athletes, their coaches, or their schools.

1.2 Formalizing the problem

The NCAA's 2016–17 DIII regional assignments had unbalanced regional sizes and strengths, resulting in dissatisfaction among coaches and wrestlers. The region assignment is illustrated in Figure 1. DIII wrestling schools are almost exclusively in the northeastern quadrant of the country. One geographic outlier sits in the northwest and two in the south.

Unfairness of the type illustrated by Table 1 was exaggerated by two realities. First, the competitive landscape among DIII wrestling teams was quite unbalanced; in the last 28 years, only two schools have won national

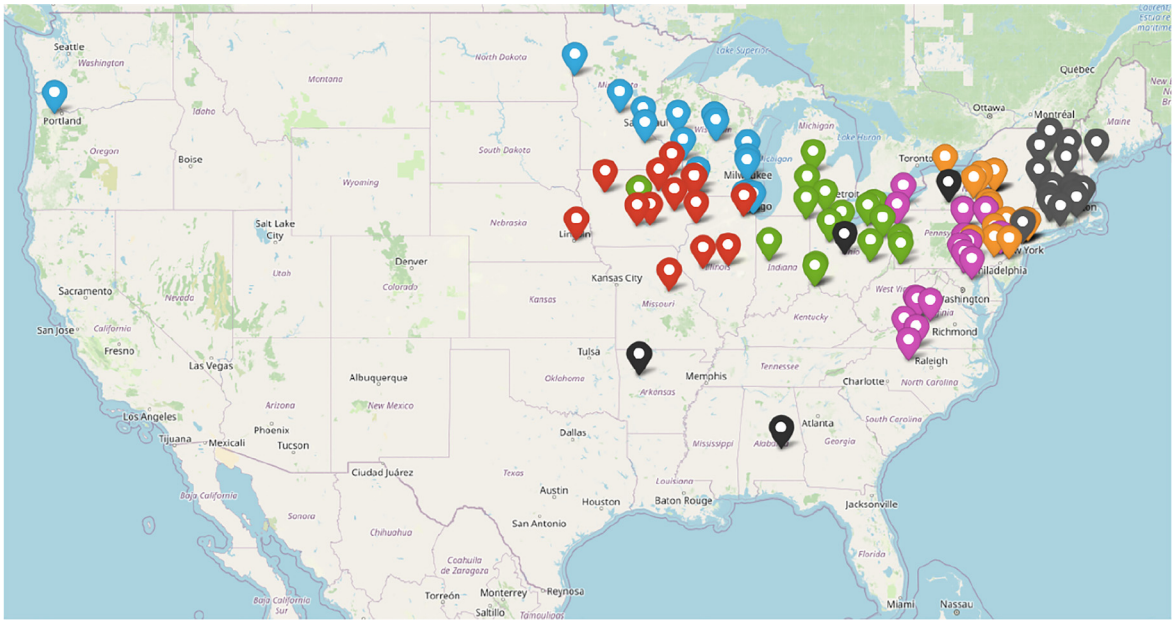

Figure 1: Original assignment of schools to regions.

titles, Wartburg College in Iowa (13 titles) and Augsburg University in Minnesota (12 titles). Second, some regions had as few as 11 teams, while others contained as many as 19, yet the same number of top wrestlers per weight class were selected from each region for the national tournament, which was the main driver for the unfairness complaints.

The Coaches Association President at the time heard the coaches' concerns and mentioned the problem to one of the authors. To diagnose the problem more precisely, the two worked with another member of our team to design and run a Survey of the DIII Wrestling Coaches in 2017. The most notable results from that survey were that 65% of coaches agreed that the existing system was unfair while 50% said they were satisfied with the system (Figure 2). Clearly at least 15% were happy with a system they admit is unfair, possibly because their students were its beneficiaries.

Furthermore, the survey results were quite polarized. For instance, one survey question asked coaches which they thought was more important: for each region to have roughly the same average strength of its schools' wrestling programs or for each region to be roughly the same size. On a seven-point scale from similar strength to similar region size, the most common answers were those two extremes. From this and other polarized survey results, it was clear that the problem was complex. It turned out that the NCAA had done a similar survey; our results and theirs came to the same conclusion. Some of the history of grappling with these challenges is described on the division's website (NCAA, 2015).

Thus the authors came together to address the following research question: Could an approach driven by data analytics partition the schools into regions in a way that reduces or eliminates this perceived unfairness?

1.3 Formulating an objective function

We thus have an optimization problem that seeks a region assignment for all the NCAA DIII Wrestling schools in a way that maximizes fairness. The number of regions was fixed at its current number, six. But before one can implement a data-driven solution, the objective function needs to be made more specific.

Of the 106 schools in NCAA DIII Wrestling at the time of our analysis, three were geographic outliers, which might skew the results of any analysis we devised. Those schools were Pacific University (Forest Grove, OR), Huntingdon College (Montgomery, AL), and the University of the Ozarks (Clarksville, AR). Since any region assignment forces these schools to travel great distances to all their meets, we omitted them from our analysis. They can be added afterward to any region that needs additional members and for which they would not imbalance its Power Rating (defined below). Or those schools could be permitted to choose which region to enter, perhaps based on flight costs to nearby airports. Thus all analyses in this document used $n = 103$ schools.

Fairness cannot be the only component to the objective function; if regions were balanced by power alone, schools that are very disparate geographically would be placed into the same "region." Thus most of these so-called regions would actually be spread across the map, creating undue travel burdens and costs on the schools. Prioritizing power balance alone would also ignore the problem of unequal numbers of teams per region.

