[[1]](#footnote-1)

Wineinformatics: Wine Region Determination Using K-Nearest Neighbor and Various Distance Metrics

Hunter Moore, Nnaemeka Newman Okereafor, Jonathan Onyumbe, Alex Riley

Department of Computer Science of Central Arkansas Conway, AR, 72034 USA

***Abstract*—Wineinformatics is an emerging domain in the field of data science. Wine production is multivariate, in that the final quality of the wine is determined by a multitude of factors. One of the fundamental determinants for the quality of a given wine is its “terroir” -- the characteristics of the soil where the wine’s ingredients are cultivated. Based on previous research, we know that the descriptive qualities for a given wine can be represented as a binary tuple, e.g. “bubbly = 1, high\_tannin = 0, strong\_finish = 1” through the Computational Wine Wheel. We have also determined that the region that a wine originates from -- Bordeaux, France or Napa, California -- can be predicted by combining distance metrics with K-Nearest-Neighbor (KNN). This paper surveys the classification accuracy of a wine dataset for various distance metrics with KNN. These results are cross-validated using the leave-one-out method for different values of k.**

***Keywords -- Wineinformatics, Data Mining, Napa, Bordeaux, KNN***

# Introduction

Humans have been producing and consuming wine for thousands of years. One of the most exciting qualities of wine is the diverse range of flavors and textures that can be produced. Over the years, wineries have experimented with various aspects of the winemaking process, such as the ingredients, fermentation time, the aging vessel, and temperature regulation. One of the most crucial elements of wine production is *terroir* -- the synthesis of environmental factors such as soil, climate, and terrain that affects the quality of the wine’s ingredients. If several wines from the same region are examined, they will likely exhibit some similar qualities. Two of the most prominent wine regions are Bordeaux, France, and Napa, California. Bordeaux represents the “Old World” of wine production -- wines produced in Europe or the Middle East. Generally, wines from this region are characterized by a narrow range of flavors, low alcohol content, and high acidity. The traditional winemaking techniques perfected in the region have been passed down from generation to generation under a veil of secrecy. The “New World” representative is Napa, California. Unlike the wines of the Old World, New World wines are characterized by high alcohol content, fruity flavors, and less acidity [1**]**. With this knowledge, the region for an unidentified wine can be determined by simply examining its descriptive qualities.

# Wine data

Wineinformatics is a subset of data mining involving the extraction of useful information from wine datasets. In previous work, the “Computational Wine Wheel” was developed to automatically process wine reviews and extract useful descriptive qualities [2]. This tool was used to process thousands of online reviews from the lifestyle magazine *Wine Spectator*, which has been publishing wine reviews since 1976. *Wine Spectator* provides a free and easily searchable online database of over 388000 wines and processes over 16000 wines per year. All wine tastings are conducted in private, controlled conditions. The identity of each of the wines is obfuscated. Only the vintage and general type of wine are revealed to the reviewer before tasting [3]. This method of blind tasting removes external factors such as cost or bias that could negatively impact the quality of the reviews. Below is an example of a typical wine review from *Wine Spectator*.

***Kosta Browne Pinot Noir Sonoma Coast 2009 95pts***

*Ripe and deeply flavored, concentrated and well-structured, this full-bodied red offers a complex mix of black cherry, wild berry and raspberry fruit that's pure and persistent, ending with a pebbly note and firm tannins.*

Wine reviews follow a standard 100-point scale:

**95-100** *Classic: a great wine*

**90-94** *Outstanding: a wine of superior character and style*

**85-89** *Very good: a wine with special qualities*

**80-84** *Good: a solid, well-made wine*

**75-79** *Mediocre: a drinkable wine that may have minor flaws*

**50-74** *Not recommended*

The most common components of each review are the sensory analysis of the flavors, acidity, tannins, weight, and finish of the wine. While there are tens of thousands of possible descriptive qualities that could exist within a single review, commonalities emerge when examining reviews side-by-side. The presence/absence of a given quality can be expressed with binary states after processing the input wine review with the Computational Wine Wheel. For the above sample wine, the processed review data will be tabulated as: “berry = 0, raspberry = 1, wild berry = 1, tannins high = 1, tannins low = 0, beauty = 0, …” Note that the Computational Wine Wheel considers the semantic and syntactic analysis of the input text data when making determinations of an input wines qualities. In this example, “wild berry” and “raspberry” flavors are present, but “berry” is not.

