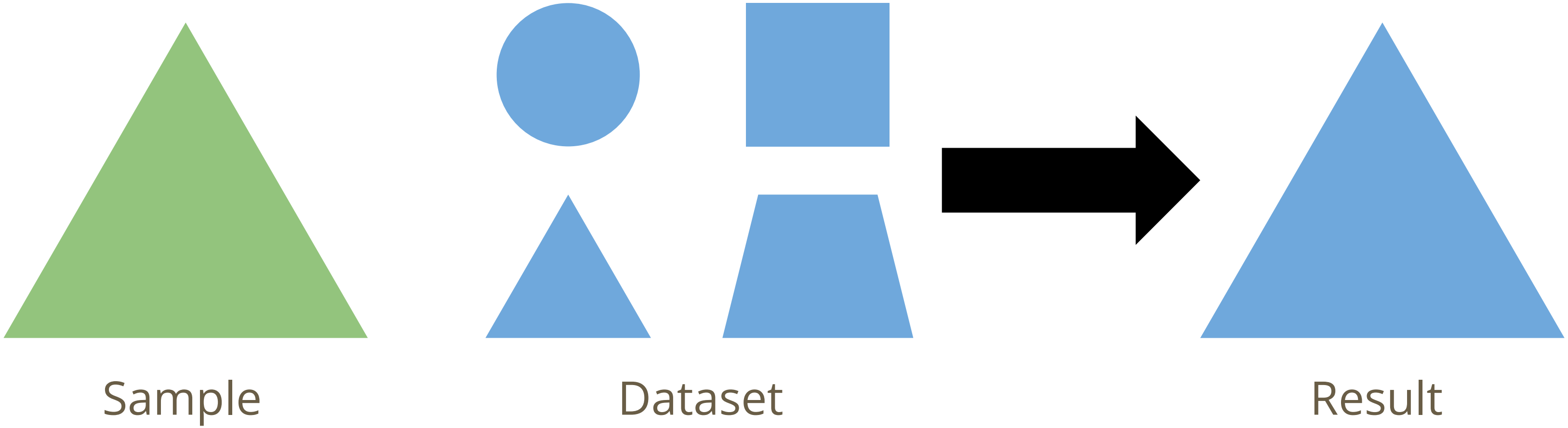
Shazam for Shapes

Using Similarity Estimation on Polygons

By Jorre Dahl

The Goal

- Given two shapefiles, one representing a single sample polygon and the other representing a dataset of many polygons, find the most similar polygon in the dataset to the sample polygon.



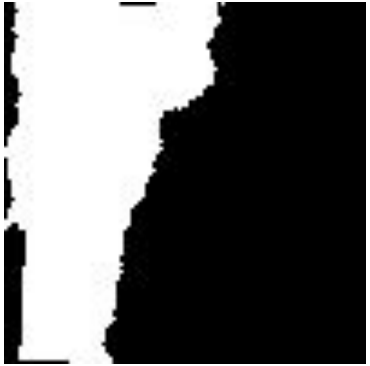
Sample

Dataset

Result

Converting a Shapefile to a Vector

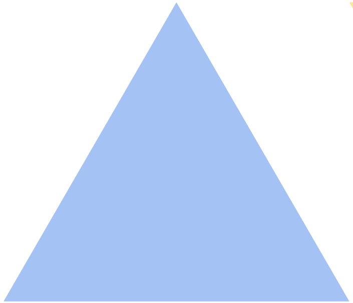
- Using small squares convert shapefile into a raster, where 1 represents pixels in the shape and 0 represents pixels outside the shape
- Convert raster into binary array.
- Fill empty space to the right side or bottom to make the array have a square resolution.



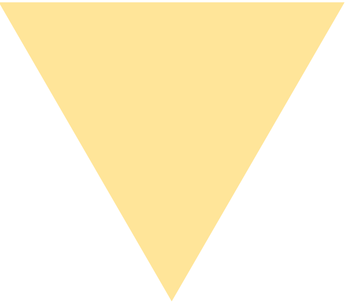
This is Vermont as a binary vector with resolution 100x100.

Effects of This Method

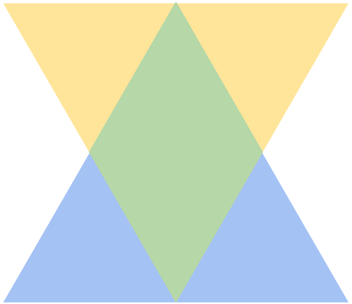
- Creates a binary array, so Jaccard Similarity can be used to find two polygons’ similarity.
- Creates a problem of rotation where rotated identical shapes will not return a high similarity:



Shape x



Shape y

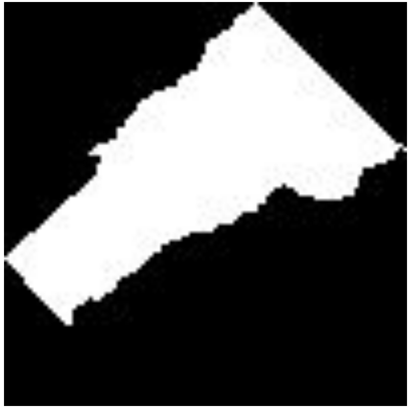


$J(x,y) = 1/3$

Intersection and Union of x and y

Solution to This Problem

- Find the maximum similarity of shape x and y for all rotations of shape x.
- Before converting to raster, convert each shapefile to a polygon and apply a rotation counterclockwise to the shape.



