Binary Indices at Various Densities

Skylar Turner OMSII, and
Joseph A. Price III, Ph.D., Dept. of Pathology, OSU-COM.

ABSTRACT

Binary similarity indices are numerical analysis methods used to compare data involving two binary vectors (lists). The scope of this project involved comparing 54 binary similarity indices methods in relationship to binary vector density using the R programming language. Matrices were created of various vector data. The matrices were then scrambled to represent random data. Finally, the data was analyzed and plotted. Vector density variation can result in large differences – in both rate of change relative to density and magnitude. Awareness of these differences is important when selecting an analysis method and understanding the effects of changing vector density on analysis of results.

INTRODUCTION

Binary data consist of two possibilities: presence and absence. A 1 represents a presence of a descriptor, while a 0 represents an absence in a binary matrix. An example would be the presence of fur on mammals. A 1 value would represent an animal with fur, while a 0 would not have fur. Vector density is the number of 1 results in comparison to 0 results for a given column. For example, a matrix column only consisting of 1s will have a vector density of 100%, while a matrix column with only 0s will results in a vector density of 0%. Comparing large data sets of true and false data is a common statistical problem, and thus numerous methods have been developed to address this issue. The method explored involve 4 factors : a, b, c, d.

a : represents data which has true values in both sets.
b : represents data where a true value is present in the first vector, but not present in the second.
c : represents data where a true value is present in the second vector, but not the first
d : represents data which has false values in both sets.

Many indices have been developed using a,b,c, and d factors to compare lists of binary data, and the properties of (dis)similarity indices have been examined repeatedly (reviewed (1)). But the exact effect of vector data density, as seen with a high incidence of d (non-matches), on the resulting similarity coefficients has not been made precisely clear. Wolda (2) examined the effect of vector length on indices, and suggested composition may have a minor effect. Lewis (3) used several vector densities and showed qualitatively that density can affect coefficients. No systematic study has previously been made.

In order to develop a better understanding of similarity results, the effects of vector density on indices was explored.

Materials and Methods

R programming language was used due to availability of 3rd party packages tailored for similarity indices and lattice plotting. The Simba R package was selected for data processing due to the wide range of available methods to conduct binary indices comparison. First, R was used to generate a series of matrices of 100 columns of "objects" and 1000 rows of "descriptors". In order to generate the matrix, columns were generated with a known vector density, and then the positions scrambled. This ensures random data of a known vector density. **Table 1** shows an example matrix with 4 rows and 5 columns of 50% vector density.

	object 1	object 2	object 3	object 4	object 5	d
Descriptor 1	1	1	0	0	1	812
Descriptor 2	0	1	1	0	0	807
Descriptor 3	0	0	0	1	1	809
Descriptor 4	1	0	1	1	0	808

Table 1 : Vector density example. Each Object represents a vector with 4 descriptors. 2 descriptors have a value of 1 which results in a %50 vector density.