Thus our objective function must seek to achieve these three separate goals: making regions similar in number of teams, similar in power, and geographically compact. Specifically, it will be some combination of the following quantities.

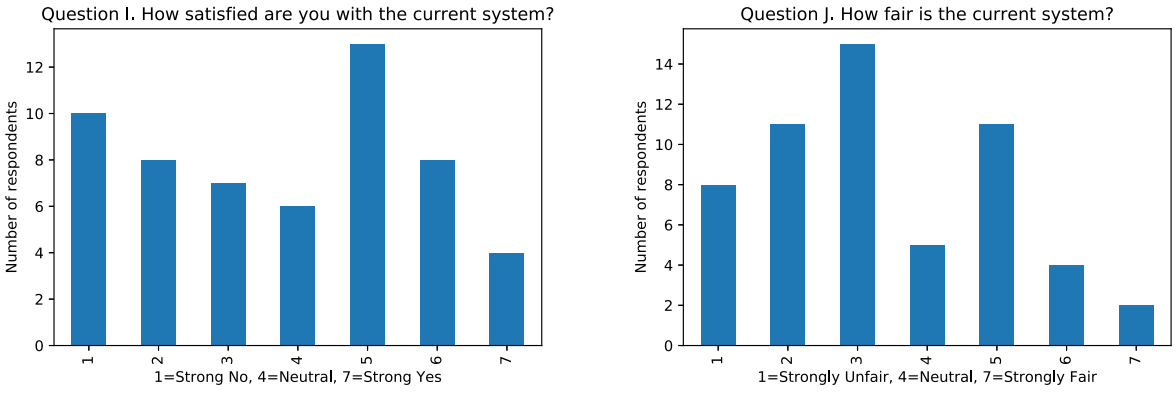

Figure 2: Results of two example questions in the survey of coaches.

1. The variance of the number of teams in each region.
2. The variance, across regions, of the average power of schools in that region.
3. The total, across all regions, of the expected travel for schools to their regional tournament.

To help us determine how best to prioritize these competing factors, the aforementioned surveys specifically asked about the relative importance of these factors. Data summarizing the coaches' responses appears in Table 2. The results shown there are close enough to equal weights that we chose to rate the three competing factors as equal.

This data from the primary stakeholders in the problem is quite valuable. If we did not know that the primary audience for our work considered the factors of roughly equal importance, the competing factors would create a set of Pareto optimal solutions among which it would be unclear whether a "best" solution existed. But with these survey results, if we can quantify each factor, we can combine them into an objective function symmetrically (In the case of Pareto optimal—or "trade-off"—solutions, strategies still exist for how to let decision makers interact with the optimization process. For instance, in the case of genetic algorithms, which we will employ in Section 3.4, Zhou et al. (2011) survey a variety of approaches. But we will not need these approaches because of the survey results.)

We are thus left with the task of quantifying each of the three factors. The first of the three is the easiest to quantify: The size of a region is the number of schools in it, a positive integer. We will measure how balanced this is by computing the variance of the region sizes and using that as a factor to minimize.

The second quantity is harder to quantify: How strong is a school? We answer this question in Section 1.4. Once we have a measurement for the power of a school, we can then compute the mean power of all schools in a region, and compute the variance of these means across regions, another quantity we will seek to minimize.

The third quantity is geographic. The location of each school as a latitude-longitude pair is publicly available data. For any grouping of schools into a region, we can calculate the centroid of the region (as well-known clustering algorithms do, though it typically won't equal any school's location). We can then calculate the distance from each school to that centroid and then sum those distances as a proxy for the typical travel time to the regional tournament. Because of varying population density in different regions of the country, some of these total travel distances may be much higher than others (especially for rural regions), but we seek to minimize their sum across all regions.

Table 2: Coaches' assessment of competing priorities' importance in region assignment.

Polls	Equalizing teams per region	Equalizing team strength	Geography
NCAA	33.5%	31.7%	34.9%
Our poll	32.8%	33.1%	34.1%

School locations are given as latitude-longitude pairs. Because the distance between latitude lines is not the same as the distance between longitude lines, we cannot use the standard Euclidean metric on latitude-longitude pairs; it would not treat north/south distances the same as east/west distances. Thus we use geodesic distance on the surface of the earth as the distance metric.

We need to combine these three factors equally, to follow the coaches' preferences. We use the technique of desirability functions, first introduced in Derringer and Suich (1980). For each factor, a suitable range $[a, b]$ is chosen with undesirable values at a and desirable values at b (by multiplying the factor by -1 to reorient first if necessary). Then each factor is linearly scaled so that $[a, b]$ maps to $[0, 1]$, values are clamped to within that interval, and the geometric mean of all factors is then computed to combine them into the objective function. This requires some subjectivity in choosing the values of a and b for each factor. We did so as follows.

1. The most desirable variance in region sizes is zero and the least desirable is the variance of the sizes 5, 10, 15, 20, 25, 29, which is $82\frac{2}{3}$. These numbers add to 103 but vary far too much if interpreted as region sizes.
2. The most desirable variance across regions of mean power within the region is zero and the least desirable is computed by placing the weakest $\frac{1}{6}$ of the schools in a region together, then the next weakest $\frac{1}{6}$ into another region, and so on, and then computing the mean power of each region, and the variance of those numbers (resulting in approximately 9.25).
3. The most desirable total travel distance across all regions is zero and the least desirable was chosen to be 500 miles times 17 (schools per region) times 6 (number of regions), giving 51,000 miles. The number 500 was selected because a region with radius 500 miles would be very large.

