Testing the validity of machine learning for ionospheric dynamics identification

Oscar Jackson

January 9, 2023

Abstract

As Earths magnetosphere interacts with the solar wind a process called magnetic reconnection occurs, this creates dynamics that can be indirectly observed within the ionosphere at the magnetic poles in the form of pulsed ionospheric flows. This project aims to test the feasibility of using machine learning methods to identify these events, in doing their rate of occurrence and repetition frequency could be determined, teaching us more about the nature of reconnection and flux transfer events within the Earth's magnetosheath. Several concepts are discussed and tested in this paper and the plans for further work are set out, preliminary tests have been positive with clustering being found to successfully split pulsed ionospheric flows into their own groups. Results have only bee possible due to a developed method of data processing for SuperDARN data.

1 Introduction

1.1 Physics and Instrumentation

Magnetosphere are a space where the solar wind is excluded by a planet's magnetic field) and due to this they are constantly changing systems effected by the solar wind, cosmic rays, and plasma source rates [MGK07]. Plasma in the solar wind enters the magnetosphere through a process called reconnection, when the magnetospheric field couples to the interplanetary magnetic field (IMF) here newly formed field lines curve and thus accelerate plasma away from the reconnection site allowing plasma to enter the magnetosphere and then the ionosphere.

This is part of a larger process called the Dungey Cycle, this process explains the interactions between a planet's magnetosphere and solar wind as a cyclic behaviour of magnetic reconnection. If day side (sun facing side of terminator) reconnection occurs at Earth, the solar wind transports magnetic flux to the night side, the path this area of reconnection (called a flux tube) crosses the polar cap, this flux must then return to the day traveling at latitudes below the polar cap, see Fig.1. As these processes carry flux they are called flux transfer events (FTE). The theory

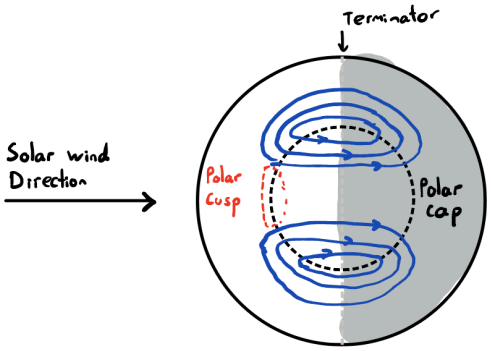

Figure 1: Diagram of the Dungey cycle, blue indicates the flow of flux. Un-shaded region is the dayside

of magnetospheric dynamics and in-turn recconection has benefited from the vast amount of data available from observations of magnetic planetary processes in the solar system that provide an

analogue of different laboratory conditions. These observations include data from satellites such as Cluster [Fea22], that pass over the polar cap, which give measurements of the components of the magnetic field in the magnetosheath, Fig.3 shows this behavior. These FTE generate dynamics in the magnetic field can be also measured in the ionosphere where the field lines effect the charged particles in the atmosphere, these signatures are known as pulsed ionospheric flows (PIFs), and have a one-to-one correlation with FTE [VOF+14].

These signatures are measured by coherent scattering radar such as that in the SuperDARN network, who's data is key to this project. Ionospheric scatter is measured by detecting back scatter from electron density structures (plasma irregularities) located in f and e region of ionosphere, see Fig.2. From this the power (signal to noise ratio), velocity and spectral width of the irregularities can be measured. If however the radar beam undergoes total internal refection it will move through ionosphere and travel to the ground (electron density in ionosphere is high enough), where it scatters in all directions, this is called ground scatter.

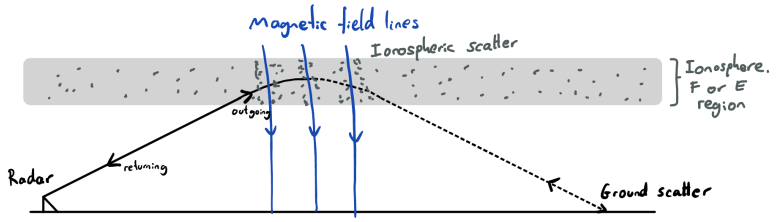

Figure 2: Diagram of the operations of a superDarn coherent scattering radar.

