Clustering of EPB using ROTI Keograms\* (use style: paper title)

\*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

*Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (*Abstract*)

Keywords—component, formatting, style, styling, insert (key words)

# Introduction (*Heading 1*)

Equatorial plasma bubbles are of significant interest to researchers in the field of space weather, since they have a significant impact on the functioning of satellite navigation and communication systems. The global goal of space weather is to learn how to predict EPBs, including where they occur. From the point of view of studying EBP, different methods are used. Thus, a number of methods are based on the use of GNSS receivers and make it possible to determine the values of total electron content (TEC) and the time derivative of this parameter - the rate of change of the TEC index (ROTI).

The presence of a large number of TEC receivers over a large area is the most effective since PES are objects that occupy large areas in longitude and latitude, and therefore they are best observed on special two-dimensional maps, or on three-dimensional keograms. . Keograms also make it possible to estimate the time duration of ionospheric irregularities and EPB in particular. Researchers use keograms to analyze the ionosphere and can obtain information about the extent and duration of EPBs, conduct their initial analysis, and identify the types of different EPBs.

Identifying the types of EPB is important in the context of analyzing the causes of EPB and influencing the longitude of their occurrence, size and latitudinal scale, their duration, and dynamic changes in size over time. Understanding these relationships between the types of EBP and their causes among geophysical features will help to identify significant ones among these features and subsequently use them in predictive models as input data. Thus, the task of identifying EBP types is an important step towards solving the problem of predicting EBP. To identify types of EBP, it is necessary to solve the clustering problem, which implies the identification of stable and reproducible clusters on a sufficiently large dataset (several years of observations) and clearly separable clusters. Clusters must have sufficiently obvious structural differences to subsequently be used in the search for statistical relationships with geophysical parameters. This work aims to develop such a clustering algorithm and, as its main goal, sets the answer to the question: how many stable clusters can be identified based on the results of observing keograms for 2021-2023, what is the optimal number of them according to standard indicators for validating unsupervised machine learning problems. In addition, this work also answers the question – of what the optimal clustering algorithm and answers it as a result of a comparative analysis of several approaches.

# Methodology

## Features extractor

Pairs of ROTI keogram images serve as the initial data, representing latitude and longitude, respectively. These keograms were generated from observational data collected at GNSS stations, specifically, KMITL and RUTI stations. The dataset comprises observations from 2021, 2022, and 2023, totaling 925 image pairs. Due to the large size of each image, direct extraction is impractical, necessitating the use of feature extraction.

Given the relatively small dataset, neural network-based methods like autoencoders were deemed risky due to potential overfitting. Instead, the approach of statistical moments was adopted for feature extraction, ensuring representativeness regardless of dataset size. The method involves dividing the original image into N small squares, within which the first two statistical moments are calculated - an estimate of the mathematical expectation and an estimate of the dispersion.



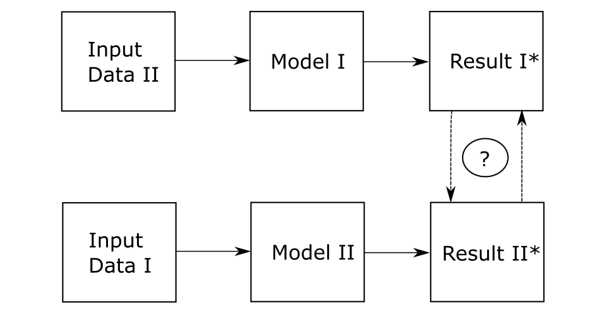
The resulting statistical features are normalized by the minimax method and concatenated into a feature vector:



Then, this vector is fed to the input of the popular clustering algorithm to the averages and the clustering procedure is performed directly

## Clustering accuracy criteria and optimization

As a criterion for clustering accuracy, it was decided to use the Silhouette Score, described in detail in []. This score is one of the classic ones for the clustering problem and allows you to estimate how similar objects within one cluster are to each other compared to objects from other clusters. The range of possible values for this score is -1 to 1, with 1 being the best value. Therefore, this speed must be maximized. The number of clusters n, as well as the size of the extractor window M, were used as parameters for optimization, and the goal of the optimization procedure was to determine their values at which the speed values are the best. In addition, it was decided to introduce one more, additional metric to assess the statistical reproducibility of clustering results



