Department of Geography and Geoinformation Science

4400 University Drive, MS 6C3, Fairfax, Virginia 22030
Phone: 703-993-1210, Fax: 703-993-9299
Email: ggs@gmu.edu Web: ggs.gmu.edu

GGS 590 - Assignment 4
Due date: April 28, 2015

The Task

The file `Olympic_torch_2012_UK.csv` contains a set of geolocated tweets that was collected during the 2012 London Olympic games in the area covering the United Kingdom and the surrounding region (as shown in Figure 1). The data file is organized as follows:

Column	Description
1	Tweet timestamp (UTC)
2	Latitude (decimal degrees)
3	Longitude (decimal degrees)
4	Tweet sentiment (integer value). Large positive values represent highly positive tweets; large negative values represent highly negative tweets.

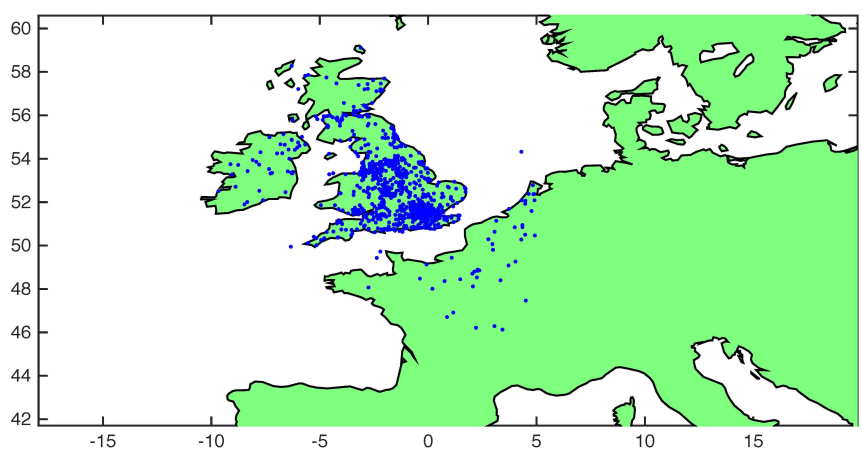

Figure 1: An overview of the `Olympic_torch_2012_UK.csv` dataset.

Using this data, your task is to identify clusters of tweets during the 2012 Olympic games. In order to accomplish this you may use any of the methods we reviewed in class and/or any other clustering method you think would be appropriate for this data. Recommended software tools

and libraries that you may build upon for clustering include (but are not limited to) PySAL[1], Scikit-learn[2], R-spatstat[3], and a variety of other cluster analysis tools in R[4]. After performing your cluster analysis please submit the following:

1. (Up to 350 words) Describe your overall approach for clustering the data set, i.e. what are the key principles that drive the identification of clusters (e.g. density, distance, attribute similarity, etc. or their combination). In addition, please provide a clear justification for your approach (i.e. why is your approach appropriate for clustering this data set). If you are using a clustering method that was not covered in class please provide appropriate references.

2. (Up to 350 words) Describe how your clustering process works. Please provide a flowchart showing the different steps of your analysis, and include key equations as necessary.

3. (Up to 250 words) Report and evaluate the results of your clustering approach. In particular, how many clusters were found? What was their size? How many points were not clustered? In your view, what was the overall performance of the method? Where did it perform well/poorly? Can you offer a possible explanation why?

4. Provide a map visualization of the study area showing your clustering results. Your visualization should clearly show the clusters that were detected, and should allow the viewer to differentiate between adjacent clusters. Your map should be included in your lab report, and should be also provided as a separate single page PDF document.

[1] https://pysal.readthedocs.org/en/latest/
[2] http://scikit-learn.org/stable/
[3] http://cran.r-project.org/web/packages/spatstat/index.html
[4] http://cran.cnr.berkeley.edu/web/views/Cluster.html