1.2 Machine Learning

Artificial intelligence is the computer science field of developing computer programs/algorithms that perform tasks that typically require human intelligence, such as problem-solving. Within this are machine learning models, this is when computer algorithms are trained to achieve specific goals without being programmed with explicit instructions on how to achieve them. They have the ability to find unknown patterns in data that can relate to a categorization of such data, an event or signal could be identified from all the background data.

In this project the goal is to develop a model that can identify particular ionospheric dynamics (more precisely PIFs) within the superDarn radar data, the model would be able to produce a list of labels that correspond to PIFs when inputted with data. There are two ways in which we could produce such a model, the first being supervised machine learning in which we train an algorithm of some example data with pre-determined labels. However, if you do not have this (which is true in our case, as the raw data is not labelled) then the second method can be useful, un-supervised machine learning uses the features of the data to make its on guesses on how it should be categorised.

2 The Problem

Time dependant nature of magnetic reconnection due to solar wind, cosmic influences and the complexity of solar wind magnetosphere interaction means that PIFs are hard to identify with simple statistical cutting methods (e.g. the do not always appear in the same position). Ultimately the goal is to develop a metod to identify the PIFs from all the noise and other ionospheric dynamics in radar data collected by the superDarn network.

Also called poleward moving auroral forms (PMAFs), PIFs are characterised by negative velocity plasma irregularities moving poleward (away from the radar) see the red streaks in Fig.7 and the arrows in Fig.3. Finding boundary's within the feature space (where in this case the features are power, velocity and spectral width) using machine learning will allow a model to predict the classification of measured ionospheric signal. Once they are identified, statistical studies (such by [PY99] and [MYP00]) can give more incites to the nature of FTE and their impact on the Dungey Cycle.

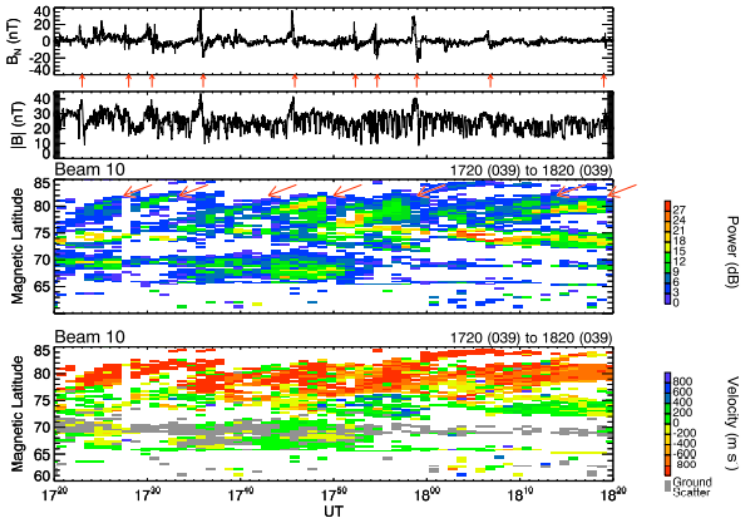

Figure 3: Time series of observations of FTE made by Cluster 3 (magnitude of the magnetic field in the magnetosheath) and corresponding PIFs observed by Prince George radar (ionospheric scattering), taken from [FTCM17].

2.1 Current Research

Currently most studies of pulsed ionospheric flows are identified manually (visually) beam by beam in hand-picked data where they are expected to occur (such as the polar cusp, see Fig.1) [PY99] and In most cases data from satellites are used to locate FTE which narrows down the search for the corresponding ionospheric dynamics. Manual identification of PIF signatures for statistical study is however tricky as SuperDARN network data consists of multiple radars each of which contain several beams directed in different directions, any of which could contain PIFs.

