Christian Felt and Jadon Wagstaff

Data Mining Spring 2017

Instructor: Jeff Phillips

April 8, 2017

Clustering and Density of Endangered Species from the IUCN Red List

Christian Felt

My goal was to determine whether the clustering techniques I learned in class could divide the ranges of endangered species into biologically reasonable groups. I assessed whether the clusters seemed biologically reasonable or not by comparing them to the areas of high biodiversity identified by Meyer et al in “Biodiversity Hotspots for Conservation Priorities” (see Figure 2). My conclusion was that Gonzalez clustering, simply by greedily choosing each new center to maximize the geographic distance from previous centers, was able to find biologically meaningful groups with a surprising degree of accuracy, though it also made notable errors, particularly where Meyer’s hotspots are located close together (in which case Gonzalez tended to combine them), or the density of endangered species was low (in which case Gonzalez tended to assign the species to clusters centered far away).

Using ArcMap software, I computed the centroids of the polygons that represent the ranges of terrestrial animals (mammals, amphibians, and reptiles) in the IUCN Endangered Species Red List shapefiles, making sure that the centroid was adjusted if necessary to lie within the polygon. Using my own Python code, I turned the latitude and longitude coordinates of these centroids into a 2-dimensional matrix and performed various kinds of clustering.

To measure distance, I used the great-circle distance as given by d from the Haversine formula:

a = sin²(Δφ/2) + cos φ1 ⋅ cos φ2 ⋅ sin²(Δλ/2)

c = 2 ⋅ atan2( ,  )

d = R ⋅ c

Since the Earth is not perfectly spherical, the great-circle distance may be wrong by up to approximately 0.3% or 22km according to the analysis at <http://gis.stackexchange.com/q/25494>. However, this error seems acceptable since in their article “Global Patterns of Terrestrial Vertebrate Diversity and Conservation,” Jenkins et al set 100 km as a reasonable lower limit for “fine-grained spatial analysis” in this field.

Hierarchical clustering of any type ran too slowly for the full dataset (74,529 points) or even for amphibians alone (18,694 points). K-means++ worked for small numbers of clusters (20 or less) but became obnoxiously slow for larger k. Lloyd’s was also too slow for more than about 20 clusters.

Gonzalez clustering, however, processed the entire dataset in less than 10 seconds. Using the “elbow principle” and my subjective sense of how many clusters looked good on the final map (adopting a scale similar to Meyer’s), I chose k=40, and obtained the following clusters:



Figure Gonzalez with k = 40

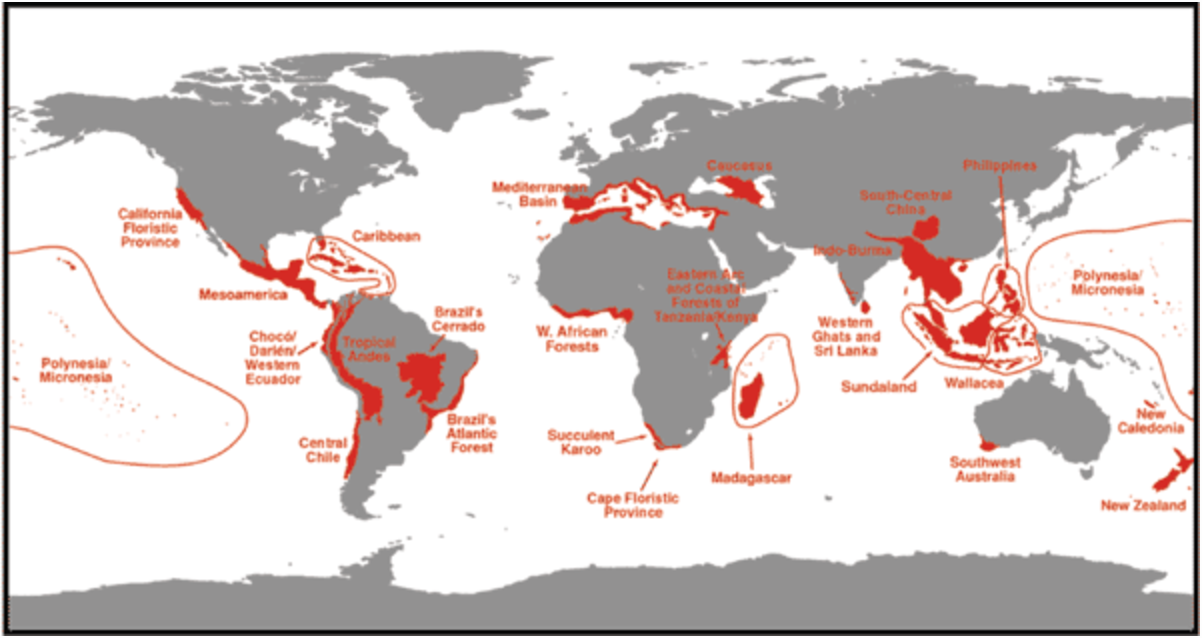


Figure 2 25 Biodiversity Hotspots from "Biodiversity Hotspots for Conservation Priorities" by Meyer, et al (2000). http://www.nature.com/nature/journal/v403/n6772/full/403853a0.html

