

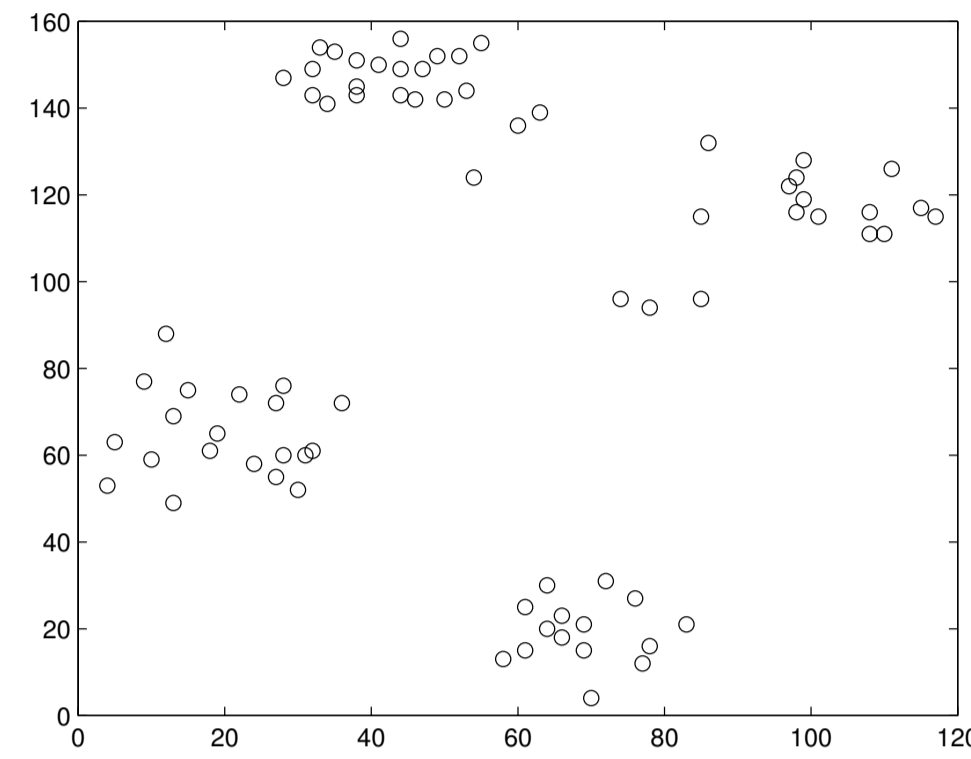


Applying Simon-Ando Theory to Data Clustering



Charles D. Wessell and Carl D. Meyer
North Carolina State University

Can You Cluster this Data Set?



- How many clusters did you find?
- Where are they?

What is Data Clustering?

- Data clustering, or cluster analysis, is the search for patterns in a data set.
- As the above example shows, the human eye is excellent at clustering when the data set is small and low-dimensional.
- Data clustering becomes more interesting and more challenging when the data set is large and high-dimensional.
 - DNA microarray data
 - Customer buying history
 - Netflix Prize data set
- Many clustering algorithms have been developed to help find hidden patterns in large data sets.

A Classic Problem in Cluster Analysis

- Some algorithms do not produce unique clusters
 - k -means
 - Nonnegative matrix factorization
- User may be unsure of best parameter values
 - Number of clusters
 - Distance measure
 - Handling of small or empty clusters
- Proposed Solution:
 1. Cluster the data set a large number of times.
 2. Store the accumulated results in a similarity matrix S where the value of S_{ij} equals the number of times data elements i and j clustered together.
 3. Use the information stored in S to cluster the original data set.
 4. The fact that S often has nearly completely decomposable structure means that Simon-Ando theory can be applied.

Simon-Ando Theory

Consider the doubly stochastic, nearly completely decomposable matrix

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1k} \\ P_{21} & P_{22} & \dots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \dots & P_{kk} \end{pmatrix}_{n \times n}$$

where the elements of diagonal blocks P_{ii} are much larger than the entries in the off-diagonal blocks.

Let x_0 be a random probability row vector and consider the evolution equation

$$x_t = x_{t-1}P.$$

Simon-Ando theory states that as t increases, x_t passes through well-defined stages.

- Initially, the large values in the P_{ii} blocks cause relatively large changes in x_t .
- Since P is doubly stochastic, as $t \rightarrow \infty$, $x_t \rightarrow$ the uniform probability distribution vector.
- Between these two extremes, the elements of x_t accumulate near k distinct values, where k is the number of eigenvalues of P near one.

This last point is essential to using Simon-Ando theory as a data clustering technique since it allows us to cluster the original data based on the clustering of the corresponding probabilities in x_t .

