MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 3 - 4, 2017 HYNES CONVENTION CENTER

A Comparison of Clustering Approaches for Automating NCAA DIII Wrestling Regional Tournament Assignments

Other Sports
Paper ID 5625

1. Abstract

This research compares the results of four analytical approaches for partitioning the NCAA DIII Wrestling teams into regional tournaments. We compare the effectiveness of balanced optimization, weighted spatial clustering, weighted optimization rectangles, and genetic algorithm approaches for organizing teams into six regional tournaments. These approaches balance critical factors identified by NCAA DIII Wrestling coaches, including region strengths, region sizes, and travel distances. Our results provide evidence that each analytical approach presents a superior alternative to non-analytical strategies. However, among the four clustering analyses, we found the genetic algorithm approach best reflects the coaches' stated preferences for creating balanced regional tournaments. The implications of this research extend beyond wresting, and these findings can be applied to provide fair and competitive distributions of athletes for a variety of other sports in which athletes compete individually, but are assigned to regional tournaments as a team.

2. Introduction

Fairness is a central theme of NCAA sports. However, the creation of competitive regional tournaments is also constrained by factors including travel distances and expenses. Rankings and regional organization play a significant role in collegiate wrestling and affect the results of national tournament performance (Bigsby and Ohlmann, 2017). While receiving scant attention in comparison to other collegiate sports, wrestling contains several aspects that make it an interesting topic of research. These unique features provide opportunities to develop strategies for regionalization that may be applicable to several other sports.

In NCAA competition, wrestling teams are composed of ten wrestlers, 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. In each weight class, the two highest placing wrestlers at each of the six regional tournaments are invited to compete at the national tournament; two 'wildcard' competitors are also invited to the national tournament.

This work was motivated by the president of the NCAA DIII Wrestling Coaches' Association, who sought recommendations for developing fair, data-driven regional tournament assignments. We started by evaluating the NCAA's 2016-17 DIII regional assignments, which had unbalanced regional sizes and difficulties, resulting in dissatisfaction among coaches and wrestlers. Some regions had as few as 11 teams, while others contained as many as 21. Additionally, complications occur when perennially successful teams are co-located into the same regions. These features are exaggerated by an unbalanced competitive landscape among DIII wrestling teams. In the last 25 years, only two schools, Wartburg College (13 titles) and Augsburg College (12 titles) have won national titles. As a consequence of competitive imbalance, some of the best wrestlers compete in the same region and

do not qualify for the national tournament. When surveyed about the current allocation of teams to regions, coaches reported the system was unfair by a 2:1 margin. Thus, our primary research interest is the equitable distribution of a limited number of invitations to the national tournament through the six regional tournaments. Our research question is: can an analytics-based approach improve the fairness of these regions by balancing region difficulty, the numbers of teams per region, and travel?

We used a multi-step analytics approach to solve this regional assignment problem. First, we surveyed active DIII wrestling coaches to understand the critical factors for determining fair regions. We used these recommendations to analyze a dataset from the 2016-2017 NCAA DIII wrestling season and develop additive compositional variables for measuring team performance. In the model, a team's success at the national tournament can be predicted by critical factors, including: winning percentages, pins and technical falls, and the number of returning all-Americans.

This research was sponsored by the Teradata University Network[1], a free resource for learning and teaching analytics, which organized a collaborative project among academic researchers from five universities. Researchers working independently at these universities applied power ratings and geolocation data about each team to cluster analyses forming alternative regional alignments. By balancing the relative importance of region strength, region size, and travel distance, these analyses provide recommendations on how to improve overall fairness.

3. Developing Power Ratings

We first surveyed the head coaches of DIII wrestling programs. Of the 103 coaches who received the survey, 56 coaches responded. We solicited opinions on the existing regional assignment process, and asked which factors should be used to determine a team's performance. The average ratings of fairness and satisfaction were 3.26 and 3.62 respectively on a scale from 1 (low) to 7 (high), with satisfaction exhibiting a bimodal distribution. This implies that coaches largely agreed the system was unfair, but certain teams benefited while others suffered from inequity. Coaches reported the most salient factors causing unfair assignments were the imbalance in the number of teams per region, an imbalance in regional quality, and unfair travel distances.

The number of teams per region and geolocation data for travel distances are publicly available. However, the coaches' responses indicated the necessity of a power rating that could balance regional tournament difficulties. Through the survey and subsequent meetings, coaches identified 51 variables that could be used to determine a team's power rating. We then collected match data from 18,669 unique matches for the 2016-2017 season from Intermat.com, FloWrestling.com, and the NCAA DIII Wrestling websites to collect values for these variables for each school.

