Evaluating Unsupervised Clustering Accuracy

Introduction

One of the most consistent guidelines to succeed with a data science project is to tap into subject matter expertise. As a former Surface Warfare Officer in the US Navy, I spent over four years operating on ships and am familiar with ship navigation and operations. Additionally, I am always interested in Geospatial Information Systems (GIS) so I choose a project that would combine my love of all things maritime and GIS.

This project analyzes shipboard Automatic Identification System (AIS) activity as collected and shared by the US Coast Guard. AIS is an automatic tracking system that uses transponders on ships and is used by ships and port authorities to avoid collisions and manage ship traffic. Because of the tremendous size of the data, I have limited my analysis to a section of data from January 2017 that includes 500 million individual position reports from 20,731 unique ship identifiers, known as MMSIs.