In other cases Fast Fourier Transforms where applied in an attempt to develop a more quantitative method [MYP00], however this was still done on hand picked data, the FFT where applied to time series data of each range gate and where able to identify PIFs. The statistical distributions of the PIF repetition periods in this paper agreed with satellite observations by [LW93] and optical observations of [FAS95].

As it can be seen there is much room for development in the identification of PIFs.

3 Current Progress

3.1 Data Processing

FITACF files are produced from RAWACF files, they are fitted data from the SuperDARN ACF fitting algorithm. Even though the data has been processed it still needs to be converted into a format that machine learning can take advantage of, this being a data set denoted by X containing individual events denoted by x each with N features. The features that we use are important to the result of the machine learning, currently the features chosen are: power, velocity, spectral width, position and time.

Position data is a topic of discussion as there are many reference frames that could be used:

- Beam dependant distance BDD: It it possible to analyse each beam of each radar individually, then it would be a one dimensional distance of only the range-gate distance (distance to radar)
- Radar centred coordinates, RCC: instead you can consider every beam together using the azimuth angle of the beam and the range-gate distance, allowing all the radar data to be fed into one model
- Altitude-Adjusted Corrected Geomagnetic Coordinates [She14], AACGM: the final step would be to convert every data point from RCC to AACGM this would allow data from all the radar stations to be fed into one model.

The reference frame is important when training our models as it is fundamental to the classifications that would be made, a model could be trained on beam dependent data it may be able

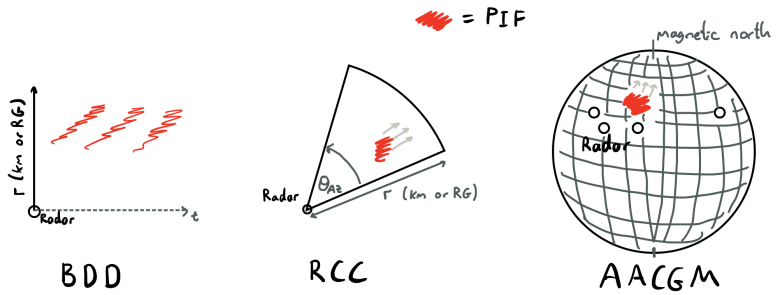

Figure 4: Diagram showing the possible reference frames used to train the machine learning models, light grey arrows show the movement of the PIFs over time, expected by the theory

to identify PIFs however the distance parameters used may mean the other important information on the global position of PIFs would not be used directly in the training, unlike what AACGM would allow for predictions to be made for certain AAM coordinates that correspond to the polar cusp, motivating the theory and confirming the models validity.

Currently Python scripts have be developed independently that convert data from FIACF files (read by pydarn) into the needed X dataset format using RCC, as current pyDarn code is not sufficient to provide pre-processed data ready for machine learning use, it can only process the read FITACF files and also plot the results. It may be worth if this research is continued to add a function in the pydarn library that does this. It is also possible to preform pre-cutting on the data or even filter in an effort to de-noise it, efforts have been made to test this by applying Gaussian filters but the complexity of the data makes it hard to do this and thus is not resulting in the required behaviour. Pre-cutting is being considered however more research needs to be done on this. Code mentioned in this section is available at https://github.com/orangeduice

3.2 Clustering

An unsupervised learning method that has shown real process is clustering, this involves dividing a dataset into groups (clusters) based on the patterns in the data. Data is split the into groups such that data points within the same group (cluster) have similar properties, while data points in different groups have relatively different ones. Clustering can be used for a wide range of applications, such as data compression, anomaly detection, and is often used as a prepossessing step for other machine learning tasks such as classification, which makes it a clear candidate for this project. (The main method currently tested is K means, definition in the appendix 6.1)