1. The initial sample is divided into two parts of comparable size without mixing

2. These samples are training samples for two independent models with identical architecture and identical hyperparameters

3. Models are trained on these samples and form reference labels

4. After training, the samples are swapped, the training sample of the first model is fed to the input of the second, the training sample of the second is fed to the input of the first

5. The label correspondence values are calculated using the scalar product principle and converted to percentages

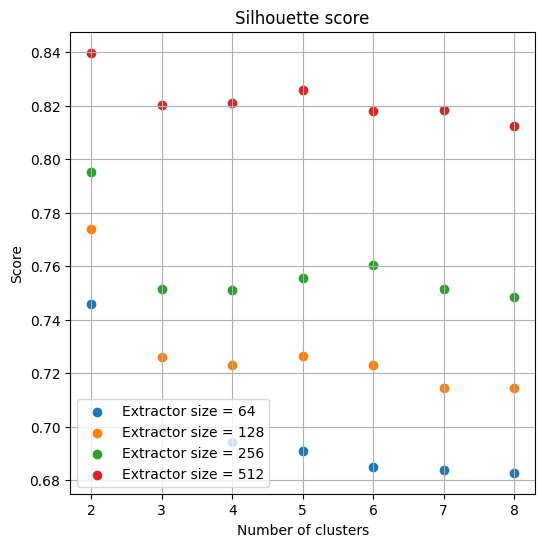
6. The experiment is statistical and steps 2-5 are repeated 100 times and the final values are averaged

This validation method is applied to two models - GMM and K-means - to select the model with the highest clustering quality

# Results

## Choose parameters for features extarcor

The first step was to conduct an analysis to find the optimal size of the feature extractor. The Silhouette speed of the clustering algorithm was evaluated for different extractor window sizes.



The figure shows characteristic dependencies for different cases of extractor window size. The highest speed value was obtained for the case when the window dimension is 512 by 512, i.e. The window size is equal to the image size. Thus, it is precisely this extractor parameter that is selected as optimal according to the criterion of maximizing Silhouette score.

|  |  |  |
| --- | --- | --- |
| Number of clusters | K-means | GMM |
| 2 | 0.641 | 0.577 |
| 3 | 0.624 | 0.587 |
| 4 | 0.620 | 0.616 |

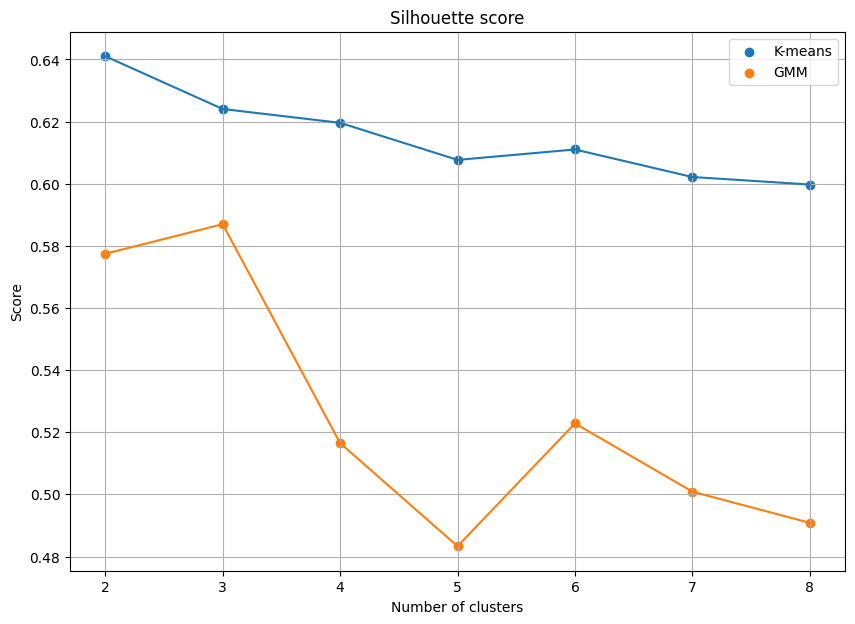
Davies-Bouldin score:

