Geospatial data using KMeans clustering

# **Geospatial data**

**Geospatial data** is information that describes objects, events or other features with a location on or near the surface of the earth. Geospatial data typically combines location information (usually coordinates on the earth) and attribute information (the characteristics of the object, event or phenomena concerned) with temporal information (the time or life span at which the location and attributes exist).

The location provided may be static in the short term (for example, the location of a piece of equipment, an earthquake event, children living in poverty) or dynamic (for example, a moving vehicle or pedestrian, the spread of an infectious disease).

Geospatial data typically involves large sets of spatial data gleaned from many diverse sources in varying formats and can include information such as census data, satellite imagery, weather data, cell phone data, drawn images and social media data. Geospatial data is most useful when it can be discovered, shared, analyzed and used in combination with traditional business data.

# **Geospatial using python**

The popular programming language Python is well suited to working with geospatial data and can accommodate both vector data and raster data, the two ways in which geospatial data are typically represented. Vector data can be worked with by using programs such as Fiona and GeoPandas. Raster data can be worked with by using a program such as xarray.

# **Types of geospatial data**

### Vector Data

Vector data is primarily made up of points, lines, and polygon shapes. Each of these elements is defined by a coordinate in space.

* **Points**: Represent specific locations like street lamps, mailboxes, or any other object that can be represented by a single x, y coordinate. For example, the location of a particular store in a shopping mall can be depicted as a point.
* **Lines**: Used to illustrate linear features such as roads, rivers, and utility lines. Each line is made up of a series of points connected in a particular sequence.
* **Polygons**: These are closed lines used to represent areas like lakes, building footprints, and city boundaries. Polygons are made up of multiple lines connected end-to-end.

Vector data is often preferred for tasks that require precise measurements and clear delineations, such as cadastral mapping, transportation planning, and network analysis.

### Raster Data

Raster data consists of a matrix of cells or pixels that cover a specific geographical area. Each cell holds a value representing information like elevation, temperature, or vegetation type. This kind of data is particularly useful in applications like satellite imagery, weather mapping, and geological surveys.

* **Cell-based Representation**: The granularity of the data is defined by the size of the cell. Smaller cells mean more detail but result in larger file sizes.
* **Value Representation**: Each cell can hold one or multiple values, offering either a simple representation, like land vs water, or a complex one like the RGB values for color images.

Raster data is beneficial for applications that require continuous data over a specific area, such as land cover mapping or climate change.

# **Application of geospatial data:**

* Navigation and mapping
* Environmental and resource management
* Urban planning and development
* Emergency response and disaster management
* Logistics
* Transportation
* Meteorology
* Forestry
* Agriculture
* Healthcare
* Ecology
* Marketing and advertising

# **KMeans clustering in geospatial dataset:**

K-Means clustering is a technique used to group geospatial data points into clusters based on their locations. Here’s a simple approach:

1. **Data Preparation**: Collect geospatial data (e.g., latitude and longitude). Ensure the data is clean and formatted correctly.
2. **Choosing Clusters**: Decide the number of clusters (K) based on your analysis goal or using methods like the Elbow Method.
3. **Algorithm Execution**: Apply the K-Means algorithm. It randomly assigns data points to K clusters, then iteratively adjusts the cluster centers to minimize the distance between data points and their respective cluster centers.
4. **Results Interpretation**: Analyze the clusters to identify patterns or trends in the spatial distribution of the data.
5. **Visualization**: Plot the clusters on a map to visually interpret the geospatial patterns and make data-driven decisions.

This method helps in uncovering spatial groupings and patterns within geospatial datasets.

For reference use the: https://github.com/Parthieshwar/Geospatial-data-using-KMeans-clustering/blob/main/geo\_location.ipynb