We therefore have an objective function; the ranges given above can be used to scale each component into the range $[0, 1]$, and the geometric mean of those three values will

therefore remain in $[0, 1]$, and it will be the value of the objective function.

In addition to being organized into regions, schools are also organized into conferences, and historically entire conferences are assigned to the same region. We did not consider conferences in our work, for two reasons. First, we were tasked only with assigning schools to regions, not conferences. Second, the NCAA DIII Wrestling commissioner told our research team that it would likely be acceptable for us to ignore conferences when selecting regions. He suspected that a principled approach that achieved all the balancing goals we incorporate into our work would be attractive enough to permit separating conferences.

1.4 Modeling the power of a school

Given the list of schools and the number of regions, the only missing data was a measurement of how strong each school's wrestling program is. One of the authors, who had been a collegiate wrestler, developed a mathematical model for the power of a school's wrestling program and shared it with the rest of our team. That model estimates a team's performance at the national tournament based on past performance, and was constructed as follows.

Websites such as Intermat.com and FloWrestling.com contain detailed, publicly-available data on past NCAA DIII Wrestling matches. As is common in applied problems, some data is missing, and a research assistant invested about 20 h to find the missing information, resulting in data from 18,669 unique matches for the 2016–2017 season. The aforementioned survey of coaches helped select 51 variables that may be relevant.

But many of those variables were strongly correlated; for example, the number of wins wrestlers have been awarded is highly correlated with the number of times they have pinned an opponent. We used the data to factor and consolidate variables, developing a parsimonious model that could predict a team's score at nationals, accounting for 67.7% of the variance (In addition to the usual reasons for preferring a model with few variables, in this case it also means less data that needs to be tracked or collected each year, if the method we develop is to be re-used in the future. The parsimonous model was obtained by dropping the non-significant exogenous factors one at a time until only significant factors remained.) This level of predictive accuracy is similar to that of other analytical approaches for predicting the outcomes of individual matches, meets, and tournament results (Bigsby and Ohlmann, 2017). The resulting model defines "Power Rating" as the number of points we expect the team to earn at the national tournament and calculates it from a combination of last year's performance and this year's performance. Using previous years' performance introduces some bias, a limitation we will discuss in Sections 5 and 6.

In year y, we write A_y for the number of All Americans a school had (wrestlers who finished in the top eight in each weight class in the individual national championship tournament), N_y for the team's score at the national tournament, W_y for the number of wins in the season, M_y for the number of matches in the season, P_y for the number of pins against the team that season, and T_y for the number of technical falls their wrestlers suffered that season. Then the model (an estimate of N_y) is as follows.

$$\text{Power Rating} = 0.376A_{y-1} + 0.151N_{y-1} + 0.621\frac{W_y}{M_y} - 0.357\frac{P_y + T_y}{M_y}$$

Because pins and technical falls are considered significant losses, the final term expresses the team's "big loss proportion." This equation contains compilation variables, which are used to develop multi-level models where team performance is derived from individual performances of players (Bliese 2000). The top ten Power Ratings across all schools are shown in Table 3 as examples.

This model turned out to be relevant when we considered how to incorporate the strength of a school's wrestling program into the question of region assignment. The assignment of teams to regions happens only once every three years or so, during the summer between seasons, so all past seasons' data are available. This infrequent reassignment happens in order to give schools time to plan ahead for who will be hosting regional tournaments. One would therefore think that the easiest measurement of a school's power at the national tournament would simply be the actual data of the number of points they scored during the previous year's national tournament.

But coaches were concerned that if the measure of school power were its most recent national tournament results, the performance of recently-graduated seniors could skew that measure, since they score the vast majority of a school's points at the national tournament, but are no longer on the team. Thus coaches preferred to incorporate only national tournament data older than the previous year. By contrast, they found it acceptable to include regular season performance for the past year in the model, because it aggregated the performance of all wrestlers, not just the team's best.

Table 3: Power ratings of ten highest-rated schools.

Institution	Location	Power rating
Wartburg College	Waverly, IA	22.9910
Messiah College	Mechanicsburg, PA	17.7960
Luther College	Decorah, IA	11.8710
Wabash College	Crawfordsville, IN	9.8450
Augsburg University	Minneapolis, MN	8.8900
University of Wisconsin La Crosse	La Crosse, WI	7.0950
Stevens Institute of Technology	Hoboken, NJ	6.6440
State University of New York at Cortland	Cortland, NY	6.4450
Elmhurst College	Elmhurst, IL	6.2500
Alma College	Alma, MI	5.9410

Thus we were able to use the same model from above, unchanged, to model school power. Its use of A_{y-1} and N_{y-1} satisfied the coaches' concern about graduating seniors (since $y-1$ is a previous year), and the other variables, W_y, P_y, T_y, and M_y, involve all team members. This has the added benefit that if the process for regional reassignment ever needs to be done before the current year's national tournament, it can be, since it is not dependent on the tournament's results. As a test of robustness, we presented our predictive model back to a group of DIII wrestling coaches who confirmed that it seemed anecdotally correct.

This model therefore fills in the one missing piece in the objective function described in Section 1.3, enabling us to look for algorithms that will minimize that function. The one exception is when new schools join NCAA DIII Wrestling, in which case they would not have a Power Rating, and yet they would need to participate in the objective function as it evaluates region assignments that include them. For such cases in our work, we used a rating of zero because no data was available. Filling in missing values well is a research project unto itself, and there are surely more accurate techniques that could be used here, such as models based on the school's location, size, etc., but this paper focuses on the optimization problem rather than the problem of creating a useful model for filling missing power data.

