# **Balanced K-Means Algorithm with Equitable Distribution of Power Ratings**

Ananthapadmanabhan Sivasankaran, Muthuraja Palaniappan, Shubhansh Jain, Matthew A. Lanham

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

asivasa@purdue.edu; palaniam@purdue.edu; [jain297@purdue.edu](mailto:jain297@purdue.edu); [lanhamm@purdue.edu](mailto:lanhamm@purdue.edu)

## **ABSTRACT**

Traditional K-means clustering algorithm helps divide the data into clusters based on distances but does not consider the minimum and maximum number of observations in each cluster nor accounts for the “power ratings” of the observations. We overcome this problem by developing a modified K-Means algorithm where the minimum and maximum number of observations in each cluster and the “power ratings” of each observation are taken as constraints.

We have implemented a heuristic algorithm(Shunzhi Zhu 2010) to transform the size-constrained and clustering problem into Linear Programming approach and develop a modified K-Means. On top of this we have added power-ratings-constraints to make the algorithm solve this problem as well.

**Keywords:** R, Linear Programming, Balanced K-Means, Constrained Clustering, Data Mining

## **INTRODUCTION**

Sports Organizations such as NCAA face the problem of having a competitive regional level competition where there must be enough number of teams in each region. They would also like to reduce the travel time of the teams. Not all teams at the regional level are at the same level of competitiveness. Rankings and regional organization play a significant role in collegiate wrestling and affect the results of national tournament performance (Bigsby and Ohlmann, 2017). Hence, there should be balancing of not only the number of teams but also the “power rating” of the participating teams so that the better teams can win at regional level and the national level will be more competitive.

Division III Men’s wrestling faces the above problems. The way the Wrestling competitions work is the winning team from each of the 6 regions proceed to the national level. Other than that, 2 other wild card entries are also allowed. So, it becomes of paramount importance that the teams are of equal power ratings at the regional level and equal number of teams are present.

With our Balanced K-Means approach we develop an optimal group of observations that are not geospatially wide apart while also balancing the imbalanced observations and improving the current region assignments.

The President of the NCAA DIII Wrestling Coaches’ Association had sought ideas for developing fair regional tournament team clusters. Few regions had as few as 11 teams, while others contained as many as 21. Moreover, it is unfair for the perennially successful teams that are co-located in 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.

The research was sponsored by Teradata University Network1. Research was performed by 4 University Teams. Researchers at these universities applied power ratings and geolocation data about each team to cluster analyses forming alternative regional alignments. They evaluated balanced optimization, weighted spatial clustering, weighted optimization rectangles, and genetic algorithm approaches and finally decided upon Genetic Algorithm to solve this problem as it gave the best performance.

Models like Genetic Algorithms are not entirely explainable to the authorities regarding the reason for clustering. Our Balanced K-Means approach overcomes this by helping with interpretability while at the same time achieving the results that were attained by Genetic Algorithm approach.

Also, this approach will also be helpful for rolling out targeted marketing efforts in various organizations where there needs to be minimum and maximum number of customers with balanced power ratings (say, purchasing power) to be targeted in each customer segment. Due to such wide-ranging usage that is possible, this problem of Balanced K-Means with equitable distribution of power ratings becomes an important problem to solve. The same issues of interpretability are important in the field of Marketing Research as well and hence, our modified K-Means approach is more likely to be implemented vis-à-vis other approaches like Neural Networks or Genetic Algorithms to solve this problem.

The remainder of this paper is organized as follows: A review on the literature on various criteria and methods used for Balanced K-Means is presented in the next section. In Section 3 the proposed methodology is presented, and the criteria formulation is discussed. In Section 4 various models are formulated and tested. Section 5 outlines the performance of our models. Section 6 concludes the paper with a discussion of the implications of this study, future research directions, and concluding remarks.

## **LITERATURE REVIEW**

The steps of K-Means are as follows:

* Create K clusters by assigning each observation to the closest centroid
* Compute K new centroids by averaging Euclidean distance between observations in each cluster
* Continue above steps until the centroids don’t change

**K-Means Algorithm Extensions**

We found a few studies that had the same theme but different from our research. Bradley, Bennett, and Demiriz (2000) investigate adding constraints to K-Means to ensure each cluster will “*have at least a minimum number of points in it*.” Essentially, they show that incorporating a lower bound to the number of observations within each cluster will result in reducing the likelihood that the K-Means algorithm will identify poor local solutions – those with one or few points within a group.