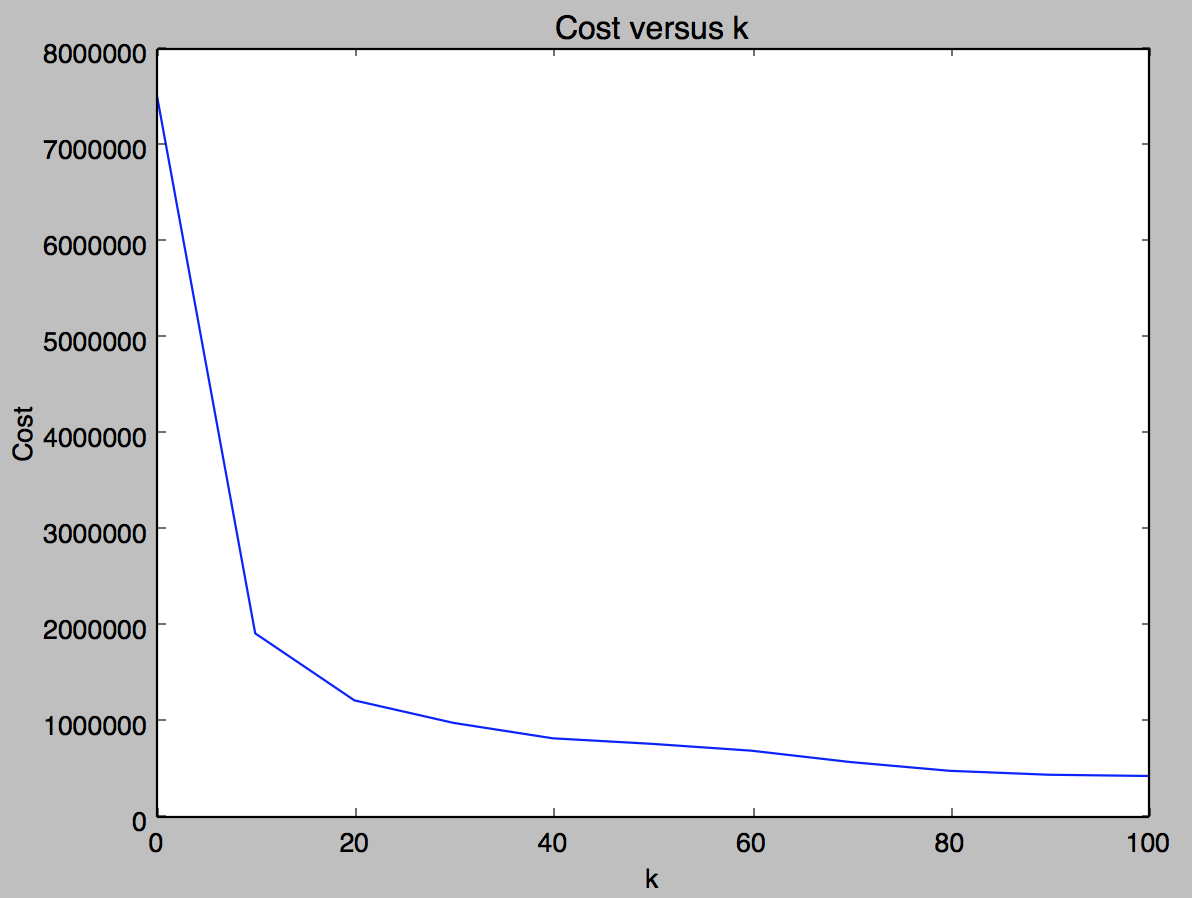


Figure 3 k-means cost in meters for Gonzalez

These clusters seem surprisingly reasonable, rarely breaking up Meyer’s biodiversity hotspots. According to Meyer, “[As many as] 35% of all species in four vertebrate groups are confined to 25 hotspots comprising only 1.4% of the land surface of the Earth…The hotspots’ boundaries have been determined by ‘biological commonalities’. Each of the areas features a separate biota or community of species that fits together as a biogeographic unit.” It is interesting, then, that Gonzalez clustering, while taking into account nothing but geographic proximity, nevertheless manages to pick out “communities of species” that are related in more biologically profound ways. For instance, Gonzalez, like Meyer, assigns Madagascar, New Zealand, and the Caucasus Mountains to their own groups, separates the “California Floristic Province” from the rest of the Rockies, the “West African Forests” from the rest of Africa, and divides up the Indonesian Islands into three different groups, with the Philippines being mostly on their own. The Aleutian and Canary Islands, Newfoundland, and Hawaii are given their own clusters, and the Andes, Himalayas, and other mountain chains are not too broken up.

Sometimes, however, Gonzalez doesn’t break groups up enough. It is blind to Meyer’s “Succulent Karoo,” located right next to “Cape Floristic Province,” and fails to separate the “Central Chile” or “Western Ecuador” hotspots from the larger “Mesoamerica” or “Tropical Andes” groups, respectively. Elsewhere, Gonzalez separates too much, for instance, breaking the Mediterranean Basin into two clusters. It also seems wrong that Gonzales has divided Siberia into so many strips (almost coinciding with Russia’s eleven time zones.) This is clearly a consequence of the low density of endangered species in Siberia; the data points there tend to be drawn into the clusters of distant, more densely-populated areas. A similar problem occurs in Northern Canada and the Sahara.

In summary, Gonzalez clustering provides a fast way to group endangered species in ways that make a surprising amount of biological sense. Since Gonzalez relies only on geographic distance, “knowing” nothing else about the animals, the explanation for why Gonzalez seems to be able to identify biological relationships must be that biological relationships are strongly conditioned by geographical proximity, and that the boundaries of terrestrial ecosystems are relatively discrete.

While my section focused more on clustering, using the centroids of the species’ range polygons so that there is only one point per species, Jadon’s section will address the question of density of endangered species, using a point cloud where each of the 534,809 points is at the center of a square of area approximately 20 km. The grid of squares covers the Earth, and the presence of a point in a square means that the polygon representing the range of an endangered species intersects with that square. This allows species with large or irregular ranges to be assigned points that map their entire range, rather than assigning them to a single centroid point that may not provide a good model of their range.

Jadon Wagstaff