Both BDD and RCC have been tested with k means clustering, both showing promising results. For BDD the data was taken from the range time plotting function in pydarn, beam 10 was used as done by [FTCM17] so that the results can be directly compared to the PIF identified in the paper, see Fig.3&7. Due to the nature of the function only one parameter could be exported if a velocity or power, first the response when using the power vaules was studied. A range of cluster numbers were chosen when fitting the clustering algorithm, 5, 6, 8, 10, 20, 40, the results can be seen in Fig.5. 6 and 8 clusters identify the PIFs into 3 groups, however some outlier points have also been grouped with them, looking at the higher cluster numbers the PIFs are not identified separately from the rest of the data this seems to be a limitation of case means clustering within this reference frame.

The response when fitting to the velocity data was vastly different, most likely due to the larger range of values having a stronger effect on the clustering. It can be seen in Fig.6 that high cluster numbers the method completely breaks down as the results become too noisy, but with lower cluster numbers the response is much better with ground scatter being consistently identified separately from the dynamics of interest and the PIFs themselves being more clearly separated as one group but with more noise.

A new method was then developed to now use both power and velocity but now also a RCC reference frame, now the data that was chosen was from a two-hour interval just before the interval used by [FTCM17] however the beginnings of PIFs were still present. The results seen in Fig.8 so a clear separation of ground scatter represented by purple while the dynamics of interests have been identified in blue, there is still noise present and some dynamics have been incorrectly grouped with the PIFs but it is an improvement over the earlier method. More research needs done on the relation between the features used for training and the results of clustering.