But, how do we convert the similarity matrix S into the doubly stochastic form needed to apply Simon-Ando theory to data clustering?

The Work of Sinkhorn and Knopp

Sinkhorn and Knopp developed the theory behind the necessary and sufficient conditions needed to guarantee that the rows and/or columns of a matrix can be scaled. Though not seen yet in testing, there are two concerns when running the Sinkhorn-Knopp algorithm:

- Convergence rate.
- Does the algorithm destroy the nearly completely decomposable structure?

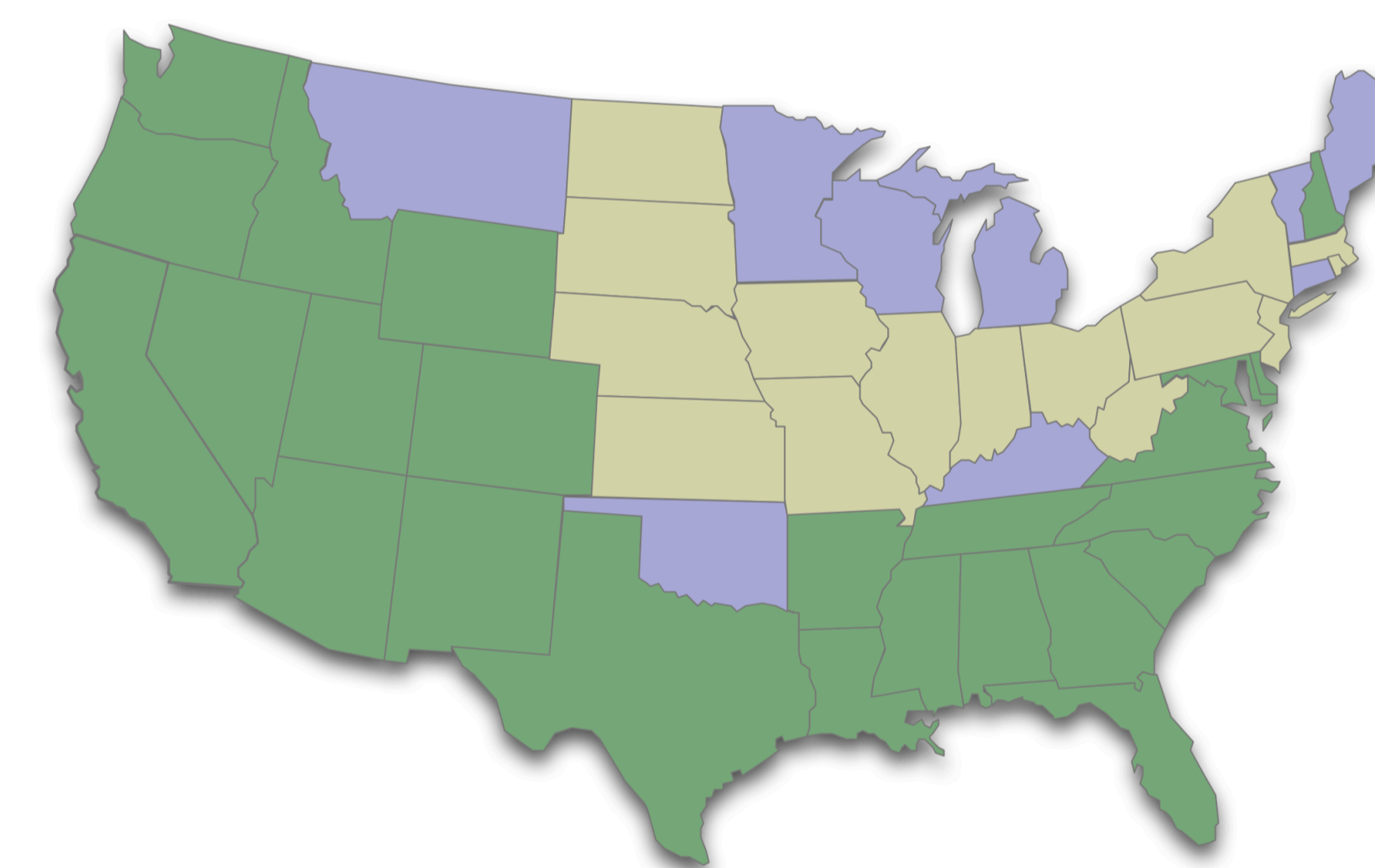
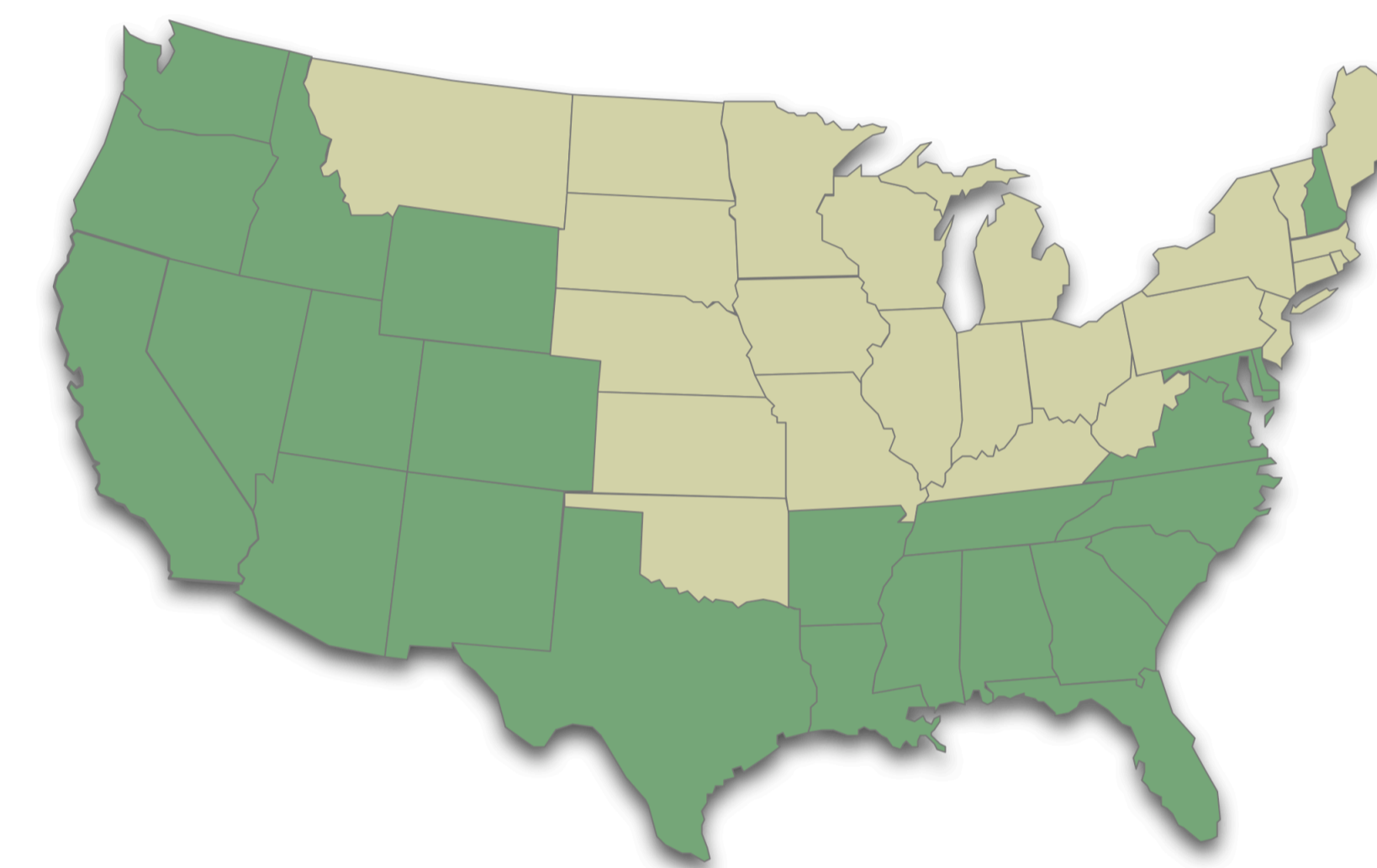
References

- H. A. SIMON AND A. ANDO, *Aggregation of Variables in Dynamic Systems*, *Econometrica*, 29 (1961), pp. 111–138.
- R. SINKHORN AND P. KNOPP, *Concerning Nonnegative Matrices and Doubly Stochastic Matrices*, *Pacific Journal of Mathematics*, 21 (1967), pp. 343–348.

Results

This new clustering method based on the work of Simon and Ando has produced results as good or better than those of test data sets with known correct answers. Here are some results for a problem with no "correct" answer.

- State-by-state presidential election data
- 1912 – 2008
- Each state represented by total votes it gave each candidate in each election
- Data matrix is 88×48
- States clustered using nonnegative matrix factorization 100 times
- Similarity matrix clustered using the Simon-Ando inspired clustering algorithm
- Maps showing results for $k = 2$ and $k = 3$ are below



Future Work

- Scaling up to larger data sets.
- Testing sensitivity to changes in the initial probability vector.
- Developing measures of nearly complete decomposability.
- Developing a tool to allow unsophisticated data analysts to use this algorithm.