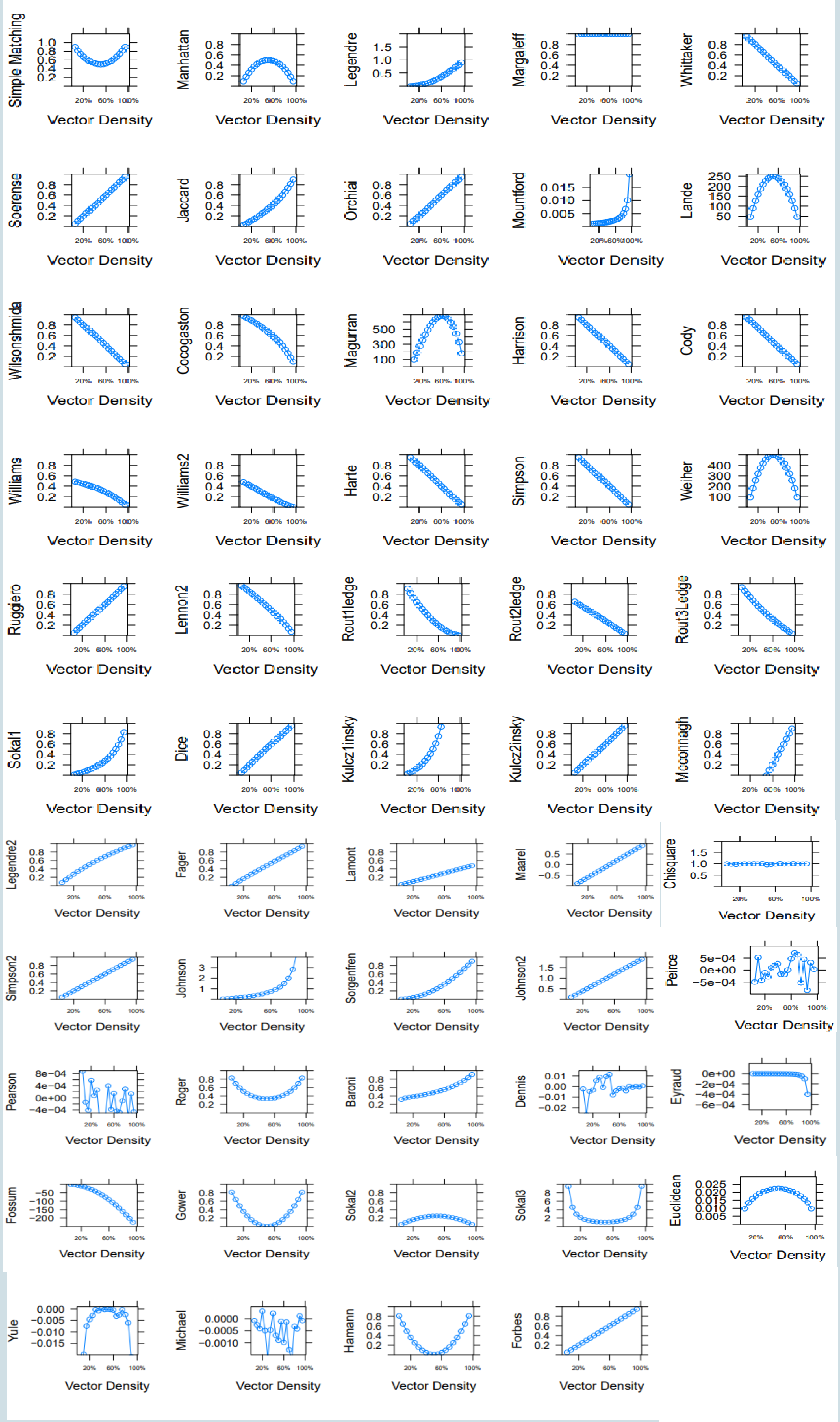

Figure 1 a & b. The X axis displays vector density and was calculated in increments of 5 percent. The Y axis depicts the average magnitude of the selected method. The methods for each vector density were calculated using the R Simba package. The figures were then created using the R library lattice.

For each vector density, 100 randomly generated vectors, each treated as an object, were compared pairwise with the R package Simba for a total of 4050 vector comparisons. Statistics "a,b,c,d" as described above were summed for each comparison, and each of 48 different similarity statistics calculated. **Table 2** displays a partial data set of the first 5 vectors with only the Manhattan method. The mean of the comparison statistic at each vector density for each method are plotted in Figure 1. Vector densities of 0 and 100% were not included.

NBX	NBY	legendre	a	b	c	d
object1	object2	0.012	12	88	88	812
object1	object3	0.007	7	93	93	807
object1	object4	0.009	9	91	91	809
object1	object5	0.008	8	92	92	808
object1	object6	0.014	14	86	86	814

Table 2 : Legendre at 10 percent vector density. The objects represent a randomized binary vector, the a,b,c, and d values are calculated based on the comparison of vectors. The legendre column is determined by a mathematical expression involving a,b,c, and d factors. The legendre average value of numerous randomized comparisons is used to represent a 10 percent legendre data point in Figure 1.

Results and Conclusions:

As expected, the results varied significantly between methods. Differences are seen in the y axis range and graph trends such as the slope. **Figure 1a** and **Figure 1b** shows the generated graphs according to method.

Several methods such as the Sokal3 (**Figure 1b :** Row 4, column 4) show a low rate of change in the 20-80 vector density range)

Methods such as the Pierce (**Figure 1b** : Row 6, column 2) show almost no correlation

Y axis ranges vary greatly among methods.

Johnsons 2 (**Figure 1b** : Row 2, column 4) and other methods show a linear correlation with density

Johnson method data shows minimal change at low vector densities (0-20percent), while Rout1ledge (**Figure 1a** : Row 4, column3) levels off at high densities (80-100 percent)

williams2, Stiles, Lennon, and Divergence were not calculated due to erratic results

Further Analysis :

How known similar/dissimilar environmental data sets would track in comparative to a random set for different densities. This would give a benchmark for understanding what the values mean and what change is significant.

A rate of change set of graphs for the methods. This would tell a researcher when comparing two data sets of differing densities what methods would possibly incur a large change due to matrix density.

Literature Cited

1. Legendre P, Legendre LF. Numerical ecology: Elsevier; 2012.
2. Wolda H. Similarity indices, sample size and diversity. Oecologia. 1981;50(3):296-302.
3. Lewis DM, Janeja VP. An Empirical Evaluation of Similarity Coefficients for Binary Valued Data. International Journal of Data Warehousing and Mining (IJDWM). 2013;7(2):44-66.