As established in previous work [2,4], the *k-nearest-neighbor* (KNN) algorithm can be used in combination with the Jaccard similarity coefficient to accurately determine the region (Napa/Bordeaux) for a particular wine. To improve upon the classification accuracy, a survey of various similarity indices was conducted. Jaccard-Needham, Kulzinsky, Dice, Correlation, Yule, Russel-Rao, Sokal-Michener, and Rogers-Tanmoto were evaluated. These distance metrics can be easily applied to the binary wine dataset. The results were cross-validated using the leave-one-out method. Our wine dataset contained 647 total wines -- 296 from Bordeaux and 351 from Napa. Our dataset contains only *classic* wines, i.e. wines with a score of 95+ from *Wine Spectator*. These classic wines are more likely to embody the known characteristics of Napa/Bordeaux wines than experimental wines that scored poorly. During preprocessing of the wine dataset, four Napa wines were discovered containing only zeros, meaning they did not find a match for any of the qualities defined in the Computational Wine Wheel. These tuples were discarded.

# Methodology: Overview of Distance Metrics

Figure III.1 defines the distance metrics surveyed. While these formulas have matching dissimilarity indices [5], only the similarity index formulas were evaluated. The term **S11** represents intersections between observations in the wine dataset -- instances where wine A and Wine B share a particular quality. The terms **S01** and **S10**represent instances where a particular quality is present in one wine and absent in the other. The term **S00**represents instances where a particular quality is not found in either wine. **N** represents the total number of observations. It is worth noting that there are many similarities between the distance metrics surveyed. The Jaccard-Needham, Kulzinsky, and Dice formulas are very similar and will produce the same accuracy for the wine dataset. Russel-Rao and Sokal-Michener are also very similar, with the latter formula simply considering S00 in the numerator.