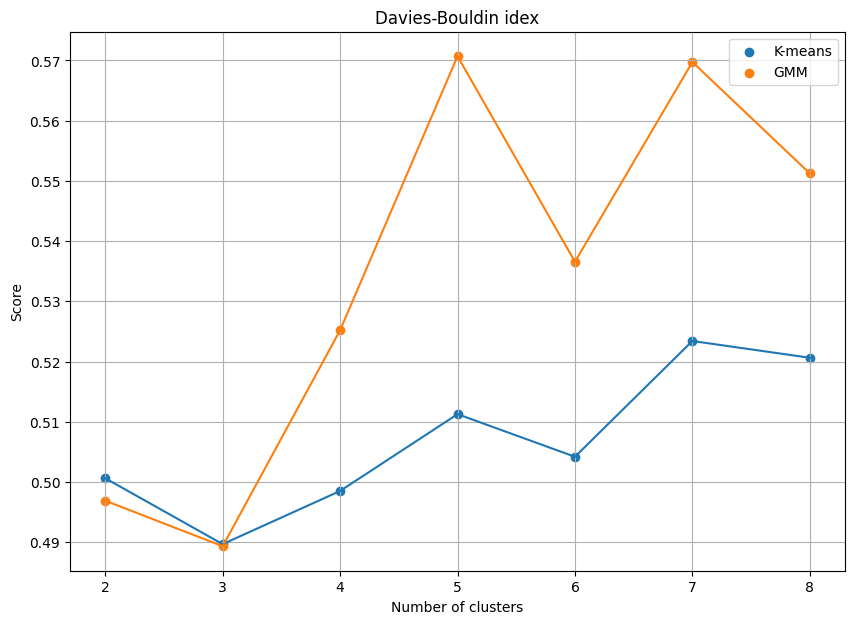
|  |  |  |
| --- | --- | --- |
| Number of clusters | K-means | GMM |
| 2 | 0.5006 | 0.4969 |
| 3 | 0.4897 | 0.4893 |
| 4 | 0.4985 | 0.5252 |

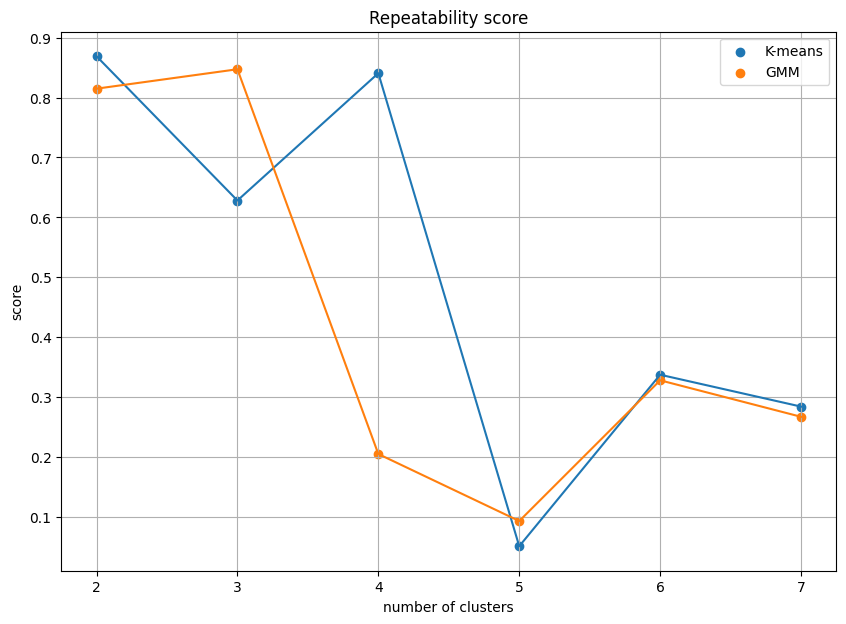
Repeatability score:

|  |  |  |
| --- | --- | --- |
| Number of clusters | K-means | GMM |
| 2 | 0.869 | 0.814 |
| 3 | 0.628 | 0.847 |
| 4 | 0.841 | 0.205 |

