KAATRU, IIT MADRAS

creation of structured data containing interpolated values and geohash strings

Suraj Bhosale

16/01/2023

Break up of problem statement

- Get coordinates (latitude, longitude) for the data points within grid
- Get interpolated values of PM 2.5 at each point in grid
- Get geohashing strings for each coordinate on grid
- Cluster kriging approach for reduction of space-time complexity
- Use of different clustering methods to partition dataset into several clusters
- Perform interpolation and geohashing operations over clustered files
- End product - Dataframe/File consisting latitude, longitude, interpolated PM2.5 values, geohashing strings

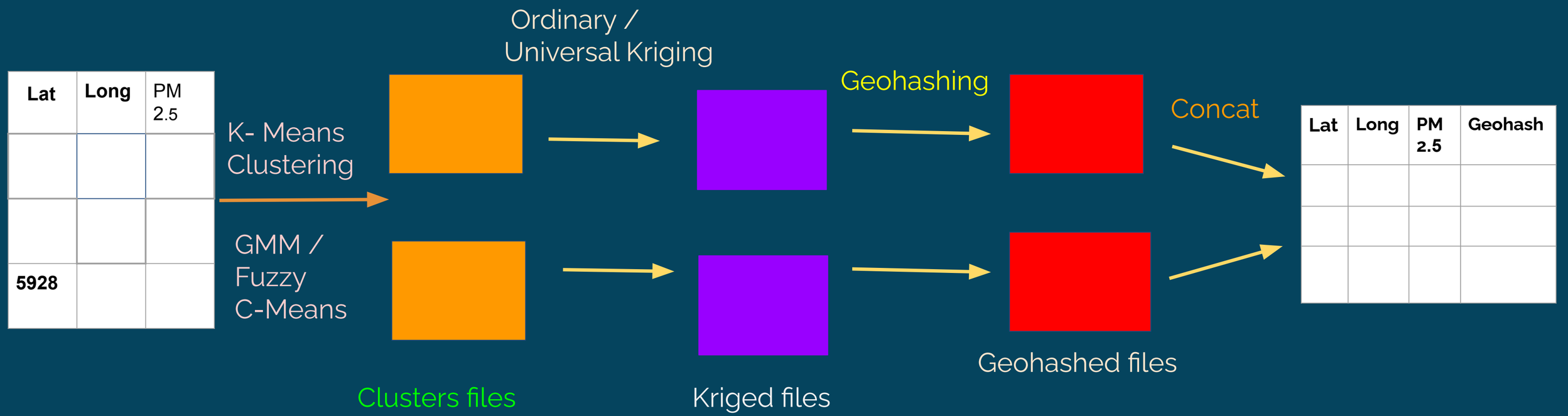
Cluster Kriging approach for space-time complexity reduction:

- *3 approaches to partition given data into several clusters*

1. *Hard clustering :*
- *K-Means Clustering*

1. *Soft clustering :*
- *Gaussian mixture models (GMM)*
- *Fuzzy C-Means clustering*

Procedure



Lorem ipsum dolor sit amet. Quo doloribus quod ut voluptas nemo in sint asperiores eum reprehenderit nihil aut eaque voluptatibus.

Ordinary Kriging

- A linear combination of the values of the regionalized variable at known locations

$$Z^*(u) = \sum_{i=1}^{n} \lambda_i Z(u_i)$$

Where Z^* is value at unknown location, λ is the weight and Z is value at known location

- Best Linear Unbiased Predictor - Minimizes the estimation variance with respect to Unbiasedness condition

Universal Kriging

- $Z(\mathbf{x}) = m(\mathbf{x}) + e'(\mathbf{x})$

Mean is a function of the site coordinates

$$m(\mathbf{x}) = \sum_{k=1}^{n} \alpha_k p_k(\mathbf{x})$$

a_k	are the local trend or drift coefficients
$p_k(\mathbf{x})$	are functions of the site coordinates (trend equations)
$\mathbf{x}$	is a two dimensional vector

Kriging and Geohashing with Multithreading (GMM clusters)

No. of records in I/P file	Grid Space	Execution time for (Kriging + Geohashing)	Execution time for Concatenating records	No. of records in O/P file
5928	0.01	51 Minutes	9 Minutes	$(32.8 \times 10^6, 4)$

Kriging and Geohashing without Multithreading (K-Means clusters)

No. of records in I/P file	Grid space	Total Execution time	No. of records in O/P file
5928	0.01	20 Minutes	$(9.96 \times 10^6, 4)$

Ordinary Kriging : 7 - 12 Minutes

Geohashing : 5 Minutes

Concatenation : 3 Minutes

Kriging and Geohashing with Multithreading (Fuzzy C-Means Clusters)

No. of records in I/P file	Grid space	No. of records in O/P file	Execution Time for (Kriging + Geohashing)	Execution Time for Concatenating records
5928	0.01	$(63.4 \times 10^6, 4)$	80 Minutes	19 Minutes

Outcomes

- K-Means clusters are on average same in size (cluster items)
- K-Means clusters takes less time for ordinary kriging and geohashing
- GMM clusters follow gaussian distribution thus there are some clusters having high density of data points while some are low in cluster items
- GMM clusters have large file sizes compared to K-Means clusters file size
- Thus time required for ordinary kriging and geohashing over GMM clusters is more compared to K-Means clusters
- Data generated through Fuzzy C-Means clusters is obviously huge compared to GMM clusters and K-Means clusters

Outcomes

- By cluster kriging approach, time required for execution of kriging and geohashing reduced
- After performing kriging and geohashing over clusters all the data aggregated in one dataframe
- Resulting dataframe contains latitude, longitude, PM2.5 concentration and geohash string at each and every point in the grid which is (structured data format) suitable for RDBMS database
- Multithreading approach becomes handy for processing large files and when input data is huge

Thank You