EpiVECS: Exploring Spatiotemporal Data Using Low-Dimensional Cluster Representations

Lee Mason (NIH/QUB) Blánaid Hicks (QUB) Jonas Almeida (NIH)

NIH NATIONAL CANCER INSTITUTE

QUEEN'S UNIVERSITY BELFAST

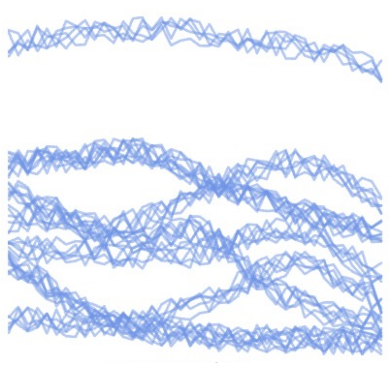
Summary

Visualizing spatiotemporal data is challenging because the spatial and temporal elements compete for the most informative visual channels. This is especially true when the data is large. In this work, we introduce a new method to visualize spatiotemporal data at scale by clustering the data, mapping the clusters onto an informative color space, and then applying these colors back onto the map. We compare several approaches on real-world public health datasets.

Cluster Embedding

We use the term cluster embedding to refer to any method which both clusters high-dimensional vector data and provides a low-dimensional representation of the clusters[1]. The classic approach is self-organizing maps (SOM), but SOMs' rigid grid structure constrains both the clustering and cluster representation. An alternative approach is to use a standard clustering method (e.g. k-means) and then apply dimensionality reduction to the cluster centroids. This often results in a representation of the data[1,2].