Wagstaff, Cardie, Rogers, & Schroedl (2001) examine constrained K-Means clustering when additional background knowledge of the problem is available. Wagstaff *et al*. (2001) modify K-Means by incorporating “*background knowledge in the form of instance-level constraints*.”

Usami (2014) also recognizes the importance of efficient algorithms that result in output with good balance between clusters. In his study, he proposes a method with lower bound constraints on cluster proportions and a direct estimation of the number of unknown clusters. However, his method still requires improvement to handle clusters that do not fulfill cluster proportions and distance among clusters.

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.

Bhattacharya, Jaiswal, and Kumar (2015) explored constrained K-Means problems by proposing an algorithm that gives a tight upper and lowers bound on the list of candidate centers. Thus, they present an alternative that intends to improve the feasibility of providing better clusters through better center candidates. K-Means (Bradley et al., 2000)

C. T. Althoff, A. Ulges, A. Dengel (2000) tried using Frequency Sensitive Competitive Learning (FSCL) algorithm to solve the K-Means algorithm. While the traditional K-Means relies on Euclidean distance, this modified version tries to balance that weight with the number of points assigned to the cluster. The paper then deviates to combine this idea with hierarchical clustering.

This modified version is further explored in Mikko I. Malinen and Pasi Fränti (2014). They tried using Hungarian algorithm to solve the assignment problem of balanced K-Means clustering algorithm. By doing so, the time complexity got reduced to O(n^3) when compared to linear programming in constrained K-means algorithm.

Shunzhi Zhu, Dingding Wang, Tao Li (2010) have proposed a heuristic algorithm to transform size constrained clustering problems into integer linear programming problems. However, this approach does not deal with Power Ratings.

Chen, Zhang and Ji (2005) have proposed an algorithm to minimize the size regularized inter-cluster similarity (this is equivalent to maximizing the size regularized intra-cluster similarity). The size regularized cut overcomes the drawback of average cut and the normalized cut that are sensitive to outliers due to the multiplicative nature of their cost functions.

Data-mining problems have demands that require balanced clusters with approximately same size or importance (Banerjee, 2006) and it is important to create K-Means variants that allow such control to increase the reliability of the clusters and its relevance to the problem.

The steps of our balanced K-Means algorithm are as follows:

* Solve K-Means as per the usual approach
* Use these as initial centroids and check for minimum and maximum number of observations in each cluster using Linear Programming approach proposed by Shunzhi Zhu, et al., 2010
* Optimize for the power ratings of the observations using this as an additional constraint in the Linear Programming

To illustrate the main contribution of our algorithm, we compare different balanced K-Means algorithms from the literature in Table 1. Based on this comparison, we want to emphasize that our method intends to be a unique solution to clustering problems, since it is meant for applications where additional external constraints are known (not only minimum and a maximum number of observations in each cluster but also the “power ratings” of each observation). By taking advantage of this additional information, our algorithm is more likely to produce a satisfactory solution.

Moreover, unlike most of the other papers (barring Shunzhi Zhu, et al.,) we have used Linear Programming approach to solve K-Means thereby making it an easily interpretable and low math complexity problem.

Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Balanced K-Means method | Easily implemented | Low math complexity | Cluster size controlling | Robustness to initialization | Applicable to large dataset |
| *Multicenter clustering*  (Liang, Bai, Dang, & Cao, 2012) | ● |  | ● | ● |  |
| *MinMax K-Means*  (Tzortzis & Likas, 2014) | ● |  | ● | ● | ● |
| *Min-Cut Clustering*  (Chang, Nie, Ma, & Yang, 2014) | ● | ● |  |  | ● |
| *Weight point sets*  (Borgwardt, Brieden, & Gritzmann, 2016) |  |  | ● | ● |  |
| *Background knowledge*  (Wagstaff et al., 2001) | ● | ● | ● |  |  |
| Undersampled  (Kumar, Rao, Govardhan, Reddy, & Mahmood, 2014) | ● | ● |  |  | ● |
| FSCL (C. T. Althoff, A. Ulges, A. Dengel, 2000) | ● |  | ● | ● |  |
| Balanced K-Means with Hungarian algorithm (Mikko I. Malinen and Pasi Fränti, 2014) |  |  | ● | ● |  |
| Heuristic with Linear Programming (Shunzhi Zhu, Dingding Wang, Tao Li, 2010) | ● | ● | ● | ● |  |
| Size-regularized inter-cluster similarity (Chen, Zhang and Ji, 2005) |  |  | ● | ● |  |
| Balanced K-Means with equitable distribution of power ratings  (method presented in this paper) |  | ● | ● | ● |  |