However, many of the variables were collinear; for example, the number of wins wrestlers have been awarded are highly correlated with the number of pins they have recorded. Through the factorization and consolidation of collinear variables, we developed a model that could predict a team's score at nationals. When reduced to its most parsimonious form, the premise of the model is that the power rating is calculated as a combination of last year's performance and this year's performance. The power rating (e.g., the expected performance at nationals) is a weighted combination of team points at last year's national tournament, the number of all-Americans at last year's national tournament,

[1] Teradata University Network (www.teradatauniversitynetwork.com) is a free resource for learning and teaching analytics. Every year, more than 5000 students and 12,000 people access the system for case studies, data sets, and homework assignments.

this year's winning percentage, and this year's "big loss" (i.e., pins and technical falls against) percentage. This equation contained compilation variables, which are used to develop multi-level models where team performance is derived from individual performances of players (Bliese, 2000). Our resulting equation contained six variables organized as follows:

$$Power\ Rating = (0.376 \times \#of\ All\ Americans\ prev\ year) + (0.151 \times nationals\ team\ score\ prev\ year) + \left(0.621 \times \frac{\sum current\ season\ wins}{\sum current\ season\ matches}\right) - \left(0.357 \times \frac{\sum current\ season\ pins\ against + \sum current\ season\ tech\ falls\ against}{\sum current\ season\ matches}\right)$$

We then tested our equation on historical data from the 2015-2016 season and found that it was possible to predict 67.7% of the variance in the number of points scored at nationals. 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). Finally as a test of robustness, we presented our predictive model back to a group of DIII wrestling coaches who confirmed that this model seemed anecdotally correct. We then calculated the teams' power ratings for the 2016-2017 season prior to regionals and used these ratings to balance regional difficulties in our alternative regionalization approaches.

4. Overview of Analyses

We used four different clustering methods to develop regional realignments, including: balanced optimization, weighted spatial clustering, weighted optimization squares, and genetic algorithmic clustering. Three of the approaches were variants of K-means clustering and the fourth approach used a genetic optimization algorithm. The four different algorithms reflected mathematically optimal solutions with variations of prioritization between balanced regional difficulties, balanced numbers of teams per region, and travel distances. These approaches primarily used the statistical software package R, while one approach used Analytic Solver. SPSS, SmartPLS, MS Excel, and Tableau were used for data cleaning, analysis, and visualization. Each solution takes about 15 minutes to converge upon an optimal solution. The specifics of the four approaches are provided below.

4.1. Balanced Optimization

We 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 developed a 6-region solution by randomly assigning teams to initial regions and calculating a centroid for each initial region. Then, the nearest centroid is calculated for each team using Euclidean distance, and that team is assigned to the closest centroid. To accommodate the business problem constraint of having a balanced number of teams per region, the number of teams assigned to each region was evaluated during each iteration of the algorithm.

Let $X = \{x_i\}, i = 1, .., n$ be the set of n d-dimensional observations to be clustered into a set of k regions, $C = \{c_k, k = 1, .., k\}$. The traditional k-means algorithm seeks a partition where the squared error among the mean of each region and the observations in the region is minimized. For each region, c_k, the squared error among the observations in c_k and their respective centroid mean is μ_k defined as $e(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$. This is iteratively minimized over all k regions as follows: $\min e(C) = min \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$. However, a traditional k-means algorithm must be extended to accommodate information about the number of teams that should be clustered together.

We extend k-means clustering by adding three tuning parameters LB, UB, and λ. LB represents the lower bound of the number of teams within a region, UB is the upper bound of teams within a region,

and λ is the percentage of teams in a region above the UB constraint and is used to reassign teams from overpopulated regions. If $C_i(x) = \sum_{i=1}^{n} \delta(x_i, c_i)$ is an indicator function counting how many teams x_i are currently assigned to region c_i, then at each iteration of the k-means algorithm we can identify for each region c_i, the number of teams $C_i(x)$ that do not fall within [LB, UB]. Thus, we randomly assign a percentage, λ of the teams within c_i to the next region that has $C_i(x)$<LB.

This algorithm iterates until the centroids converge and the number of teams within a region are within the [LB, UB] parameters. Lastly, because traditional k-means clusters can be negatively affected by the random start, we also ran our modified algorithm multiple times with varying random starting centroids and averaged the final converged centroids to obtain final regional assignments.