Method Overview

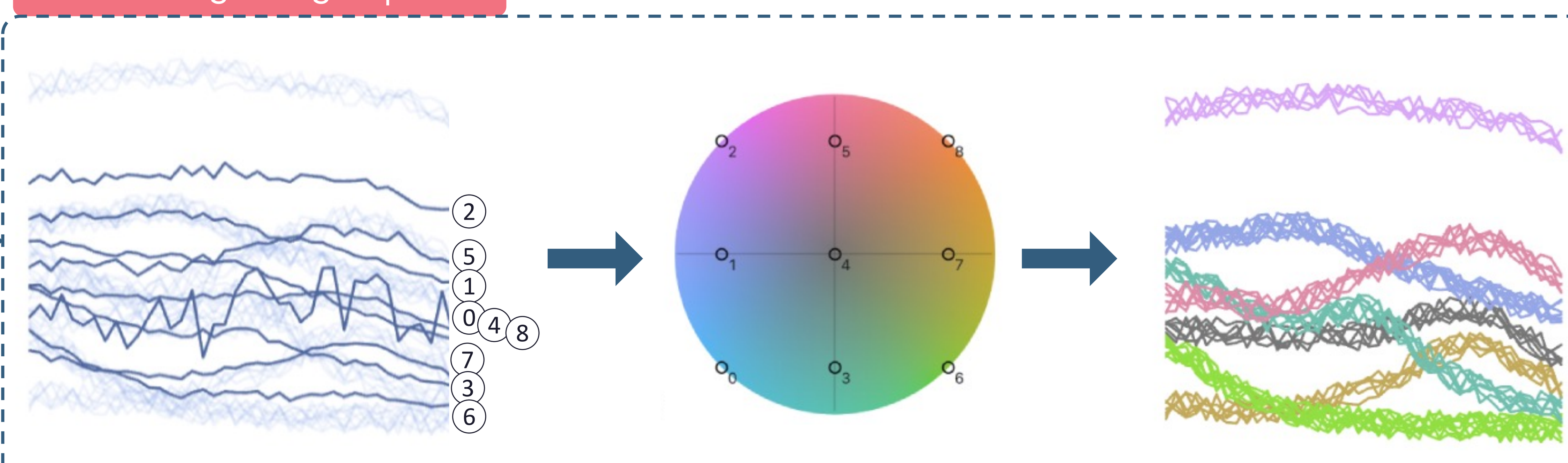


Input vector data

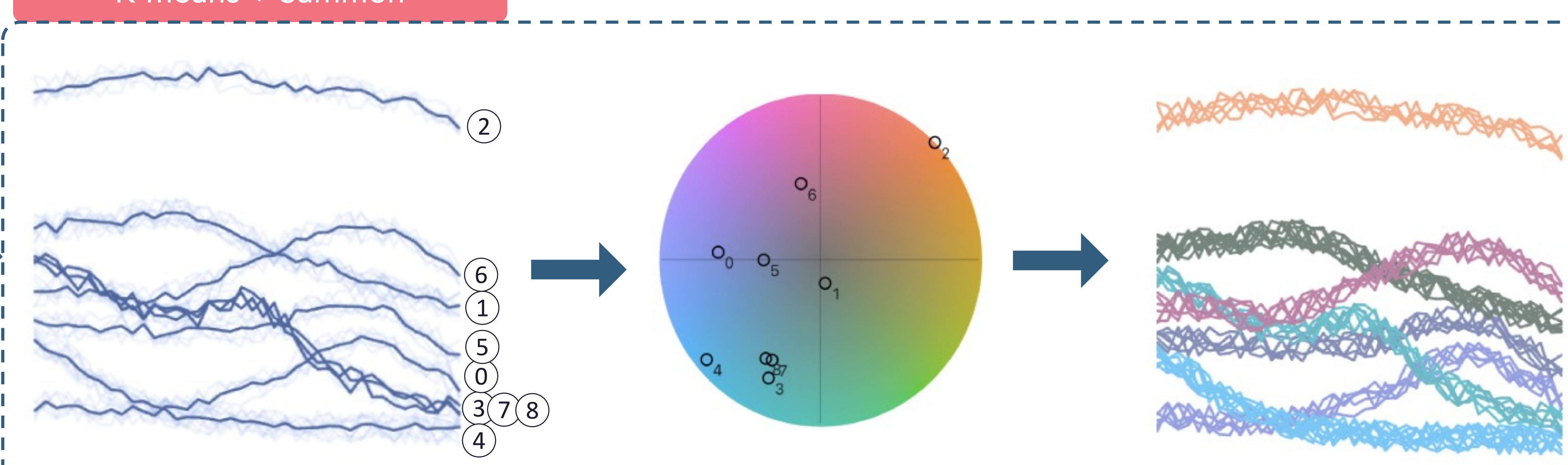
We compare two main approaches in this work. Here we show how each works on simple data.

1. **Cluster embedding** is performed on the time-series vectors.
2. The 2D cluster representations are mapped onto a **perceptually uniform color space** (OKHSL)
3. Each vector is assigned its corresponding cluster's color. These colors can be used across plots