**DATA**

The dataset is publicly available at NCAA website. The dataset consists of the following data:

Table 1: Data used in study

|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Longitude | Numeric | Longitude of the wrestling team |
| Latitude | Numeric | Latitude of the wrestling team |
| Power Rating | Numeric | Power Rating (similar to Elo rating) of the wrestling team |

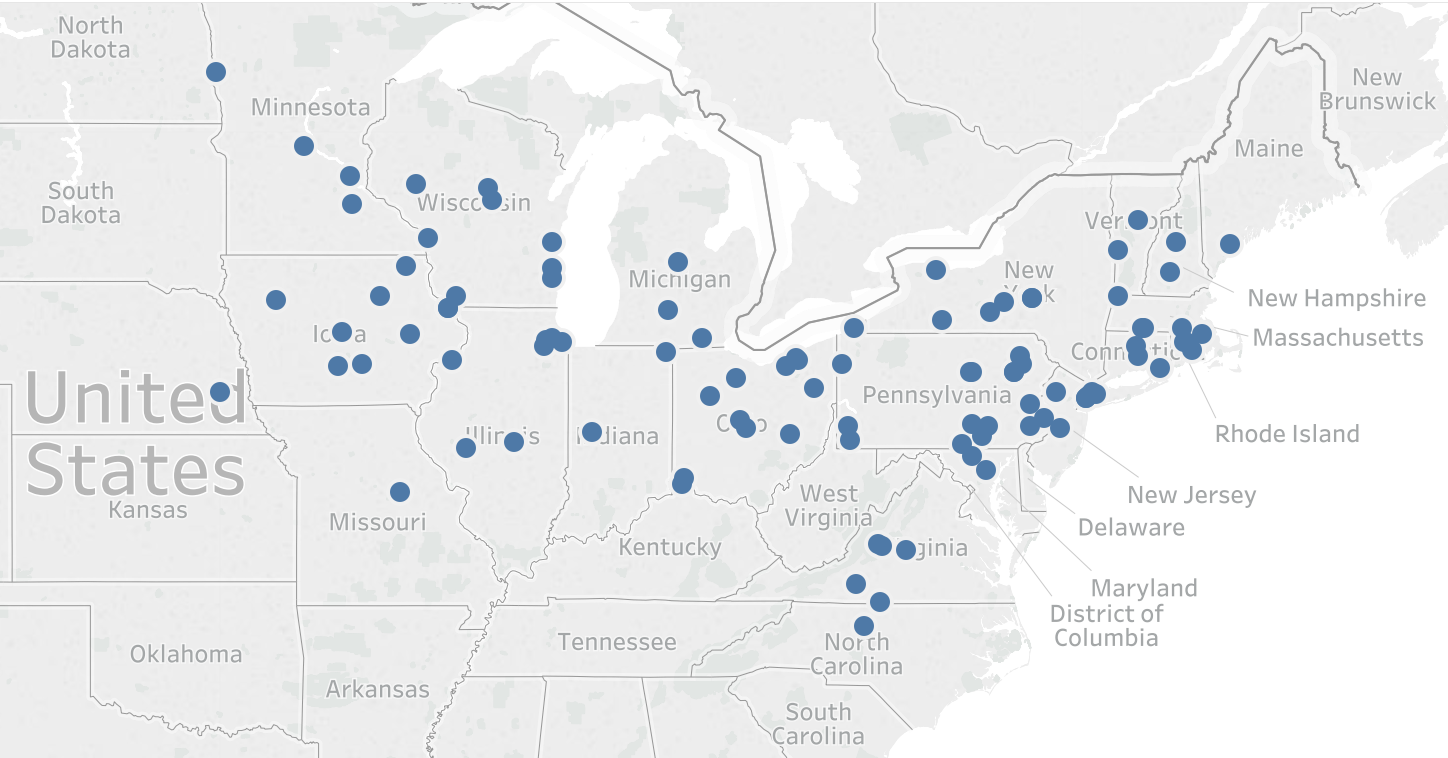


Figure 1: The school locations involved in this study

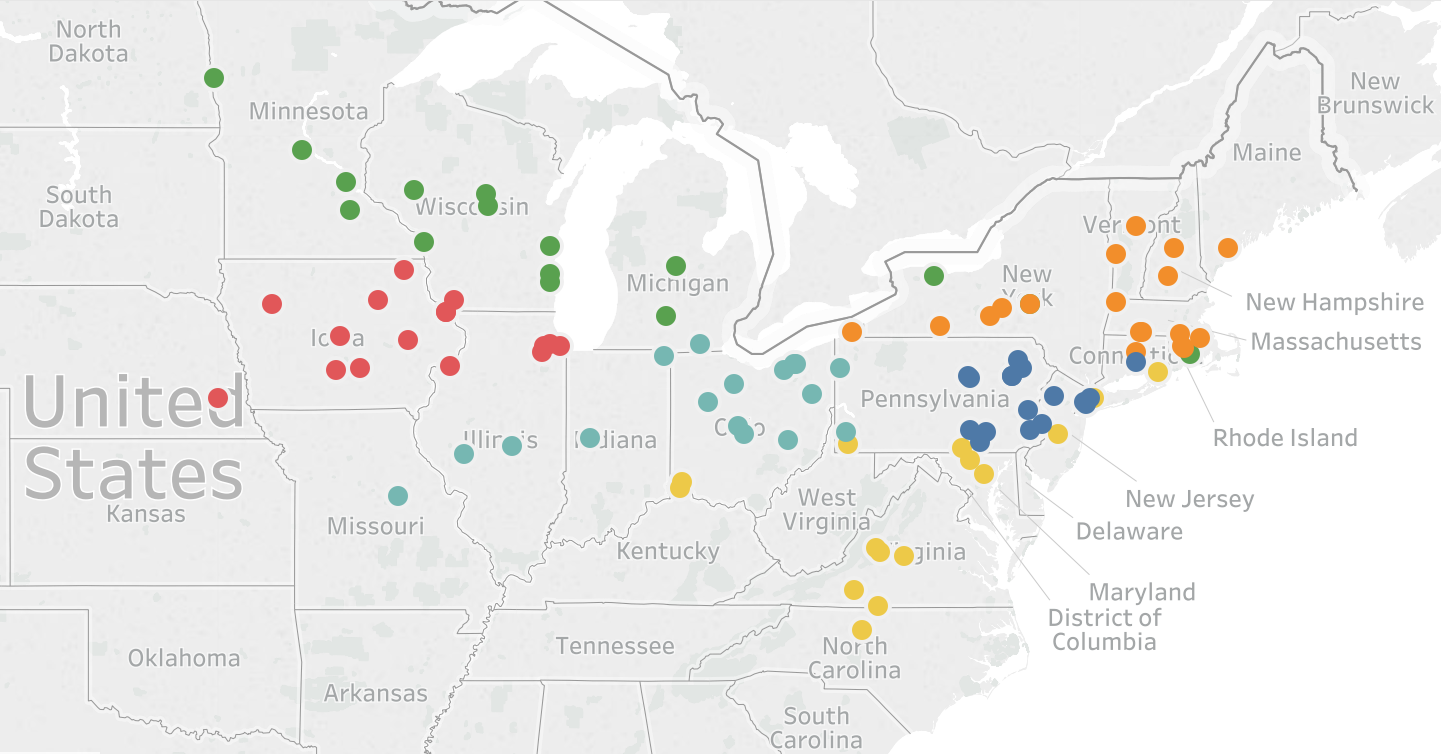


Figure 2: Clusters generated by the standard K-means algorithm

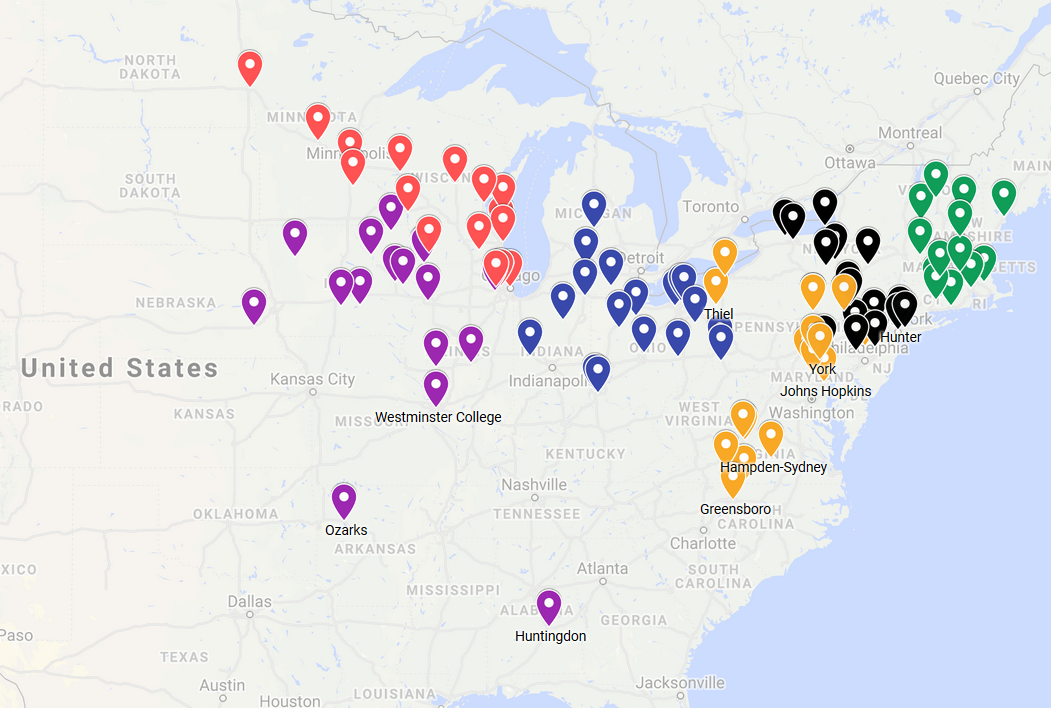


Figure 3: Current clustering as per 2017-18 assignments

## **METHODOLOGY**

We have implemented Linear Programming approach to optimize the constrains of maximum and minimum number of elements in each cluster as well as equitable distribution of power ratings in each cluster.

**Optimization Parameters:**

The user must specify the minimum number of elements in each cluster and maximum number of elements in each cluster. If minimum is set too high, the model will not be able to converge. Similarly, if the maximum is set too low, the model will fail to converge.