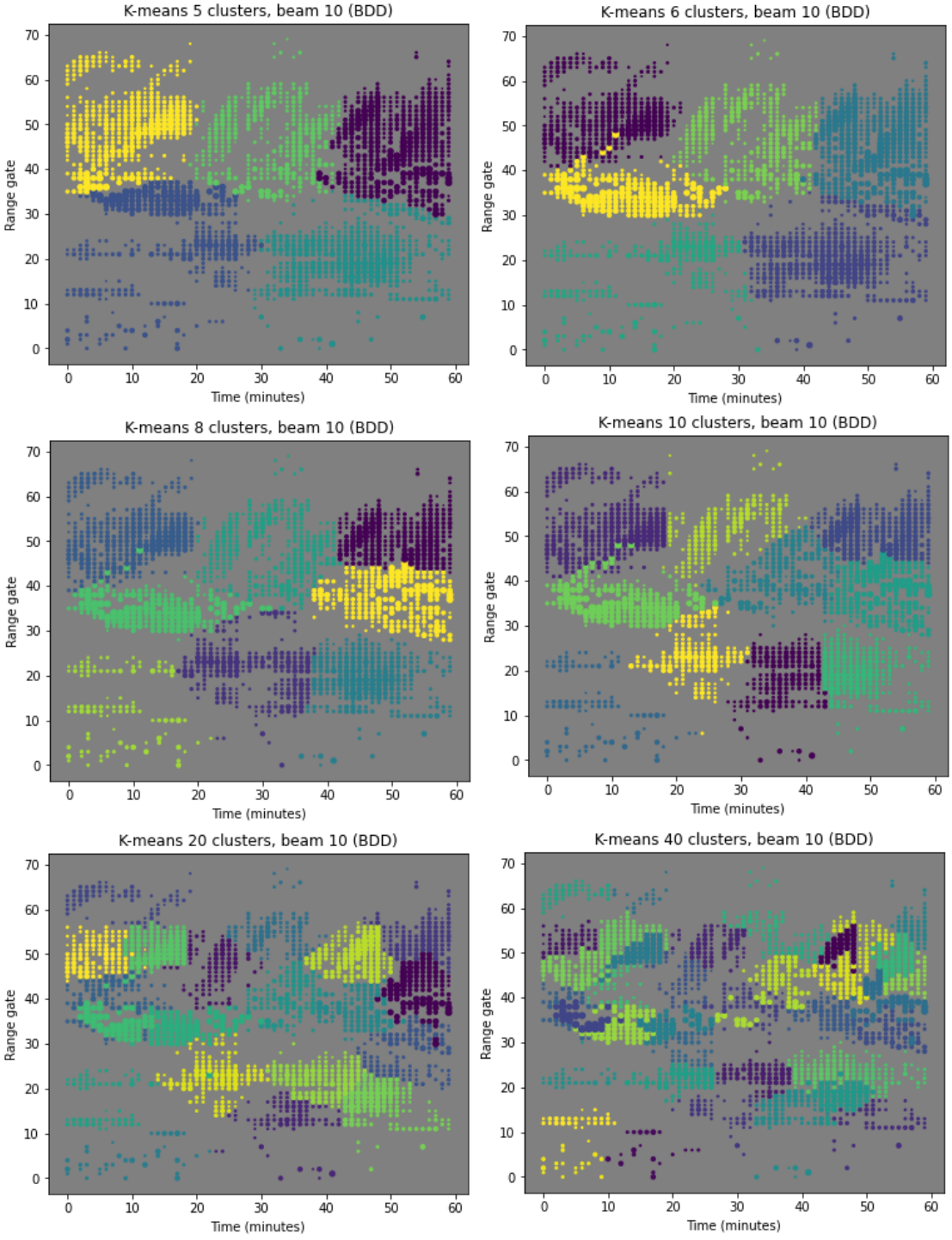

Figure 5: Grouped time-range plot with of SuperDARN radar data in the BDD reference frame by a selection of k mean algorithms (with different cluster amounts) trained on power, currently colors are randomised this is a WIP