## Determining the optimal number of clusters





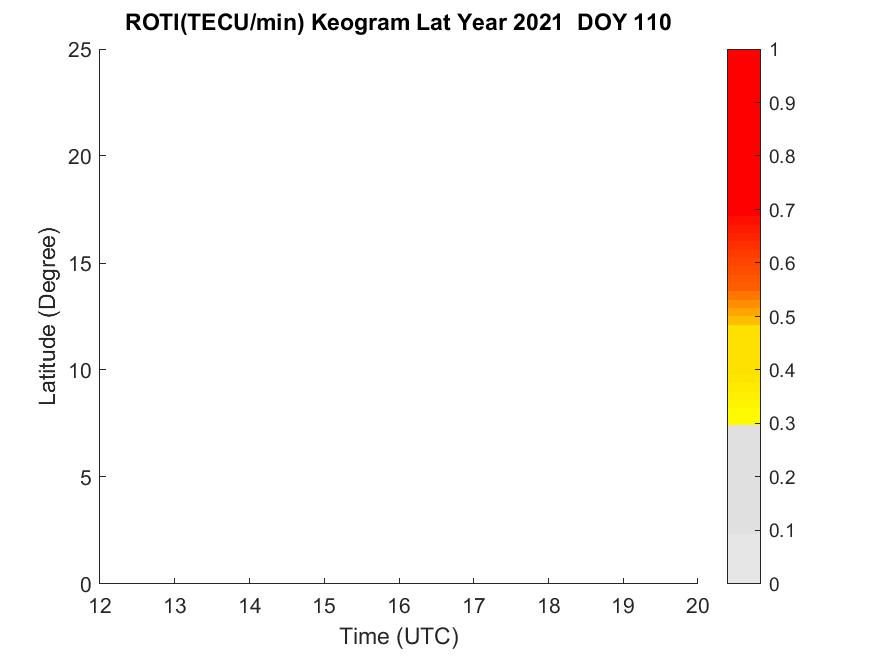


During a series of tests for two models, gmm (orange dots) and kmeans (blue dots), it was revealed that both models demonstrate high repeatability of the result when the number of clusters is three or four. At the same time, the accuracy of gmm with the number of clusters equal to three exceeds 90%, and the highest local maximum for kmenas is achieved with the number of clusters equal to four. Thus, we can conclude that a choice should be made between 3 and 4 clusters.

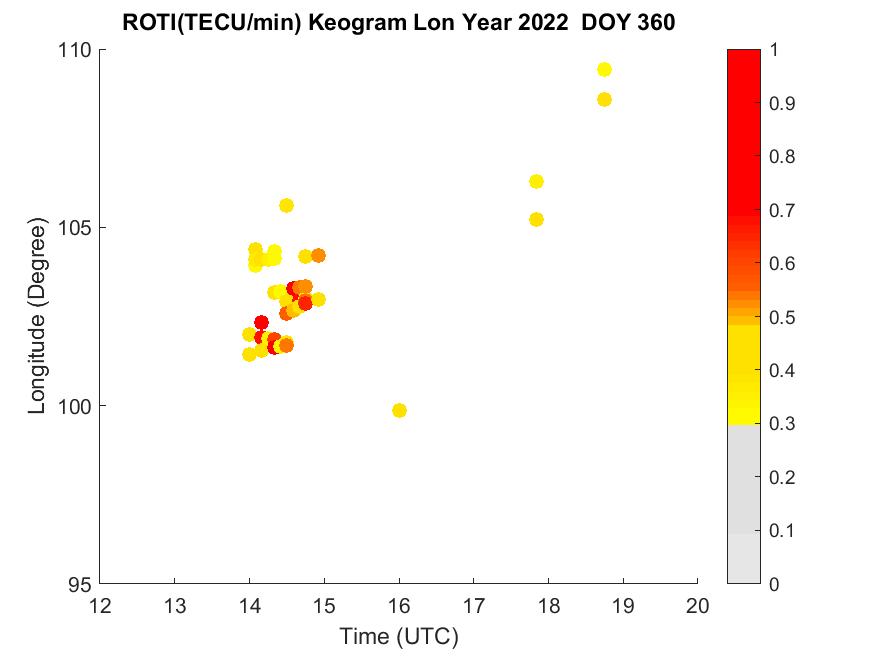
Since the rate is higher for the gmm model, we can choose it and conclude that exactly three clusters are the optimal number.