2 Existing clustering methods

Because we aim to group the 103 schools in our analysis into regions, we are facing a cluster analysis problem. There are an enormous number of clustering algorithms in existence; over 30 are listed for the use of biologists alone in Andreopoulos et al. (2009). But none of the existing algorithms will address our complex objective function. There are a few reasons for this.

First, k-means clustering with the standard Euclidean distance metric, which provides only a local minimum on each run, is NP-hard to solve exactly. Adding more constraints makes it even more difficult.

Second, cluster analysis is often framed in terms of grouping like or similar objects together, while keeping unlike or dissimilar objects in separate clusters, using some natural criterion of similarity (Duda and Hart, 1973). Some authors even define cluster analysis as being restricted to such problems (Han, Kamber, and Pei 2012). But the objective function in our problem cannot be boiled down to a mere similarity measure; some of its components are inherently about the relationship among clusters, not just individuals.

Estivill-Castro (2002) says that one reason that there are so many clustering algorithms is that there are many definitions of what clustering is, and those definitions can be domain-specific. He suggests at least two options for coping with domains that are more complex than existing clustering algorithms can handle.

First, one might just use existing methods anyway, as heuristics that we hope will give a good enough answer, and then compare their results to the correct objective function to see how well the heuristic performed. For example, if one wanted to do k-means clustering with equal-sized clusters (just one of our constraints), one can begin with the standard algorithm and add a step called the "Hungarian algorithm" to balance the size of clusters (Malinen and Fränti, 2014). A similar approach is taken in Zhu et al. (2010). We will later refer to this approach as "extending existing algorithms."

Second, genetic algorithms (and hybrid approaches) have been shown to be superior to local search methods if they give a better tradeoff between quality of answer and computational resources invested. We applied both of these approaches suggested by Estivill-Castro (2002) to the

clustering problem for the NCAA, to see which would be most effective in this situation.

3 Algorithms

Our team comprises researchers from various colleges and universities; the researcher teams at each school independently considered the problem posed at the end of Section 1.2. Each team began by understanding the objective function, then developed a solution unique to that team, which solutions we compared afterwards across teams, using the objective function as a judge. Most teams used the approach of extending existing algorithms, but one used a genetic algorithm. Tools used in various teams' algorithms include R, Python, and Analytic Solver. SPSS, SmartPLS, MS Excel, and Tableau were used for data cleaning, analysis, and visualization. Each solution takes under 30 min to converge to a best solution, depending on hardware. We review the four approaches in this section, then provide a comparison of how each performed.

3.1 Balanced optimization (based on *k*-means)

Part of our team applied a balanced k-means clustering optimization approach motivated by the cluster analysis methods described in Bradley et al. (1998) and Wagstaff et al. (2001). This approach modified k-means to also prioritize balanced region sizes, then once that had converged, made adjustments to improve balance in region power.

Let $T = \{t_1, \ldots, t_n\}$ be the set of teams to be clustered into regions, where each t_i is a point in the plane, representing the school's location. Begin by randomly assigning teams to initial regions and calculating a centroid for each initial region, $C = \{c_1, \ldots, c_k\}$, with $k = 6$ in this case. As in the standard k-means algorithm, calculate the nearest centroid for each team and assign the team to that centroid. Write T_i for the set of teams assigned to centroid c_i.

The traditional k-means algorithm seeks a partition that minimizes the mean of the squared distances from each team in a region to that region's centroid. For each region T_i, the squared error among the teams in T_i and their respective centroid c_i is $e(T_i) = \sum_{t_j \in T_i} \|t_j - c_i\|^2$. But this does not balance the number of teams in each cluster.

Thus we extend the k-means algorithm by adding three tuning parameters, LB, UB, and λ. The LB parameter represents the lower bound of the size of each region, UB is the corresponding upper bound, and λ is a proportion of teams in too-large regions that will be randomly reassigned to too-small regions, as follows. If we write $|T_i|$ for the size of region T_i, then after each standard k-means step that computes new centroids c_i and assignments T_i, we add this step: For any region for which $|T_i| > \text{UB}$, randomly assign $\lambda \cdot |T_i|$ of those teams to the next region T_j for which $|T_j| < \text{LB}$.

The resulting algorithm is thus a variation of k-means that iterates the above two-step process to convergence. Because k-means can be negatively affected by the random initial clustering, we ran the modified algorithm multiple times with varying random starting centroids and averaged the final converged centroids.

This algorithm uses k-means to satisfy geographic proximity constraints and our modification of k-means to balance region size, but it does not address balance of region power. Thus after convergence of the above algorithm, we modify the regional assignment by reassigning teams near the boundary of two regions to another neighboring region if doing so increases the balance of power across regions. An illustration of this concept appears in Figure 3, with schools that were reassigned circled in red.

This final step is an integer programming problem with binary decision variables $y_{ij} \in \{0, 1\}$, for whether to assign a team t_i on the boundary between two regions to region T_j. Let F stand for the set of fixed teams (those well within the interior of their assigned region) and B stand for the set of boundary teams, which may get reassigned to new regions. Let ρ_i stand for the power of team t_i. Then the total power P_j of region j can be expressed as follows.

$$P_j = \left(\sum_{t_i \in T_j \cap F} \rho_i \right) + \left(\sum_{t_i \in B} y_{ij} \rho_i \right)$$

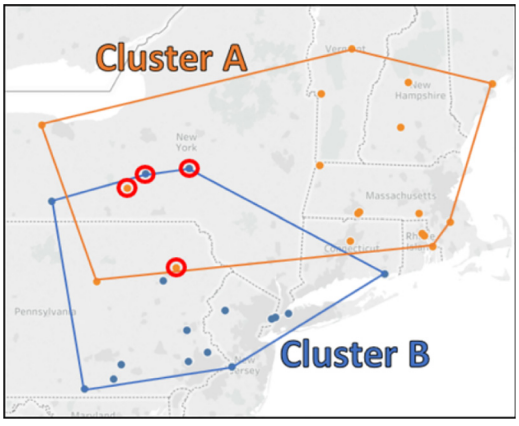

Figure 3: Example of border reassignment.