Self-organizing map



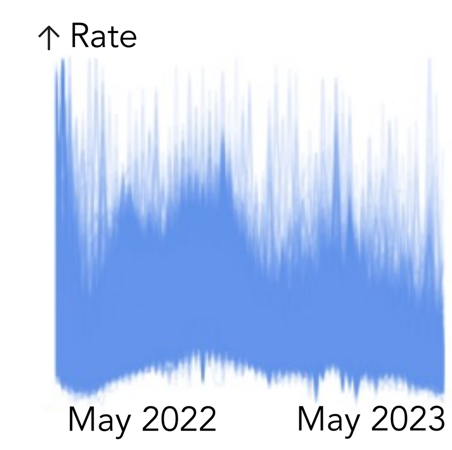
K-means + Sammon



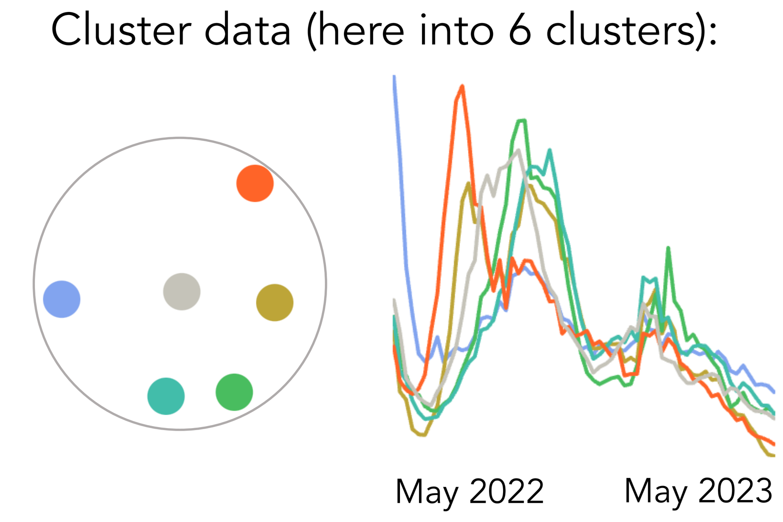
Real World Example

We apply the K-means + Sammon approach to real-world COVID-19 incidence data:

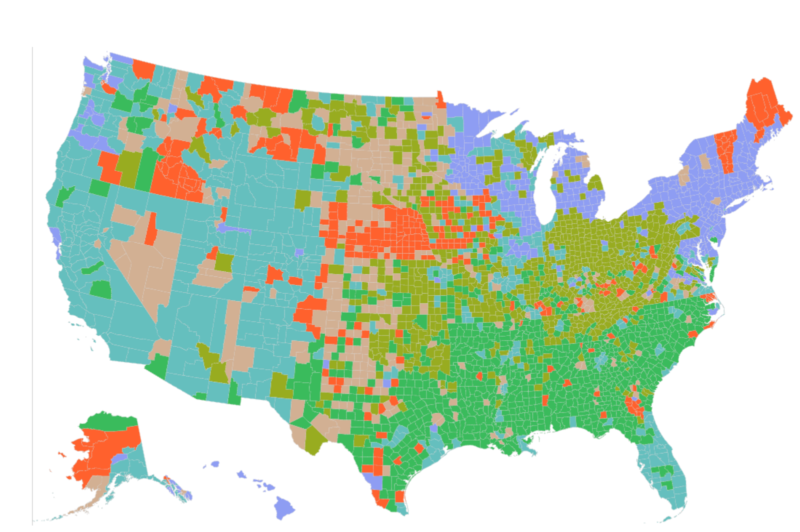
3,143 county rate time series vectors, normalized



Cluster data (here into 6 clusters):



Apply the colors to the map:



Variant comparisons

We compared the SOM approach to several k-means based approaches. We performed a comprehensive comparison on 14 real world spatiotemporal public health datasets, on the county and state level. We used a variety of internal cluster and dimensionality reduction metrics to make the comparison. Here, we summarize the results using a rank score across the metrics. K-means outperformed SOM, and PCA was the best DR technique to use with K-means.

Clustering

Method	Rank Score
K-means	**8**
SOM	7

Embedding

Method	Rank Score
PCA	**27**
Sammon	25
SOM	18
UMAP	14
t-SNE	6

EpiVECS Library and Tool

We have implemented these methods in a JS library: **`epivecs`**. We have also provided an online web-tool, EpiVECS, which performs these methods and presents them to the user in an interactive dashboard. Try the dashboard for yourself by scanning the QR code:

episphere.github.io/epivecs

Discussion

The strength of our method is that it provides a **two-tiered summary** of the data. We capture similarity of the vectors through clustering, then capture similarity of the clusters by positioning them in 2D space. By then mapping the clusters onto a color space, their assigned colors now convey the similarity between clusters.

Future work could explore other cluster embedding techniques or alternative ways to map the clusters onto a color space.

1. Mason, L., Hicks, B., & Almeida, J. S. (2023). EpiVECS: exploring spatiotemporal epidemiological data using cluster embedding and interactive visualization. *Scientific Reports*, *13*(1), 21193.
2. Flexer A. Limitations of self-organizing maps for vector quantization and multidimensional scaling. Advances in neural information processing systems. 1996;9.