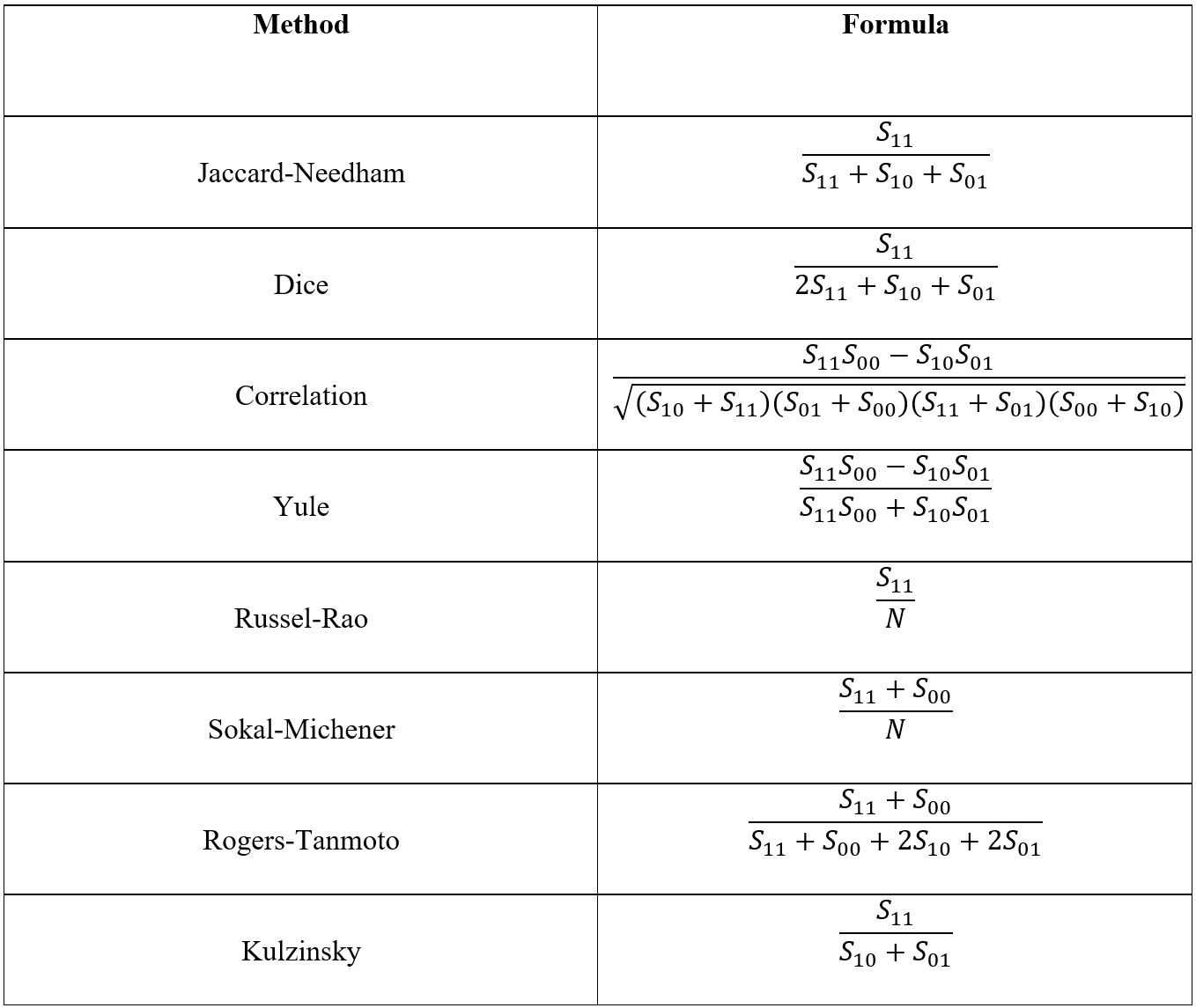


Fig III.1: Distance metrics definition.

# Results

Using accuracy as the metric for determining the quality of the various distance metrics, Figure IV compares these accuracies for various values of K.

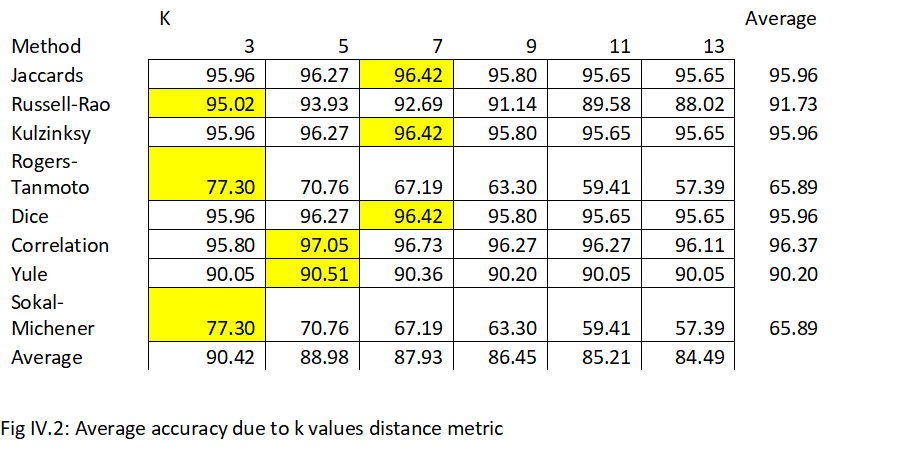


Figure IV.1 depicts the various distance metrics, K values, and two different average values. The Average column contains the mean accuracy for all values of K for each corresponding distance metric. The Average row shows mean accuracy for all distance metrics for each corresponding value of K. The overall trend shows the Average row decreases as K increases.