We then define the integer programming problem's objective function to be the following, which we seek to minimize subject to $|T_j| \in [\text{LB}, \text{UB}]$ for all j. It expresses the average squared distance of each region's power from the mean region power.

$$\frac{1}{k}\sum_{j=1}^{k}\left(P_j - \frac{1}{k}\sum_{j'=1}^{k}P_{j'}\right)^2$$

3.2 Weighted spatial clustering

Other members of our team tried an approach using weighted spatial clustering to partition regions, an extension of standard cluster analysis to address multiple constraints (Gan et al., 2007). In our case, this approach allows a clustering algorithm to account for balancing teams per region, minimizing travel distances, and balancing power across regions.

Before the algorithm began, it took the total power of all schools (219) and divided it by the number of regions that should be formed (six) to produce a goal power of about 36.5 for each region, if power were to be equitably distributed. It then chose balanced region sizes and randomly assigned teams to initial regions of those pre-selected sizes. Centroids were calculated for each region.

Then each team was proposed for reassignment to a new region as follows. The school that was farthest from the center of its assigned region was proposed for trade to a new region, thus preserving all region sizes as balanced. The distance from that school to the center of each potential new region was calculated and reassignments were done only when they decreased the average distance to the center of the region. This process was iterated until the average distance of schools to the center of their regions stopped decreasing.

In the next stage, the cluster solution produced so far was further modified to balance power across clusters. Some region with power below the goal of 36.5 was selected and some of its members were swapped with neighboring clusters in order to increase its power, closer to the goal. This step was iterated until the power of each region was within a chosen threshold of the goal, 36.5.

This method ensured perfect balance of region sizes and scored well on geographical compactness, but less well on region power balance.

3.3 Weighted optimization rectangles

The third approach focused on minimizing the total area of a set of rectangles, one rectangle for each region, defined as the smallest rectangle that bounds all the schools in that region, treating latitude-longitude pairs as points in the plane. This method of modeling region compactness was inspired by techniques in Xu and Wunsch (2008). This method used that total area as the objective function to be minimized, while imposing constraints on power balance across regions and number of teams in each region.

That is, the objective function was formed only to optimize region compactness, and the other two goals (similar numbers of teams and similar powers across regions) were converted into reasonable constraints instead. For instance, using the notation from earlier, one could specify that each region's size $|T_i|$ must satisfy $16 \le |T_i| \le 18$, or some other bounds. The user can manipulate these bounds and experiment with which values lead to solvability. From the solutions produced in this manner, the one we report later in Table 4 is the one with minimum total travel distance. This approach performed acceptably in terms of travel distance and well in the other two criteria; it was the second-best algorithm overall.

The approaches in Sections 3.1 through 3.3 follow the first strategy that Estivill-Castro (2002) suggests, extending existing algorithms. The other strategy he suggested was genetic algorithms, as used by Cowgill et al. (1999) and Hall et al. (1999). We cover it in the following section.

3.4 Genetic algorithms

Genetic algorithms offer some benefits over other cluster analysis strategies when attempting to simultaneously resolve competing interests. According to Hruschka et al. (2009a, 2009b), traditional cluster analysis cannot optimize a solution without implicitly prioritizing some constraints over others, as we've seen in the previous three approaches.

In contrast, a genetic algorithm approach begins with a number of random "solutions." Those solutions, each of which in this case is a way to partition the 103 schools, are bred together to create new offspring solutions that inherit some combination of features from their parents. Then, according to the objective function, only the best new solutions are retained and allowed to breed in the next iteration. Further details of this process appear below and in Cowgill et al. (1999). What is most important is that we need only an objective function. From Sections 1.3 and 1.4, we have one, which equally weights variance in region size, variance in region power, and total travel distance.

To apply the genetic algorithm approach to the NCAA problem, each partition was represented as a vector whose length was the number of teams $n = 103$ and whose k

Table 4: Four approaches to regional assignments.

	2016–2017 Regional alignment			Balanced optimization			Weighted spatial clustering		
Region	***N***	**Power**	**Dist.**	***N***	**Power**	**Dist.**	***N***	**Power**	**Dist.**
West	15	3.65	125.72	16	2.66	67.8	17	2.14	76.04
Central	17	2.34	250.28	18	1.01	110.32	17	2.16	173.55
Midwest	17	2.07	143.35	18	3.32	146.96	18	2.01	94.59
Mideast	18	2.24	141.95	15	2.21	176.04	16	2.25	187.51
East	17	1.74	82.86	18	2.68	292.02	17	2.18	141.75
Northeast	17	1.16	72.53	17	1.01	181.03	17	2.16	225.0
Components	0.99	0.73	0.93	0.98	0.67	0.9	1.0	0.7	1.0
Obj. func.		0.874			0.8412			0.8873	
	Weighted Optimization Rectangles			**Genetic Algorithm**					
Region	*N*	Power	Dist.	*N*	Power	Dist.			
West	17	2.52	144.36	17	2.39	102.28			
Central	16	1.53	66.24	18	2.41	137.23			
Midwest	16	3.12	264.48	17	1.89	75.26			
Mideast	18	1.22	74.93	17	1.86	84.39			
East	16	1.58	109.68	17	2.43	276.51			
Northeast	18	3.04	138.58	17	1.77	143.36			
Components	0.99	0.74	0.93	1.0	0.72	0.99			
Obj. func.		0.8774			0.8942				

entries were in the set $\{1, 2, 3, 4, 5, 6\}$, thus partitioning the teams into six regions. That is, the "DNA" of a partition is a map P: $\{1, \dots, n\} \to \{1, \dots, k\}$.

