# A new method of selecting K-means initial cluster centers based on hotspot analysis

CHEN Qu, YI Hong, HU Yujie, XU Xianrui, LI Xiang

[a]*School of Geographical Sciecnes, East China Normal University, Shanghai , China;*
[b]*Department of Real Estate, East China Normal University, Shanghai , China;*
[c]*School of Geosciences, University of South Florida, Tampa, FL, USA;*

**Please Contact: chenq_gis@hotmail.com**

## INTRODUCTION

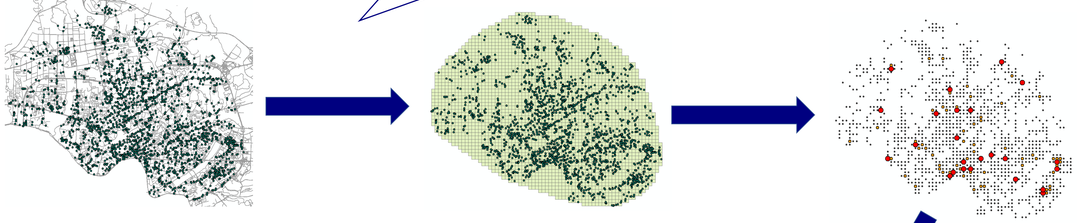
Clustering, different from classification, is process to group events into different clusters according to their characteristics. A high-quality clustering result has less within-cluster difference while more between-cluster difference in terms of event attributes. Most existing K-means clustering methods randomly select initial cluster centers from data samples. The random initialization may lead to significantly different clustering results even if the same clustering algorithm is used. By this means, the quality of solutions cannot be ensured.

On the other hand, hotspots refer to some specific areas with high occurrence of certain events or values, and some works have been conducted to apply spatial clustering methods to discover hotspots. In the contrast, if hotspots can be identified first with other analysis quickly, and then, in turn, these hotspots may be employed as initial cluster centers and may lead to solutions outperforming existing ones.

## METHODOLOGY

Experiment data

Build contagious squares with equal area that fully cover the study area, named cells. Cells are analytical units of hotspot analysis.

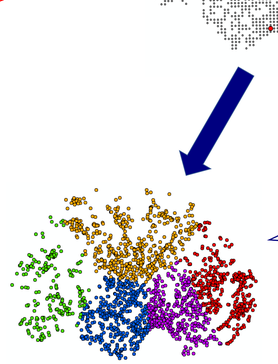


$$C_N^2 = N * (N-1)/2$$

$$S_d = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d(x_i, x_j)$$

$$D_a = S_d / (C_N^2 * f)$$

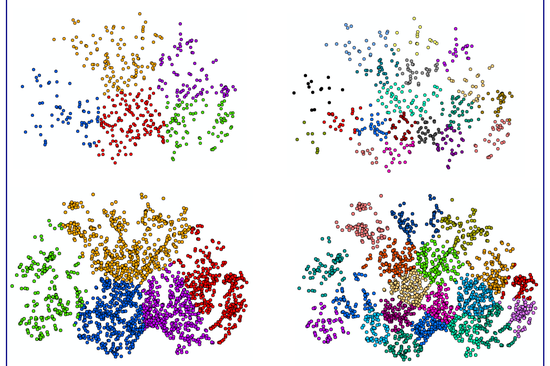
For each hotspot, we set a search radius $D_a$ to find other hotspots that fall in the search area. Then, through comparing the attribute values and confidence levels of these hotspots, identify one hotspot as the initial cluster centers. Repeat the above procedure, until the number of initial cluster centers is *K*.



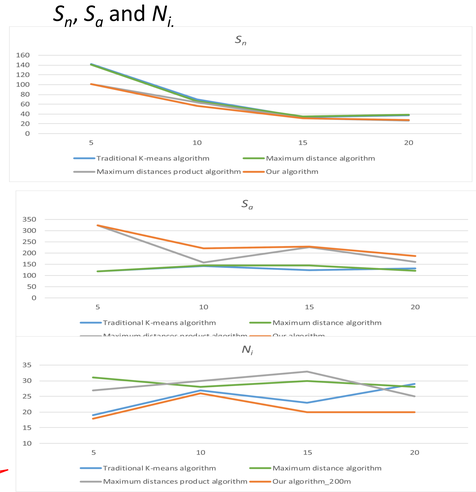
Cluster based on K-means algorithm

## RESULT

There are the clustering results when the number of retail stores is 500 or 2000, the cell size is 200 meters and the number of clusters is 5 or 20.



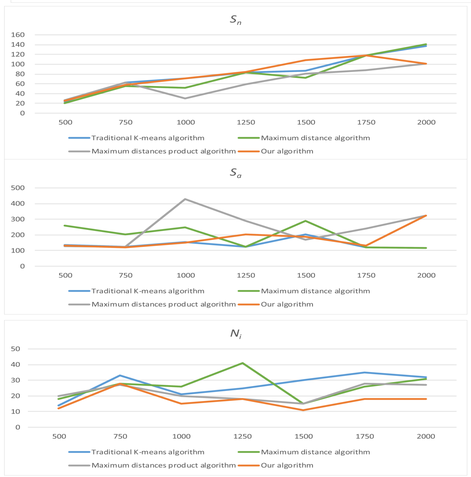
For the purpose of comparison, three statistics are calculated to evaluate the performance of the algorithm, including $S_n$, $S_a$ and $N_i$.




For the purpose of comparison, three other cluster approaches are implemented. There are two scenarios:

1)When the number of stores is fixed as 2000, *K* is set as 5,10,15,20 respectively.
2)When *K* is fixed as 5, increase store numbers.

In terms of $S_n$ and $S_a$, it is hard to figure out the best approach. All of them lead to clustering results with similar quality. $N_i$ with our algorithm is the least in almost all cases. Furthermore, the changes of its $N_i$ are less dynamic than others when the number of stores or clusters changes..




## CONCLUSION

The proposed algorithm is compared with three other existing algorithms in our experiments. Results demonstrate that, when the same ending conditions are employed, all algorithms can generate clustering results with similar quality, while our algorithm is the most efficient. Besides, our algorithm exhibit stable performance for different problem scales or cluster numbers.