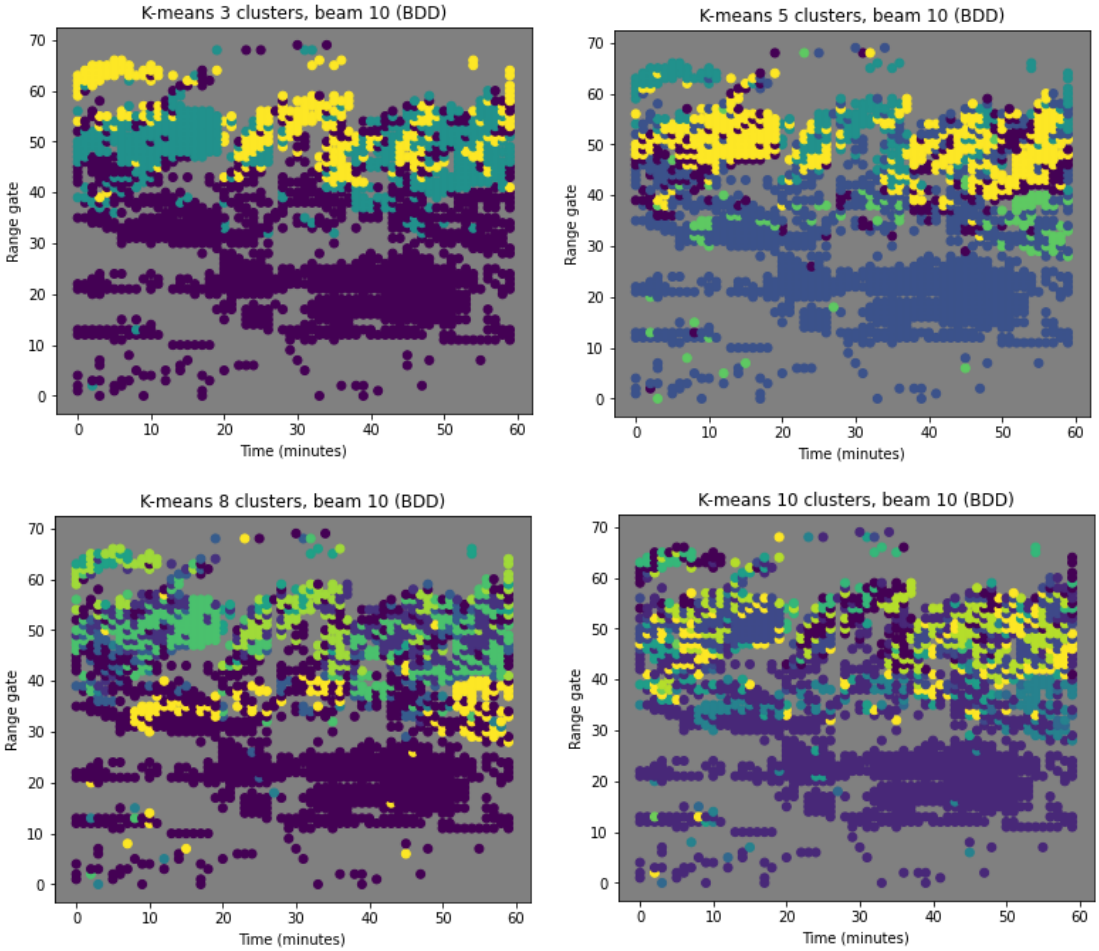

Figure 6: Grouped time-range plot with of SuperDARN radar data in the BDD reference frame by a selection of k mean algorithms (with different cluster amounts) trained on velocity, currently colors are randomised this is a WIP

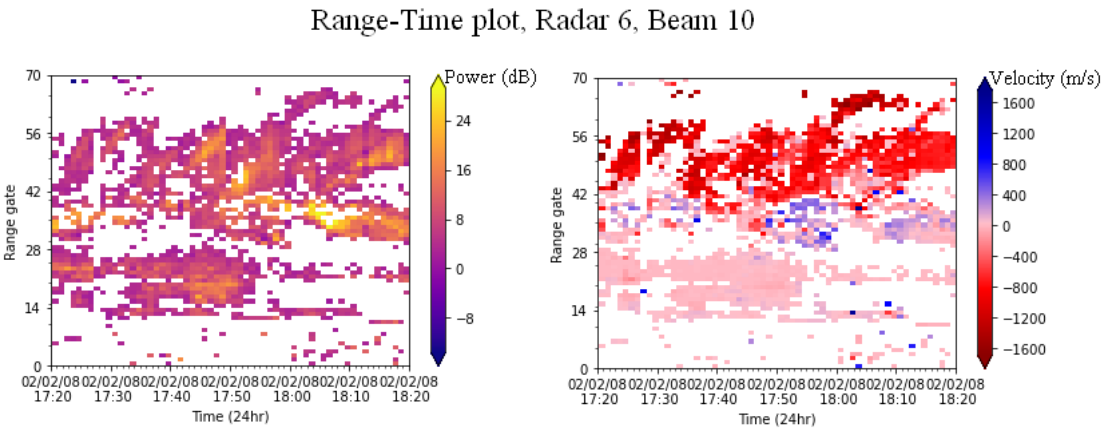

Figure 7: Time range plots from the same time period of Fig.3 of power and velocity