where µ and are the average and standard deviation values of the power rating (target variable) and d is the distance matrix between points and center of each cluster. b is the matrix of optimal cluster allocation that we want to find. p is the power rating (target variable) that we would like to maintain near the mean and delta is an user defined value for tolerance in p.

When we have a low delta value the model tries to force-fit elements in such a way that power ratings are closer to mean of all the power ratings. This is a very restrictive condition. This leads to the points being widely dispersed in terms of distance.

As we relax this condition by increasing delta, the model fits the elements in a more natural way. For balanced K-Means, we see that the points are not dispersed too wide geospatially. Once the convergence is realized, we can fix the delta.

## **MODEL**

To achieve our goal, we have used custom defined K-Means clustering. As mentioned before, this model involves using K-Means clustering to first arrive at a solution set for clusters and then optimizing this cluster allocation to arrive at the most optimal result.

Our motivation to arrive at this methodology was mainly motivated by the paper published in ‘Data Clustering with Size Constraints’(Shunzhi Zhu 2010), where the heuristics were used to arrive at a similar yet balanced cluster from the preexisting allocation. We realized that instead of simple allocation, distance matrix would be a better suitable candidate for optimization. Also, since all the constraints are integral coefficients, the solution must also have integral coefficient as solution.

Regarding the constraints on the number of elements in each cluster, according to the WagStaff et al [7], imposing a minimum constraint in the number of elements is enough in most of the cases as balancing the clusters automatically gives clusters around the optimal size. But, for generalization, we have included both the boundary constraints in our models.

