Dimensionality Reduction, Clustering, and PlaceRank Applied to Spatiotemporal Flow Data

Blake Shaw – blake@cs.columbia.edu
Tony Jebara – jebara@cs.columbia.edu
Columbia University
Sense Networks

On average over 250,000 yellow taxi cab trips are recorded every day in New York City, capturing the collective movement patterns of millions of individuals. This massive high-dimensional dataset necessitates tools such as dimensionality reduction and clustering algorithms in order to better understand the flow patterns of an urban area. Each city block can be described by a high-dimensional vector of counts representing how much flow there is to and from all other city blocks for each hour of the week. From these high-dimensional vectors we can apply Minimum Volume Embedding (MVE) [3] and spectral clustering [1] to visualize the similarities between places. Furthermore, we can identify the hubs and authorities of the flow network of places. Similar to how PageRank [2] finds the most authoritative places on the web, we can compute PlaceRank to find the most authoritative places in the physical world. These algorithms applied to this unprecedented spatiotemporal dataset offer a unique perspective on the collective behavior of millions of people moving around New York City.

The dataset consists of the latitude, longitude, and timestamps for the start and endpoints of 22.5 million New York taxi cab trips spanning 6 months between January and June 2009. For our analysis we consider the 2000 busiest city blocks, each of which has a minimum of 20 pickups or dropoffs per day. The flow data is represented as a directed adjacency matrix $A^t \in \mathbb{R}^{N \times N}$ for each of the 168 hours in a week where each entry $A^t_{i,j}$ counts the number of trips from block i to block j in weekhour t. The similarity between places can be expressed as a linear kernel: $K = \sum_{t=1}^{168} \frac{1}{2} \left(A^t A^{t\top} + A^{t\top} A^t \right)$, where two places are considered similar if they have similar amounts of flow to other places at similar times of the week.

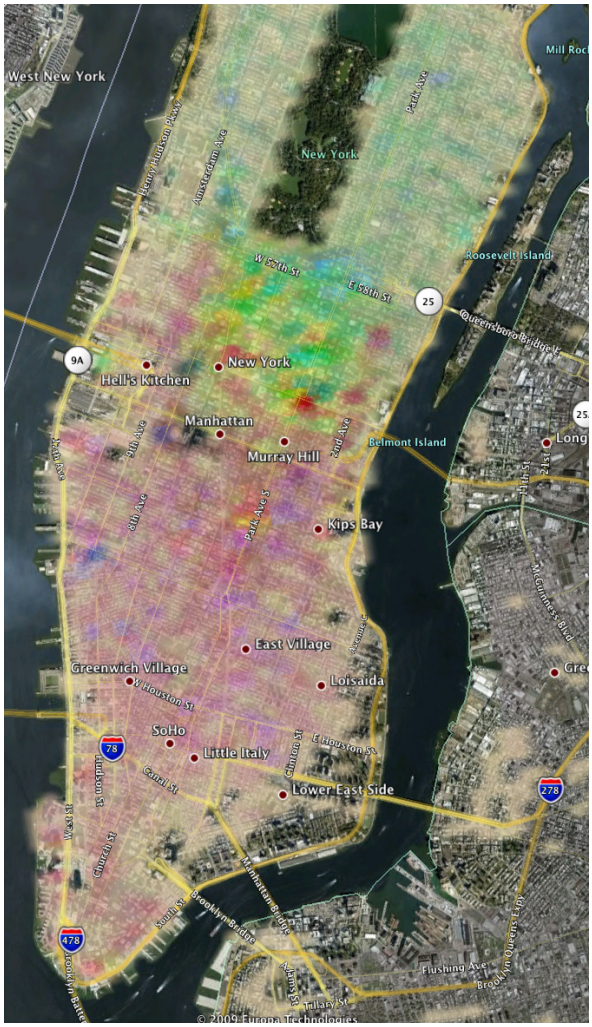
Given K, we can apply Minimum Volume Embedding (MVE) to learn low-dimensional coordinates for each place which best capture the variance of the original high-dimensional data: $\vec{y}_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$, where $d \ll D$, and for this data $D = 2 \times N \times 168 = 672000$. We find that over 99% of the variance of the original data can be preserved using only the top 3 MVE dimensions. In Figure 1(a) we translate these 3 coordinates to RGB to color code each city block, allowing us to visualize the smooth modes of variation in the flow patterns of places in the city. Figures 1(b) and 1(c) show the result of applying spectral clustering to the kernel K to assign one of $k = 12$ labels to each place, thus visualizing the natural neighborhoods that emerge out of the data. Figures 1(d) and 1(e) show heat maps representing the hubs and authorities in the cab flow network which were computed by finding the stationary vector of the total flow adjacency matrix $A = \sum_{t=1}^{168} A^t$ and its transpose using the power method. Places with high authority values have inbound traffic from many other high authority places, similarly places with high hub values have outbound traffic to many other hubs.