## Clusters

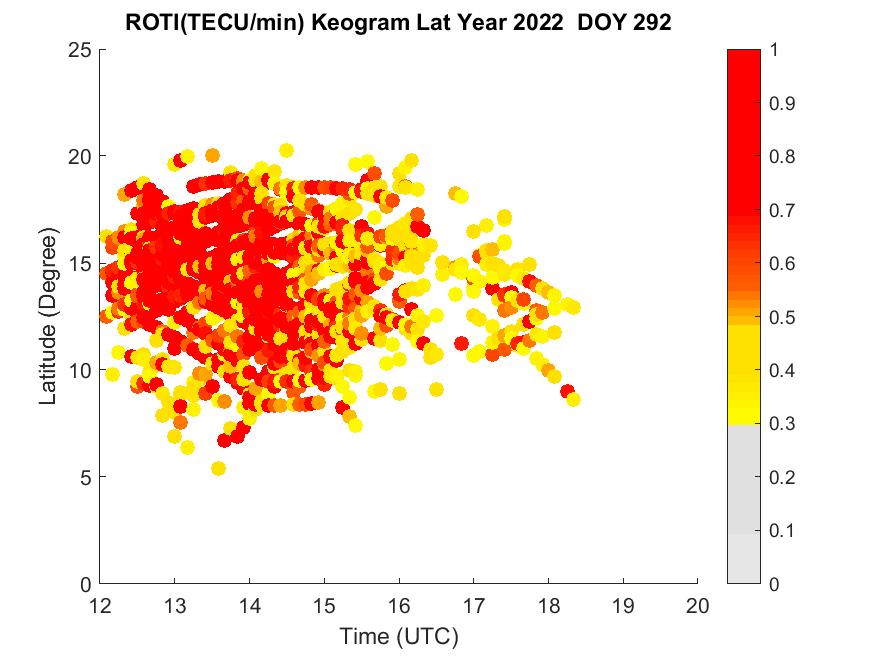
1. A cluster with a low ROTI value (few or no points) is a case when EPB is not observed. This may be a case of small irregularities that cannot be classified as EPB



1. The second case is an EPB localized in part of the observed area. It has a relatively small size.



1. Large EPB is an EPB that occupies a significant area of the keogram. This case is especially important since such phenomena are relatively rare.



|  |  |  |  |
| --- | --- | --- | --- |
| Year | Small  Irregularity | Medium EPB | Big EPB |
| 2021 | 25 | 28 | 1 |
| 2022 | 58 | 59 | 62 |
| 2023 | 37 | 63 | 94 |

# Conclusion

The keogram clustering challenge has been successfully addressed. Through the optimization of hyperparameters in the two considered models, it was discerned that there exist three distinct types of keograms, clearly expressed and reproducible in a statistical experiment.

The first type is characteristic of situations where ROTI values are small, and their distribution density on the keogram is low. This scenario indicates the absence of EPB observation (though irregularities may still be present). The second type encompasses keograms with relatively high ROTI values and a concentrated density of these values in a localized area, representing a typical EBP. The third type is characterized by a high density of large ROTI values across a broad localization area, signifying a large EPB with substantial dimensions in both longitude and latitude.

This outcome signifies the development of an algorithm capable of automatically labeling keogram data for subsequent applications. Specifically, it can be employed to address the challenge of predicting the occurrence of a specific type of EPB. Furthermore, these results can be utilized in the future not only to predict the type but also to forecast the longitudinal localization area for each EPB type.

the drop down menu to differentiate the head from the text.

##### References

[Put laterr)