Like ridge and lasso extensions of OLS, we proposed a penalty constraint for deviating from the mean of target variable. This factor used as delta is user defined and would determine how important is it for the clusters to have mean target variable values. Having a low delta would force the clusters to be very haphazard at the cost of achieving close mean values of target variable. Having a large delta would defeat the purpose of having the target variable in the first place. Typically, a value of 0.5 to 2 is recommended.

For the linear optimization part, we used the predefined library ‘lpsolve’ from R. The distance calculation among the points were Euclidean and ‘pdist’ library was used to implement the same.

## **RESULTS**

After implementing our algorithm, we obtained the following map organization in Figure 5. Clusters produced by our Balanced K-Means algorithm satisfy the problem requirements (whereas the standard K-Means algorithm does not).

The Figure 5 clearly shows that the clusters that we found are much better than the clusters that are currently assigned by NCAA. This will increase the competitiveness of the sport and increase fan following at the national level. Also, the fairness of the sport will be restored as the strong teams need not face each other at the regional level and miss out on reaching national level.

This new organization of schools is aligned with the NCAA’s expectations in terms of distance between schools and average competitiveness. Both these constraints are taken care of in our clusters. Thus, we verified that our modified version of the K-Means algorithm can be implemented successfully in problems within this domain.

The decision support for Market Segmentation where there needs to be balancing of the number of customers in each segment as well as their power ratings, this algorithm can be used.

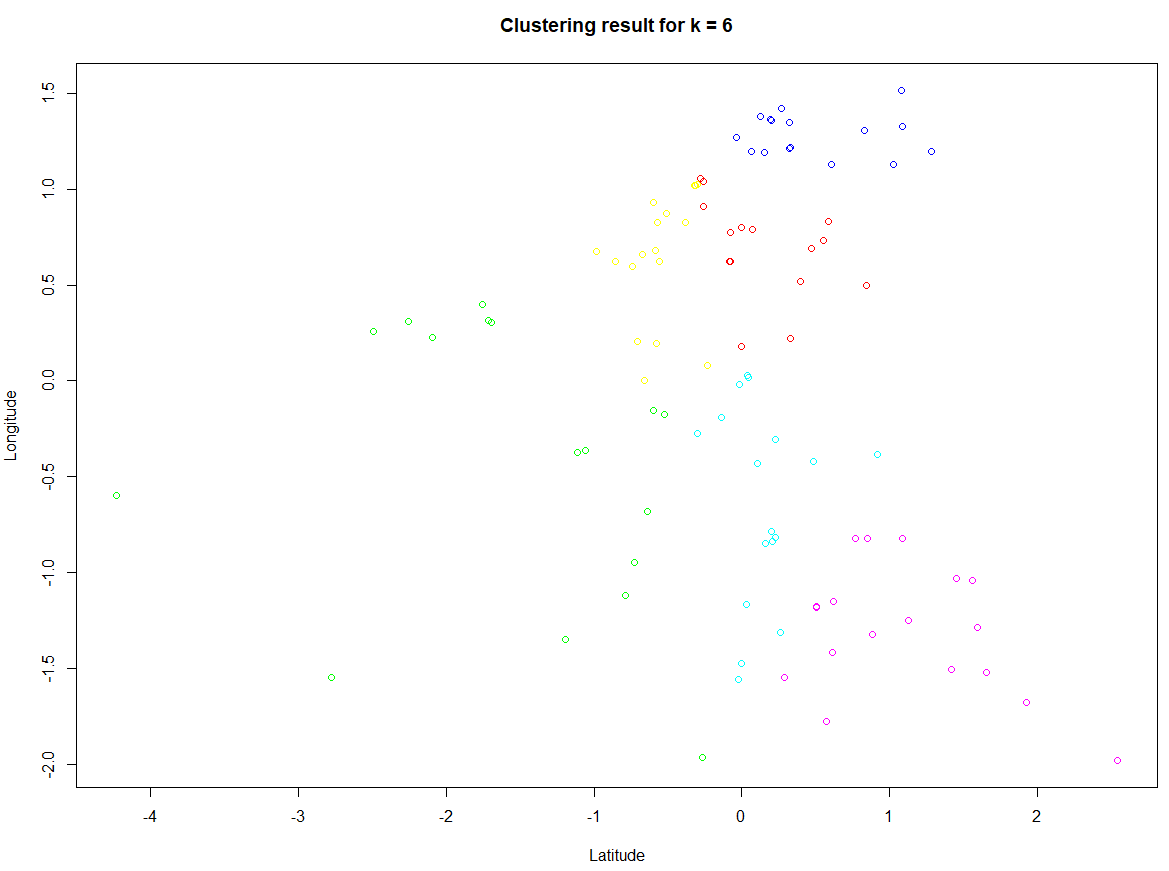


Figure 5: Clusters produced have observations clubbed together in terms of distance and equitably distributed power ratings