Prediction accuracy based Jaccard’s distance metric, Kulzinksy’s distance metric and Dice’s distance metric increases as the number of K increases. However, there is a sharp decrease in accuracy as the value of K increases beyond 7. The Rogers-Tanmoto’s and Sokal-Michener metrics gives decreasing accuracy value as the value of K decreases across board. Prediction accuracy decreases with increasing values of K when using the Yule and Correlation distance metrics.

Jaccard’s, Dice’s, and Kulzinksy’s metrics have an average accuracy of 95.96 percent across the values of K, while the Yule and Russell-Rao distance metrics maintained fair average accuracy values of 90.20 percent and 91.73 percent respectively across different values of K.

Correlation distance metric gave the highest accuracy value of 97.05 percent when K is 5 and has an average accuracy value of 96.37 percent across the different values of K. The Sokal-Michener and Rogers-Tanmoto distance metrics gave the worst accuracy values when the value of K is 13.

The majority of the distance metrics, with the exclusion of the Correlation metric, achieved their highest respective accuracy values when the value of K is 7.

Finding that the Correlation metric provides the highest accuracy overall raises some interesting questions. The formula takes attributes into account that are not present in the review of either wine being compared. This suggests that words not used to describe a wine may be important as well. An example of this might be comparing two wines where both are *not* described with a certain word e.g. “earthy”. This lack of “earthy” could signify that these wines are more similar than just their matches and non-matches would suggest.

# Conclusion and future works

Various distance metrics can be used in conjunction with KNN to predict whether a wine might originate from Bordeaux or Napa. The Correlation metric proved to give the best results in terms of accuracy and raised questions regarding the lack of attributes when comparing two wines.

Possible future works related to wine regions are the application of various classifiers rather than just KNN. These could include K-Means clustering, Naive-Bayes, SVM, Logistic Regression, and many others. Another consideration could be simply adding more distance metrics. Correlation outperformed Jaccard but there could still exist a distance metric that provides even greater accuracy. This research could potentially be applied to other areas of study such as soda, cereal, tea etc. There are many options to further the exciting field of wine region identification.

References

[1] Teeter, A. (2019). *Explaining Old World Wines Versus New World Wines*. VinePair. Available at:

<https://vinepair.com/wine-blog/understanding-difference-old-new-world-wines>.

[2] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel," *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 142-149

[3] Wine Spectator Home | Wine Spectator. Available at: <https://www.winespectator.com/>

[4] B. Chen, C. Rhodes, D. Fink, and T. Atkison*, "Wineinformatics: Wine Regions Determined by K-Nearest Neighbor," SWDSI 2017 Annual Conference,* Little Rock, AR, March 2017

[5] B. Zhang, S.N. Srihari, "Properties of Binary Vector Dissimilarity Measures", *Proc. JCIS Int'l Conf. Computer Vision Pattern Recognition and Image Processing*, Sept. 2003.

[6 ]Mehmed Kantardzic. “Data mining: concepts, models, methods, and algorithms.” *John Wiley & Sons* Aug. 8, 2011.

[7] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “From data mining to knowledge discovery in databases.” *AI Magazine* Vol. 17, No. 3 (1996)

[8]http://winearomawheel.com/Websites/aromawheel/Images/use rguide\_2010.pdf

Further Reading

E. Nasibov, C. Kandemir-Cavas. ”Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction.” *Computational Biology and Chemistry* Vol. 33, Issue 6 (2009): 461-464

Urtubia, Alejandra, J. Ricardo Perez-Correa, Alvaro Soto, and Philippo Pszczokowski. “Using data mining techniques to predict industrial wine problem fermentations.” *Food Control* 18, no. 12 (2007): 1512-1517.

C. Paulo, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. “Modeling wine preferences by data mining from physicochemical properties.” *Decision Support Systems* 47, no, 4 (2009): 547-553.

Ishibuchi Hisao, Tomoharu Nakashima, and Manabu Nii. “Classification and modeling with linguistic information granules; advanced approaches to linguistic data mining.” *Springer*, 2005.

1. [↑](#footnote-ref-1)