In this approach, geographic regions form that are close together and are more balanced in terms of number of terms per region. However, to satisfy the additional business constraint of having competitive-based regions as well, the last step in this balanced optimization approach was to reassign teams that could reasonably be assigned to another neighboring region to yield regions that are more competitively similar. This was done by identifying all possible schools that overlapped another geographic region, indicated by red circles as shown in Figure 1.

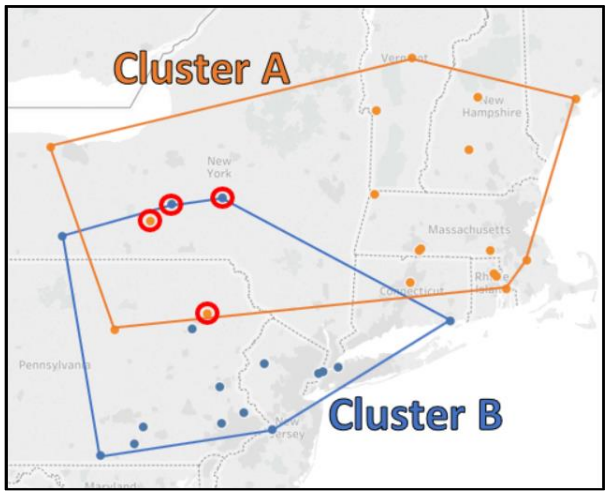

Figure 1. Example of Border Reassignment

This competitive balance optimization step is essentially an integer programming model, where the binary decision variables, $y_{ij} \in \{0,1\}$ assign team i to region j, with the objective to minimize the difference in average power rating by region. The function is constrained so that each region must still have between the LB and UB. This approach prioritized a balanced numbers of schools per region and travel distances, but had more inter-region variability in power.

Decision variables
y_{ij} ≔ assign candidate overlapping region team i to region j

Parameters
LB ≔ the minimum number of teams per region
UB ≔ the maximum number of teams per region
ρ_i ≔ power rating of team i
ρ_{ij} ≔ power rating of team i that is fixed to region j
f_j ≔ number of teams fixed to region j
$n_j = \sum_i y_{ij} + f_j$ ≔ total number of teams assigned per region j
N ≔ total number of regions j

Objective function & constraints

$$\min \sum_{j=1}^{N} \left[\sum_i (y_{ij}\rho_i + \rho_{ij})/n_j - \left(\sum_{j=1}^{N} \left(\sum_i (y_{ij}\rho_i + \rho_{ij})/n_j \right) /N \right) \right]^2 /N$$

Subject to:
$LB \leq n_j \leq UB \quad \forall\, j$

4.2. Weighted Spatial Clustering

Our second approach used weighted spatial clustering to partition regions. Traditional spatial clustering techniques are useful for segmenting data based on the similarity of an attribute. A weighted spatial clustering approach extends this type of analysis to address multiple constraints (Gan et al., 2007). Specifically, this approach allows a clustering algorithm to account for (1) teams per region, (2) minimized travel distances, and (3) balanced power between regions.

This approach set the region size and then randomly assigned teams to the initial regions. The Euclidean centers were calculated for each region using the Haversine formula to account for the curvature of the Earth. Then each team was reassigned to the closest region with an average power rating below the average power of all regions combined. For a 6-region solution, each region was limited to a score below the expected average region power of 36.5 (219 total power/6 regions).

During the team reassignments, the farthest school from the center of its assigned region was proposed for trade to a new region to balance the number of schools in each region. After proposing reassignments, the sum of the distances from each school to the center of their newly proposed regions were recalculated to ensure reassignments would only be implemented when they decreased the average distance to the center of the region. This process was continued until all schools had been assigned to a region and any additional iterations stopped reducing the average distance of schools to the center of their regions. By integrating the average power of the regions into the assignment algorithm, the weak and strong teams are distributed evenly among regions. This method prioritized equalizing the number of teams per region and balancing regional power, but produced more variance in travel distances.

4.3. Weighted Optimization Rectangles

In our third approach, we focused on minimizing the total size of a set of rectangles containing all of the schools in each region. The rectangles follow latitude and longitude lines and identify regions, and the rectangles were defined as the minimum bounding rectangles around schools in the region.

This solution differs from traditional clustering approaches because regions are not calculated on straight-line Euclidean distances, but instead schools are assigned to regions in a manner that minimizes the sum of rectangle areas for each region (Xu and Wunsch, 2005). By adjusting the objective function, this technique allows prioritization and balancing between factors that affect clustering. This method prioritized minimizing total travel distance, while constraining the difference in regional power and number of teams per region to acceptable levels.