Breeding two solutions was done by uniform crossover, a common genetic algorithm breeding method in which the child's ith entry is either the mother's ith entry or the father's ith entry, equally likely. Other standard crossover techniques were investigated (e.g., one-point crossover and others involving optimal permutations) but all evolved the pool at about the same rate.

A mutation in this case was performed in 10% of all offspring, and was one of two procedures, chosen randomly at the time of the mutation, each equally likely. The first procedure was designed to move the solution towards any nearby local optimum: choose two random entries, $i, j \in \{1, \dots, n\}$, and swap their partition assignments. The result is a "nearby" solution and thus becomes a candidate for exploration to see if it has a better value under the objective function. The second procedure helps reduce the chance that the evolutionary process will get stuck in a local optimum: it randomly changes precisely one entry in the recently born child. Mutation rates lower than 10% were tested, but did not produce convergence as quickly, probably due to the complexity of the objective function, which may have many local optima.

Over time the "gene pool" of partitions improves as they advantageously inherit beneficial features from their parents or fortunately get them from mutation, and as poor solutions are removed from the gene pool through natural selection. This process does not always reach a global optimum. Indeed, genetic algorithms are typically applied in situations where no algorithm for achieving a global optimum is known. But the random mutations create opportunities for exiting local optima by introducing a solution that is not necessarily "near" any of the other solutions in the pool.

In our analysis, 10 candidate solutions were used for each round of breeding, and the process continued for 10,000 generations. By that time, evolution stopped yielding substantive changes in the objective function (Rapid improvement took place in the first 500 generations, slowing until about 4000 generations, after which it plateaued.) We choose the best-scoring partition from the gene pool when the algorithm terminates.

The resultant solution, shown in Figure 4, was the best one from the final gene pool, as judged by the objective function. The reader may note that several regions overlap slightly, which helps balance region power at a small cost of travel time.

4 Results

Each team reported their recommended regional assignments, then we compared the performance of the four approaches. The results appear in Table 4. Each of the

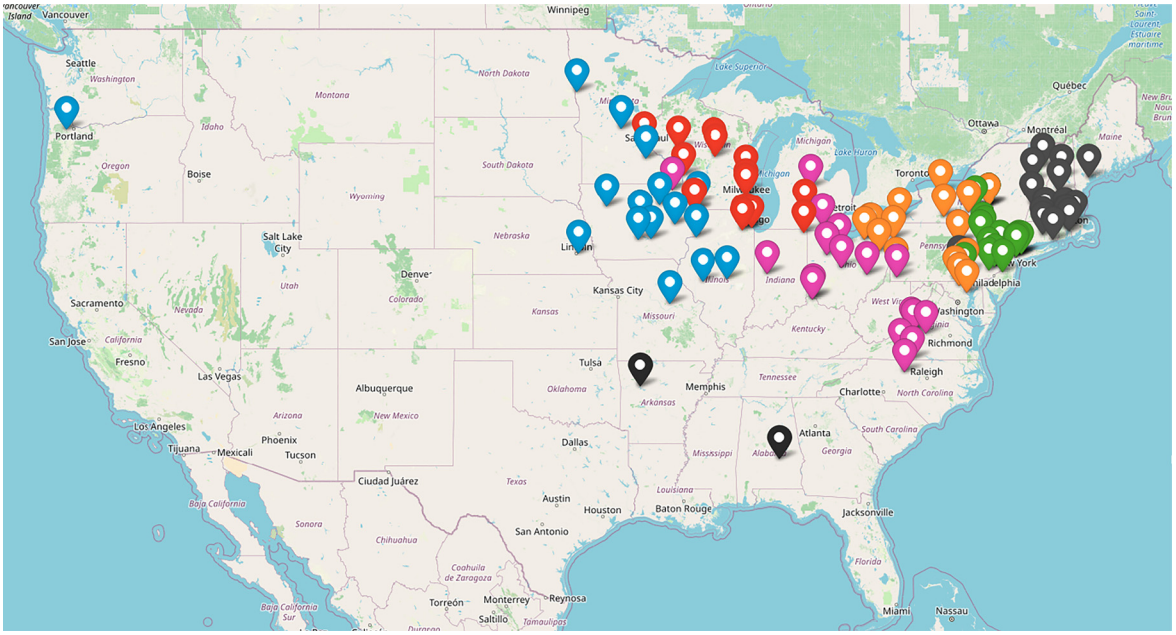

Figure 4: Genetic algorithm regional assignment result.

five major headings represents a way to assign regions, beginning with the 2016–2017 Regional Alignment in place before our work began, and followed by the four approaches documented in Section 3. For each assignment, N represents the number of schools in each region, Power the average power rating of the schools in the region, and Dist. the approximate mean travel distance for all schools to the regional tournament. The "Components" row in each column shows the corresponding values of each component of the objective function, standardized as described in Section 1.3. The "Obj. func." row shows the value of the objective function, which is the geometric mean of the three components.