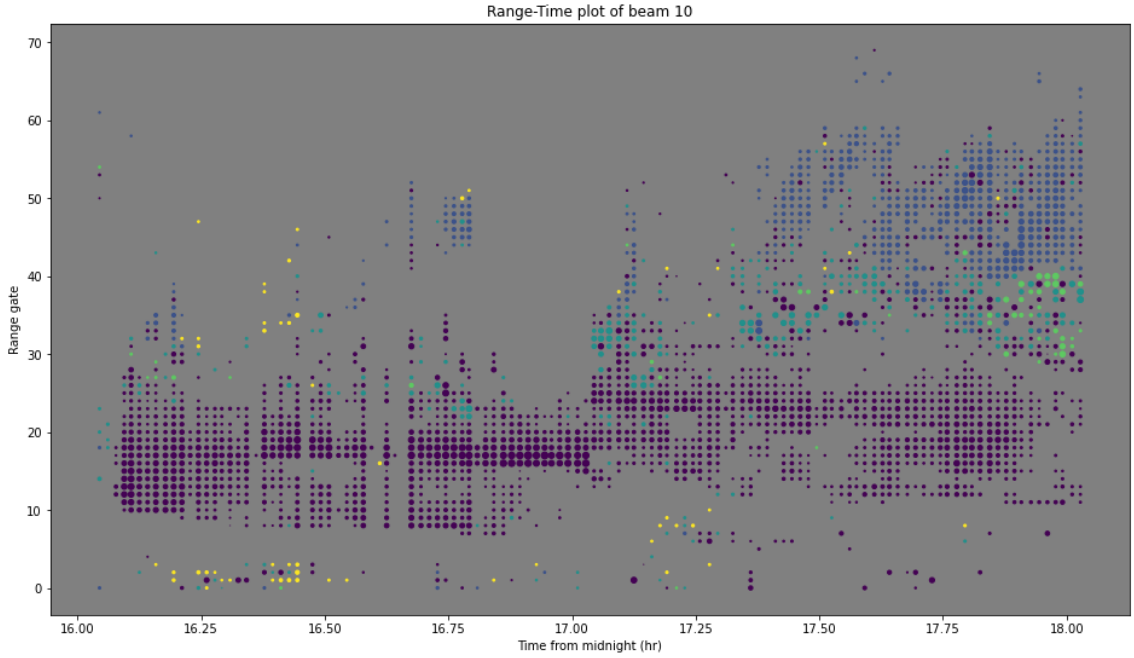

Figure 8: Grouped time-range plot with of SuperDARN radar data in the RCC reference frame by a k mean algorithm of 5 clusters

4 Future Plans

4.1 Auto Encoder

In a effort to de-noise and simply the input data an autoencoder can be trained, they are designed to learn a lower dimensional more compact representation of unlabelled input data. Autoencoders are made from two components: an encoder, which maps the input data to this compact representation, and a decoder: that maps the compact representation back to the original data space. Implemented as a neural network with a bottleneck of neurons, the goal of training to learn the weights of these networks that will reconstruct the data (from the compact version) as similar as possible to the original. They are often used for dimensionality reduction, data denoising and anomaly detection, which could improve the response of other methods such as clustering.

4.2 AAML Coordinates

To improve the clustering algorithms and to provide more data for a single machine learning model to train on I will introduce altitude adjusted magnetic coordinates and combine data of multiple radar stations. AACGM will allow crosschecking with our understanding of the polar cusp and the Dungey Cycle, for example the model should be able to predict that dynamics near the polar cusp are more likey to be PIFs.

4.3 Supervised Machine Learning

Unlike un-supervised, supervised learning is highly accurate and trustworthy method, for full pulsed ionospheric flow identification it will be necessary to develop such a method. New pipeline must create labelled datasets from the raw data, this could be done manually with the PIFs labelled by hand or be done using clustering algorithms and some user input (selection of clusters to be labelled as signal). This could feed into to a feedback loop which improves the clustering algorithm that in turn improves the un-supervised classification.

4.4 Timeline

Current progress is good and the next steps are clear, what follows is a timeline of the next steps to test the feasibility of developing an effective machine learning algorithm.