4.4. Genetic Algorithm

A genetic algorithm approach offers some benefits over other cluster analyses when attempting to simultaneously resolve competing interests. Traditional cluster analytics cannot optimize a solution without implicitly prioritizing some constraints over others (Hruschka et al., 2009). In contrast, a genetic algorithm approach generates a number of imperfect solutions. Those solutions, presented as potential partition strategies, 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. The objective function used to partition teams into regions was as follows:

$$partition\ goodness = -\big(c_1(power\ variance) + c_2(size\ variance) + c_3(dispersion)\big)$$

The coefficients c_1, c_2, c_3 are chosen to make the range of each factor roughly equal.

- **Power variance:** Compute the mean power of each region in the partition, and then compute the variance of that set of means. Higher scores indicated an uneven balance of power.
- **Size variance:** A region's size is the number of schools it contains, and this is the variance of that set of sizes across a partition. Higher scores indicate a greater disparity of schools per region.
- **Dispersion:** The dispersion of a region is the sum of distances between each pair of schools in it; a measure of how much travel will tend to happen in that region. The dispersion of the partition is the sum of the dispersions of its regions. Higher values represent a greater geographic dispersion of teams within the region, and thus greater travel time and cost.

In this case, each solution was represented as a vector whose length was the same as the number of teams and whose entries were in the set {1,2,3,4,5,6}, thus assigning each team to a partition numbered 1 through 6. Breeding two solutions was done via uniform crossover (the child's i^{th} entry is either the mother's i^{th} entry or the father's i^{th} entry, equally likely). Other standard crossover techniques were investigated (e.g., uniform crossover and others involving optimal permutations) but all evolved the pool at about the same rate. A mutation in this case was a random change to any entry in a solution vector, and was performed randomly, and infrequently.

Over time the "gene pool" of partitions improves as they advantageously mutate or inherit beneficial features, and poor solutions are removed from the gene pool. After a sufficient number of generations, a near-optimal choice evolved for partitioning the schools into regions. For this analysis, 20 potential partitions were used for each round of breeding, and the process continued for 20,000 generations. By that time, evolution stopped yielding substantive changes in the objective function. The resultant solution, shown in Figure 2, was the best from the final gene pool. The genetic algorithm approach resulted in a balanced solution with a relatively equal prioritization of region strength, average distance traveled, and number of schools per region. This technique could also be adapted to prioritize other aspects by tweaking the values of c_1, c_2, c_3 in the formula above.

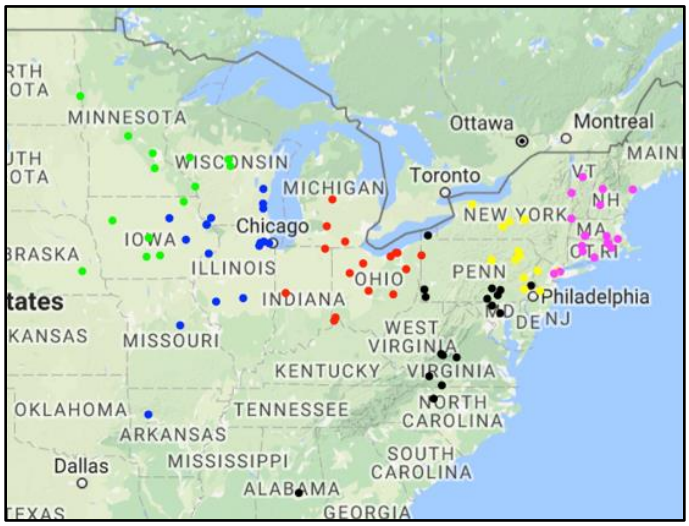

Figure 2. Genetic Algorithm Regional Assignments[2]

[2] Three schools are excluded from this figure because they fly to events and are located outside the range of the map.

5. Discussion

As displayed in Table 1, all the analytical approaches embody improvements when compared to previous regional assignment strategies, and all provided mathematically optimal solutions based on slightly different prioritizations of key factors (Wagstaff et al., 2001). The Balanced Optimization approach prioritized balancing the number of schools and minimizing travel distances, but had more variations in regional power. In contrast, the Weighted Spatial Clustering approach prioritized a balanced number of schools and regional power, while offering more variation in travel distances. The Weighted Optimization Rectangles approach prioritized balanced regional power and travel distances at the cost of more variation in the number of schools per region. When ranked by variance, the Genetic Algorithm approach is the most balanced optimization solution offering a relatively even weighting of regional difficulties, travel distances, and teams per region. This balance matches the preferences of coaches, who had indicated that an equal weighting of these three factors was ideal. Accordingly, we recommend implementing a regional realignment strategy that uses a genetic algorithmic clustering approach.