The Weighted Spatial Clustering approach prioritized a balanced number of schools and therefore tied for the highest score in that component. It also achieved the highest score in the travel distance component but at the cost of achieving a low score in the power component. The genetic algorithm approach is the most balanced, scoring equal or very close to the best scores in all three components. It also achieves the highest score on the objective function.

The same data is summarized visually in Figure 5, with the approaches sorted from left to right in increasing order of their value on the objective function. We can see that the genetic algorithm only slightly outperforms some of the other options, but does so by achieving a greater balance of the objective function's components. Specifically, compared to the next-best solution (weighted

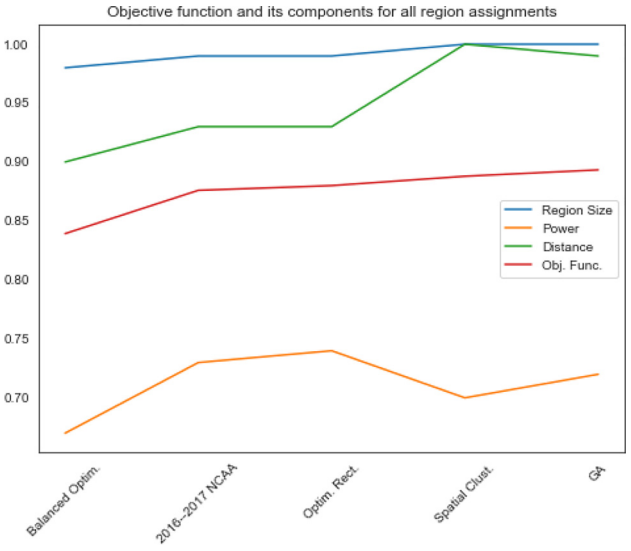

Figure 5: Visual summary of some of the data from Table 4.

spatial clustering), the genetic algorithm achieved a better balance of power at the expense of sightly more travel distance. This improved the value of the objective function because travel distance is near perfect optimality for both solutions, while power balance had more room for improvement.

Because it scored the best on the objective function, which equally weighted region strength, average distance traveled, and number of schools per region, this method

was the one we chose to report to the NCAA. It also had the advantage that it contained no steps that required expert intervention (such as tuning of parameters). Code implementing this method in Python in a Jupyter notebook can be found on GitHub (Carter 2020).

If a different balance among the components of the objective function were desired, it could be achieved using the same method, but introducing weights on each component. Such flexibility is not (easily) possible with the other algorithms we devised.

Due to the inherent subjectivity of scale in the objective function, we compared our results to those obtained through other approaches as well. For instance, we crafted an alternative objective function by sampling 2500 random region assignments and measured each component of the objective function on those, then used them as data to compute z-scores for $\frac{x-\mu}{\sigma}$ rescaling of the components before doing a simple sum to form the objective function. This method ended up assigning the power component too large of an influence, and yet the genetic algorithm approach still yielded the best results. We also tried replacing the geographic component of the objective function with one that measured not distance to the centroid of the region, but average distance among pairs of schools in the region. In that case as well, the genetic algorithm's results were those that maximized the objective function. These experiments provided confirmation that our choice was robust with respect to alternate formulations of the problem.

When we reported our initial results to the NCAA in September of 2018, we spoke with the Commissioner for DIII Wrestling and a coach on the DIII Wrestling committee. We found, to our surprise, that they had spent the intervening months working extensively among themselves to update the regional assignments to a new version for 2017. We therefore re-ran the genetic algorithm method on the latest data to create a comparable region assignment for 2017; it remained measurably superior to the NCAA's assignment, but took only a few minutes of computing time to achieve, rather than the numerous meetings and considerable time the NCAA wrestling committee.

The response was very positive, including confidence that coaches would be supportive of assigning teams to regions in this way. We met again in May 2019 with the DIII Wrestling Commissioner, who said that the committee is interested in using our genetic algorithm approach to calculate the 2020 regional assignments.

This study demonstrates that multiple analytical methods exist that can produce data-driven approaches to assign schools to regions that balance Power Ratings and the numbers of teams per region while minimizing travel distances. The techniques in Section 3 show how the orders of operations within cluster analysis can result in the relative prioritization of certain factors (Wagstaff et al. 2001).

Despite the limitations in the following section, some of the optimization approaches offer clear improvements over the 2016–2017 regional allocation system, and thus implementation would improve the fairness of DIII wrestling while simultaneously reducing travel costs. The automated and transparent nature of these algorithms reduces the influence of political considerations, and can be applied dynamically as teams are added or removed from DIII competition. These considerations increase the attractiveness of the flexibility of genetic algorithms.

This paper has thus also shown a real problem that demonstrates the value of genetic algorithms for clustering. It is certainly not the first such example; the value of genetic algorithms was first established in Cowgill et al. (1999) and the more recent survey paper of Hruschka et al. (2009a, 2009b) covers examples of their applicability in domains such as image processing, computer security, and bioinformatics. This paper adds to that list of domains by providing an example clustering problem from sports in which a genetic algorithm worked well.

5 Limitations

One limitation in our work is that actual travel time to the regional tournament is impacted by which school hosts the regional tournament, to which all teams travel. Our models did not take this into account, but used the proxy of total distance to the region's centroid. The robustness check mentioned in the previous section mitigates this limitation some.

A secondary limitation is that when the NCAA did region assignments, to which we compared our own results, they used a slightly different set of schools because they had more up-to-date information on which (three) schools were entering or departing the division, which happens every year. Consequently, comparing our region assignments to theirs is not a perfect apples-to-apples comparison even if we apply the same objective function, because the inputs are qualitatively different, even if only by just a few schools.