This is Vermont as a binary vector with resolution 100x100 rotated 315 degrees counterclockwise.

Application of This Program

- Given the shape of a state, I want to find the most similar country.

- Comparison requires the use of the Mercator Projection.

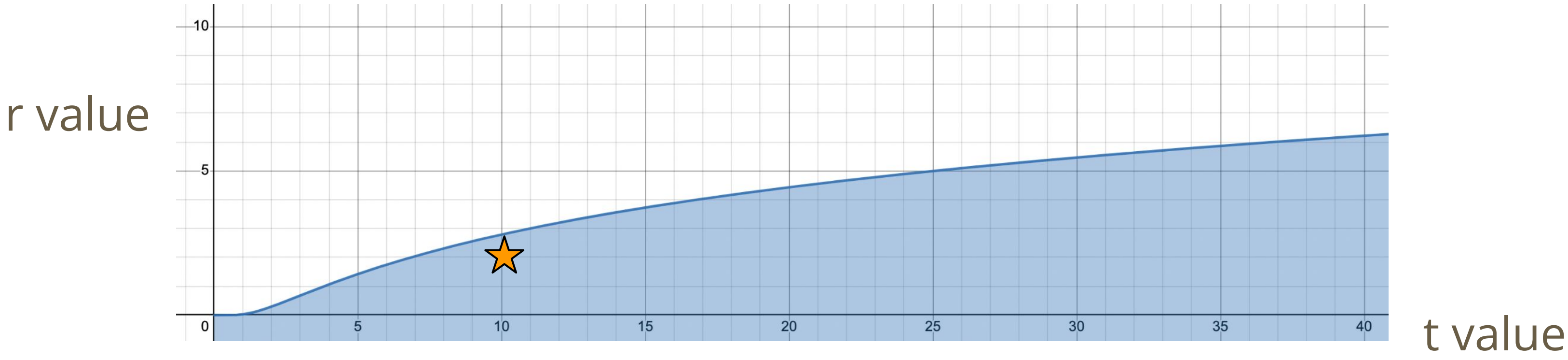
- Locally Sensitive Hashing is useful to reduce the number of comparisons made.

- Uses r bands of the minHash function then hashes to an m by t table

Doing the Math

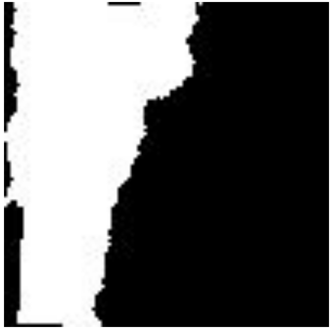
- From our class notes, the probability that a vector y from the dataset is checked from sample vector x is **Pr(find y) = 1 - (1 - (J(x,y))^r)^t**
- If we want the probability of checking y to be over 0.99 for all J(x,y) = 0.7, we can say the ratio of r to t is:

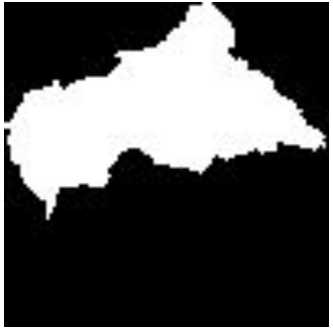
r ≤ (log(1 - (0.01)^t)) / (log(0.7))

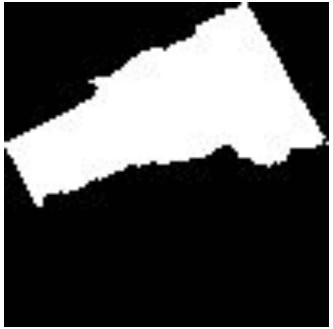


Results

- The most similar country in shape to Vermont is…
- Central African Republic (~0.73 similarity)







Results

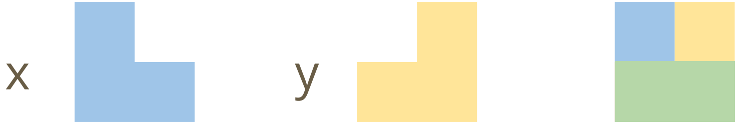
- The most similar country in shape to Texas is...
- Nigeria (~0.71 similarity)

Results

- The most similar country in shape to California is…
- Hungary (~0.77 similarity)

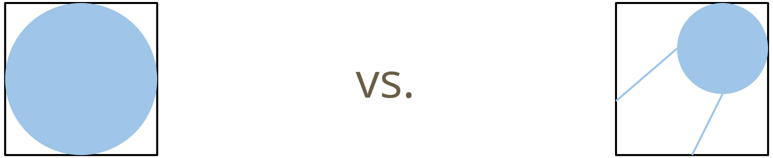
Problems and Things to Improve

- What about reflected shapes? Reflected across which axes?



$J(x,y) = 1/2$

- Does bounding shapes in a box really capture similarity? Why not use the centroid of a shape and give all shapes the same area?



- Can I change the code to instead read in a drawn polygon as my sample vector instead of a shapefile?
- My code is not very accurate...

Thank You