Table 1. Comparison of Clustering Results

	2016-2017 Regional Alignment			Balanced Optimization			Weighted Spatial Clustering			Weighted Optimization Rectangles			Genetic Algorithm		
Region	Schools	Rating	Dist.	Schools	Rating	Dist.	Schools	Rating	Dist.	Schools	Rating	Dist.	Schools	Rating	Dist.
West	11	1.6	148.9	18	2.7	148.6	17	2.1	183.8	16	2.8	154.7	18	1.7	131.5
Central	12	4.5	104.3	18	3.3	144.3	17	2.2	224.8	18	1.8	143.1	16	3.2	176.6
Midwest	19	2.3	190.8	16	2.1	189.2	17	2.1	76.0	14	3.1	135.5	16	2.7	140.0
Mideast	21	2.2	149.1	17	1.0	172.1	18	2.0	94.5	18	2.1	129.2	17	2.2	163.2
East	21	1.5	153.5	16	2.7	143.2	17	2.2	141.6	19	1.6	97.0	19	1.4	92.5
Northeast	19	1.5	197.3	18	1.1	187.9	17	2.2	173.4	18	1.7	95.9	17	1.8	99.8
Rank	5	5	4	2	4	1	1	1	5	4	2	2	3	3	3

Note: Distances are measured in miles

Our algorithms have some limitations. First, we used the geographic locations of schools to determine travel distances; however, average travel distances used to compare solutions were calculated to mathematical centers of the regions, which may not represent the precise locations of the hosts of the regional tournaments. Next, the relative power of teams in DIII wrestling is strongly skewed so that a few teams are extremely dominant and many of the other teams are largely interchangeable in regional assignment (i.e., teams that are not expected to have any wrestlers qualify for the national tournament). As a result, it may be useful to normalize the distribution of power scores to lessen the possibility that the most dominant schools are assigned a region of full of "tomato can" opponents to balance the regional difficulties. Future research may also consider using a power rating based on a multilevel variation of the Elo (1978) approach based on point differentials rather than wins and losses, which was adapted by Bigsby and Ohlmann (2017) for predicting the performance for DI wrestlers. Slight improvements to the power ratings may not provide substantial changes to regionalization recommendations, but could incrementally improve the precision of models.

Despite these limitations, each of the optimization approaches offers substantial improvements over the 2016-2017 regional allocation system, and we are confident that implementation would represent a substantial improvement to the fairness of DIII wrestling while simultaneously reducing

travel costs. Finally, 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 or other factors (e.g. conference membership) are added for consideration.

6. Conclusion

This study compares various clustering methods applied to the problem of NCAA DIII wrestling regionalization. It 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. Finally, this research has interesting theoretical implications about how the orders of operations within cluster analytics can result in the relative prioritization of certain factors (Wagstaff et al., 2006). These findings suggest that an order of precedence within priorities could be reflected in the strategic arrangement of background factors in clustering algorithms to produce solutions that reflect the relative prioritization of semantic rules.

All of the analytical approaches in this study offered superior solutions in terms of competitive equity and minimized travel costs when compared to historical implementations. In particular, we found that a genetic algorithm approach produced a solution most desired by the NCAA DIII coaches. Jon McGovern, President of the National Wrestling Coaches Association for NCAA DIII Wrestling, reflects these sentiments in the following statement, *"The NCAA Championships Committee responsible for arranging the new alignments in the years ahead (2019-2024) will have an opportunity to get feedback and new data based on the work of [this research team]. The team has already given the NCAA Championships Committee and the NWCA NCAA III Coaches body some very useful information and the hope is that this relationship will continue in the years ahead."*

In addition, the findings of this study can be generalized beyond wrestling to other individual-based sports where regional assignments occur at the team level. The same regional alignment problems occur in gymnastics, cross-county, track, skiing, bowling, and swimming. Many of these sports suffer from the same difficulties as wrestling where competition occurs at an individual level, but constraints require entire teams to be collocated during competition.

References

[1] Bigsby, K. G., & Ohlmann, J. W. (2017). Ranking and prediction of collegiate wrestling. *Journal of Sports Analytics*, *3*(1), 1-19.
[2] Bradley, P. S., Fayyad, U. M., & Reina, C. (1998). Scaling Clustering Algorithms to Large Databases. In *KDD* (pp. 9-15).
[3] Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.
[4] Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.
[5] Hruschka, E. R., Campello, R. J., & Freitas, A. A. (2009). A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(2), 133-155.
[6] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *ICML* (Vol. 1, pp. 577-584).
[7] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.