Finally, as mentioned in Section 1.4, our model for the power of a school's wrestling program includes some bias because it uses previous years' national tournament performance. That performance was measured at a time

when regional allocations were not done using the more objective standards of a computer model, and thus may reflect some schools' powers inaccurately or unfairly. This limitation may not be as significant as it first appears, however. Recall that the unfairness covered in Section 1.1 was that strong schools weren't able to send as many wrestlers to the national tournament as they might deserve. So on the one hand, a strong school has fewer wrestlers at the tournament, but on the other hand, they are competing against weaker wrestlers than they might otherwise have done. How much of the bias is mitigated by these competing issues remains unknown.

6 Future work

An open question about the power model introduced in Section 1.4 is whether it remains consistent from year to year for a school. While such an analysis is time-intensive to compute, given the data gathering work required, it would be interesting to see the results, because they could allay the concerns of coaches about which years should have their data included in the model.

The value of the work documented herein is not restricted to the sport of wrestling. Recall that the only inputs to our algorithms are a set of schools, their locations, and estimates of their teams' power, and such data could be obtained for many team sports. The same regional alignment problems occur in gymnastics, cross-county, track, skiing, bowling, and swimming. Many of these sports have in common with wrestling the fact that competition occurs at an individual level, but constraints require entire teams to be collocated during competition.

Regarding other types of future work, the optimization problem solved herein is only one of the many faced by coaches in NCAA DIII Wrestling. They have also communicated to us challenges surrounding optimal allocation of extra slots in each weight class at the national tournament, optimal seeding for the regional tournament, and optimal seeding for the national tournament. Such problems are particularly challenging because there is often very little or no data about head-to-head competitions among the individuals who need to be seeded or ranked.

Acknowledgement: Three anonymous reviewers provided suggestions that improved the clarity of this article, connected it better to the existing literature, and improved how we check the robustness of our methods.

Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Andreopoulos, B., A. An, X. Wang, and M. Schroeder. 2009. "A Roadmap of Clustering Algorithms: Finding a Match for a Biomedical Application." *Briefings in Bioinformatics* 10 (3): 297–314.

Bigsby, K., and J. Ohlmann. 2017. "Ranking and Prediction of Collegiate Wrestling." *Journal of Sports Analytics* 3 (1): 1–19.

Bliese, P. D. 2000. "Within-group Agreement, Non-independence, and Reliability: Implications for Data Aggregation and Analysis." In *Chapter in Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, 349–81. Jossey-Bass.

Bradley, P., U. Fayyad, and C. Reina. 1998. "Scaling Clustering Algorithms to Large Databases." In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*: 9–15.

Carter, N. C. 2020. Python Code Applying Genetic Clustering Algorithms to NCAA Division III Wrestling. Online Also available at https://github.com/nathancarter/clustering-for-ncaa.

Cowgill, M., R. Harvey, and L. Watson. 1999. "A Genetic Algorithm Approach to Cluster Analysis." *Computers and Mathematics with Applications* 37: 99–108.

Derringer, G., and R. Suich. 1980. "Simultaneous Optimization of Several Response Variables." *Journal of Quality Technology* 12 (4): 214–9.

Duda, R. O., and P. E. Hart. 1973. *Pattern classification and scene analysis*. New York: John Willey & Sons.

Estivill-Castro, V. E. 2002. "Why So Many Clustering Algorithms: A Position Paper." *SIGKDD Explor. Newsl.* 4 (1): 65–75.

Gan, G., C. Ma, and J. Wu. 2007. "Data Clustering: Theory, Algorithms, and Applications." In *Society for Industrial and Applied Mathematics*. Philadelphia, Pennsylvania: SIAM.

Hall, L. O., I. B. Ozyurt, and J. C. Bezdek. 1999. "Clustering with a Genetically Optimized Approach." *Trans. Evol. Comp* 3 (2): 103–12.

Han, J., M. Kamber, and J. Pei. 2012. *Data mining concepts and techniques*, 3rd ed. India: Elsevier Ltd.

Hruschka, E., R. Campello, and A. Freitas. 2009a. "A Survey of Evolutionary Algorithms for Clustering." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39 (2): 1133–155.

Hruschka, E. R., R. J. G. B. Campello, A. A. Freitas, and A. C. F. Ponce Leon. 2009b. "A Survey of Evolutionary Algorithms for Clustering." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39 (2): 133–55.

Malinen, M. I., and P. Fränti. 2014. "Balanced K-Means for Clustering". In *Structural, Syntactic, and Statistical Pattern Recognition*, edited by P. Fränti, G. Brown, M. Loog, F. Escolano, and M. Pelillo, 32–41. Berlin: Springer Berlin Heidelberg.

NCAA. 2015. *Regional Alignment and the Growth of Division III Wrestling*. Online Also available at https://www.d3wrestle.com/regional-alignment-and-the-growth-of-division-iii-wrestling/.

NCAA. 2020. *Division III Wrestling Website*. Online Also available at https://www.ncaa.com/sports/wrestling/d3.

Wagstaff, K., C. Cardie, S. Rogers, and S. Schrödl. 2001. "Constrained *K*-Means Clustering with Background Knowledge." *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning* 1: 577–84.

Xu, R., and D. Wunsch. 2008. *Clustering*: Wiley-IEEE Press.

Zhou, A., B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang. 2011. "Multiobjective Evolutionary Algorithms: A Survey of the State of the Art." *Swarm and Evolutionary Computation* 1 (1): 32–49.

Zhu, S., D. Wang, and T. Li. 2010. "Data Clustering with Size Constraints." *Knowledge-Based Systems* 23 (8): 883–9.