## **CONCLUSIONS**

Division III Men’s wrestling team assignments are imbalanced leading to competitions being skewed against the stronger teams and the matches being uncompetitive at national level. This calls for a better clustering approach that not only looks at approximately equal number of teams but also the equitable distribution of power ratings of the teams within each cluster.

The conversion of K-Means to Linear Programming approach with minimum and maximum number of observations in each cluster as constraints with additional constraints for balancing the power ratings solves the problem of balancing the clusters.

This simplified approach is beneficial to the business as it is easily understood and also clusters the teams appropriately leading to healthy competition at regional and national levels. This could also find applications for Market Segmentation where there needs to be a balance between the clusters.

We have made generalized the algorithm for any number of columns and assumed that the number of features is atleast 2.

Since it is a Linear Programming approach, as the number of features increases, the time taken to solve the problem increases. This is the limitation in our study.

Ways of increasing the speed of the algorithm as the number of features increases could be looked upon in the future.

Alternatively, instead of Linear Programming approach, we could explore ways to implement Hungarian Algorithm that Mikko I. Malinen, et al., proposed and include power ratings constraint in that. We could see which method gives faster and accurate results.

## **ACKNOWLEDGEMENTS**

The authors would like to thank Division III men’s wrestling, Dr. Dave Schrader, and Teradata for providing us an opportunity to work on this exciting conference realignment problem.

## **REFERENCES**

1. Shunzhi Zhu , D. W., Tao Li. 2010. Data clustering with size constraints. 7. ELSEVIER: ScienceDirect.

2. Bigsby, K. G., & Ohlmann, J. W. (2017). Ranking and prediction of collegiate wrestling. *Journal of Sports Analytics*, *3*(1), 1-19.

3. Liang, J., Bai, L., Dang, C., & Cao, F. (2012). The Kmeans-type algorithms versus imbalanced data distributions. IEEE Transactions on Fuzzy Systems, 20(4), 728–745. <https://doi.org/10.1109/TFUZZ.2011.2182354>

4. Tzortzis, G., & Likas, A. (2014). The MinMax Kmeans clustering algorithm. Pattern Recognition, 47(7), 2505–2516. <https://doi.org/10.1016/j.patcog.2014.01.015>

5. Chang, X., Nie, F., Ma, Z., & Yang, Y. (2014). Balanced Kmeans and min-cut clustering.

Eilermann, M., Post, C., Schwarz, D., Leufke, S., Schembecker, G., & Bramsiepe, C. (2017).

6. Borgwardt, S., Brieden, A., & Gritzmann, P. (2016). A balanced Kmeans algorithm for weighted point sets.

7. Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained Kmeans clustering with background knowledge. International Conference on Machine Learning, 577–584. <https://doi.org/10.1109/TPAMI.2002.1017616>

8. Kumar, N. S., Rao, K. N., Govardhan, A., Reddy, K. S., & Mahmood, A. M. (2014). Undersampled Kmeans approach for handling imbalanced distributed data. Progress in Artificial Intelligence, 3(1), 29–38. <https://doi.org/10.1007/s13748-014-0045-6>

9. C. T. Althoff, A. Ulges, A. Dengel, “Balanced clustering for content-based image browsing”, GI-Informatiktage 2011. Gesellschaft f¨ur Informatik e.V., March 2011.

10. Mikko I. Malinen and Pasi Fränti, “Balanced K-Means for Clustering”, Volume 8621 of the series Lecture Notes in Computer Science pp 32-41 and Proceedings of the Joint IAPR International Workshop, S+SSPR 2014, Joensuu, Finland, 2014.

11. Y. Chen, Y. Zhang, X. Ji, “Size regularized cut for data clustering”. Advances in Neural Information Processing Systems, 2005.