References

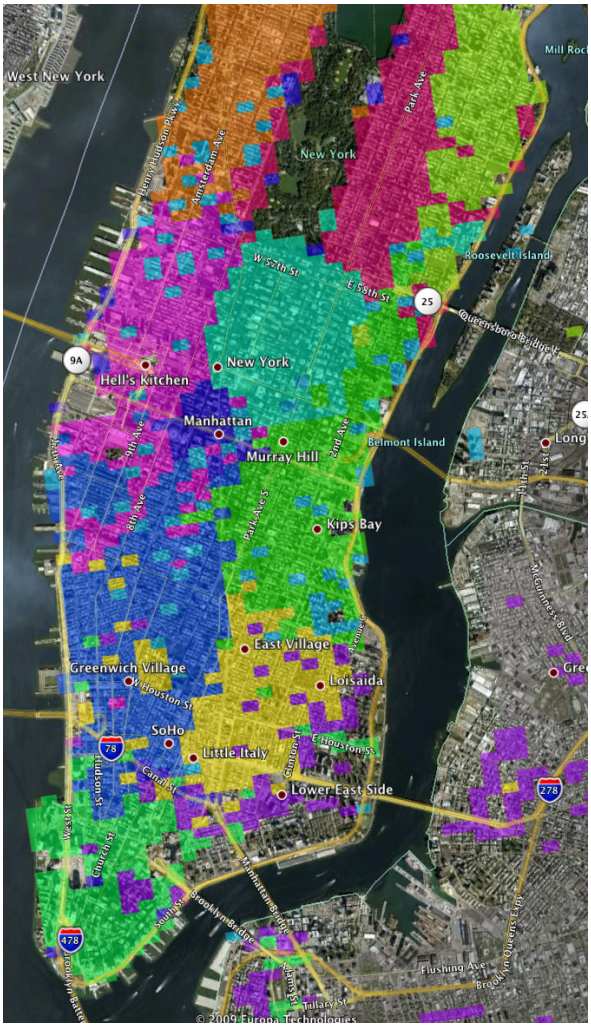
[1] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems 14*, 2001.

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

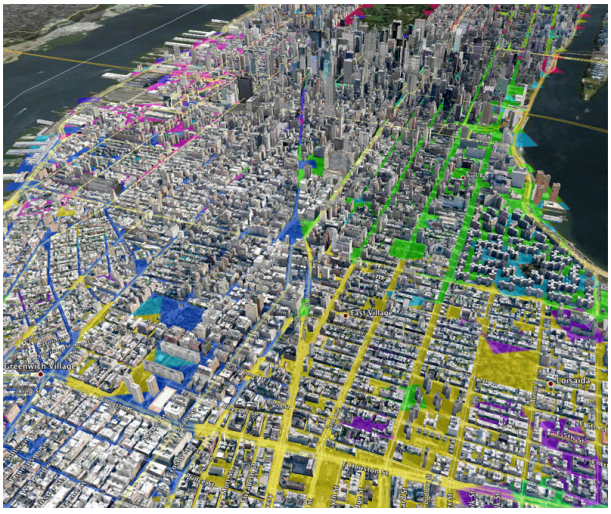
[3] B. Shaw and T. Jebara. Minimum volume embedding. In Marina Meila and Xiaotong Shen, editors, *Proc. of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2 of JMLR: W&CP, pages 460–467, March 2007.



(a) MVE



(b) Spectral Clustering



(c) Spectral Clustering (downtown)

(d) Hubs

(e) Authorities