1. Improving clustering algorithms by finding best hyperparameters (e.g. grid search) *Main Goal*
2. AACGM coordinate testing to combine data from multiple radars *Optional Goal*
3. Autoencoder testing *Main Goal*
4. User labelled data testing *Main Goal*
5. Unsupervised learning testing *Main Goal*
6. Final system development *Stretch Goal*
7. Testing classification *Stretch Goal*
8. Comparison with other statistical surveys *Stretch Goal*

5 Conclusion

Overall, the preliminary testing has been positive, clustering have shown that these PIFs can be separated from the rest of the ionospheric dynamics even with a limited amount of data. BDD and RCC reference frames have shown to be applicable for this application of machine learning, with the improvements of RCC showing the need for more features to be used in the future. There are several ongoing issues with the complexity of radar data and the complication of different operation modes and frequencies this could introduce unwanted bias and will be an ongoing issue throughout the project. More work needs to be done on what time scales should be used for the time parameter, as it has shown to affect the clustering results (when time is defined in such a way that it does not change much over the appearances of PIFs the algorithm struggled to cluster the data correctly).

The feasibility of machine learning has been tested and everything points to its potential in this task, this can be developed into deep learning due to the large amount of data available (2 solar cycles). Current progress is also showing that pre-processing is incredibly important to the results and getting the data into a state where machine learning can be applied should be imperative, a streamline data processing system should be set up so that all data can be used, as the more data we have the better the models will be.

Continuing with this project I hope to develop a pre-processing system to convert all radar data into AACGM format so that I can experiment with training supervised and unsupervised machine learning methods with as much data as possible, included in this will be looking at the potential of autoencoders. Current results show a model can be made to identify PIFs and if this is done meaningful comparisons can be made with other statistical surveys (such as ones done by [PY99],[PY99],[LW93] and [LW93]) to validate the results.

6 Appendix

6.1 K-means clustering

K-means clustering is an unsupervised learning algorithm that is used to spit a dataset into K clusters, where each cluster is represented by its own centroid (mean). The algorithm works by assigning each data point to the nearest centroid, and then recomputing the centroid for each cluster based on the data points that are assigned to it. This process is repeated until the centroids no longer move to a determined degree. The number of clusters must be specified in advance. K-means is simple to implement and can scale to large datasets.

References

[FAS95] GJ FASEL. Dayside poleward moving auroral forms - a statistical study. *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, 100(A7):11891–11905, JUL 1 1995.

[Fea22] R. C. Fear. Joint cluster/ground-based studies in the first 20 years of the cluster mission. *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, 127(8), AUG 2022.

[FTCM17] R. C. Fear, L. Trenchi, J. C. Coxon, and S. E. Milan. How much flux does a flux transfer event transfer? *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, 122(12):12310–12327, DEC 2017.

[LW93] M LOCKWOOD and MN WILD. On the quasi-periodic nature of magnetopause flux-transfer events. *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, 98(A4):5935–5940, APR 1 1993.

[MGK07] Fran Bagenal Margaret Galland Kivelson. Encyclopedia of the solar system. Planetary Magnetospheres:519–539, 2007.

[MYP00] KA McWilliams, TK Yeoman, and G Provan. A statistical survey of dayside pulsed ionospheric flows as seen by the cutlass finland hp radar. *ANNALES GEOPHYSICAE-ATMOSPHERES HYDROSPHERES AND SPACE SCIENCES*, 18(4):445–453, APR 2000.

[PY99] G Provan and TK Yeoman. Statistical observations of the mlt, latitude and size of pulsed ionospheric flows with the cutlass finland radar. *ANNALES GEOPHYSICAE-ATMOSPHERES HYDROSPHERES AND SPACE SCIENCES*, 17(7):855–867, JUL 1999.

[She14] S. G. Shepherd. Altitude-adjusted corrected geomagnetic coordinates: Definition and functional approximations. *JOURNAL OF GEOPHYSICAL RESEARCH-SPACE PHYSICS*, 119(9), SEP 2014.

[VOF+14] A. Varsani, C. J. Owen, A. N. Fazakerley, C. Forsyth, A. P. Walsh, M. Andre, I. Dandouras, and C. M. Carr. Cluster observations of the substructure of a flux transfer event: analysis of high-time-resolution particle data. *ANNALES GEOPHYSICAE*, 32(9):1093–1117, 2014.