# Unsupervised Clustering of Magnetospheric Dayside Data

James “Andy” Edmond[1], Joachim Raeder[1], Banafsheh Ferdousi[1], Maria Elena Innocenti[2], Matthew Argall[1]
1: University of New Hampshire, Durham, NH, USA | 2: Ruhr-Universität Bochum, Bochum, NRW, Germany

**Contact Email:**
james.edmond@unh.edu

## I. INTRODUCTION

Earth’s dayside magnetosphere is composed of several regions, such being the magnetosphere, the magnetosheath, and the solar wind, and their boundaries, the magnetopause and the bow shock. Individually identifying crossings from data between these regions is often easy but time-consuming, and so it would be beneficial to automate this process.

We perform this automation on dayside THEMIS data via an unsupervised machine learning clustering algorithm, the gaussian mixture model (GMM)[1]. Once the data is clustered, magnetopause and bow shock boundary crossings can be inferred from one of two instances.

**Instantaneous crossings:** The spacecraft immediately crossed a boundary
**Gradual crossings:** The spacecraft entered a region where data is clustered as a mix of the two adjacent plasma regions

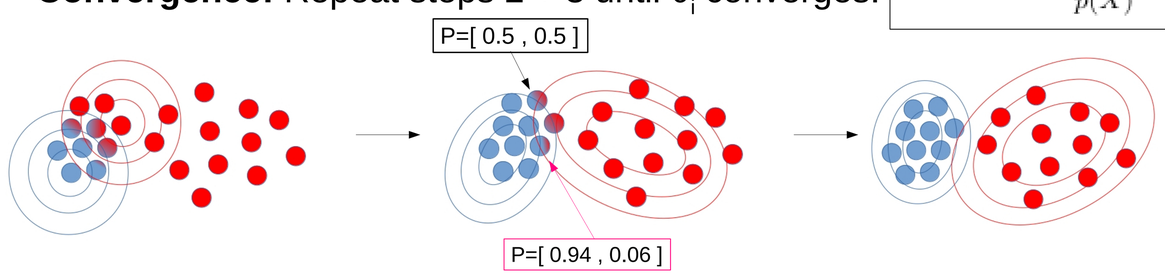
## II. METHOD

While others[2,3,4] have automated the classification of magnetospheric data into plasma regions, the methods used thus far have used supervised learning. One consequence of this is that such methods need to be retrained for every mission’s dataset. To avoid this, and given the recent success of a similar endeavour[5], we use an unsupervised clustering method, the gaussian mixture model.

**QUALITATIVE DESCRIPTION:**
In essence, one provides **X** data and an integer *k* indicating how many gaussians to decompose **X** into. The algorithm then “learns” the proper parameters (e.g. **μ**, **Σ**, and π) for each distribution so as to best describe **X**

**QUANTITATIVE DESCRIPTION:**
More explicitly, the method performs the following steps:
**0 – A Priori Assumption:** All data **X** is assumed to be generated from a linear combination of normal distributions $\{N_i\}$, each with mean vector $\boldsymbol{\mu}_i$, covariance matrix $\boldsymbol{\Sigma}_i$, and mixture scalar $\pi_i$ (these three parameters are collectively referred to as $\theta_i$).
**1 - Initialization:** Make initial guesses for $\theta_i$ for each distribution
**2 – Expectation:** Compute the bayesian posterior probabilties $p(\theta_i \mid X_q)$, i.e. probability that normal distribution $N_i$ is correct given data point $X_q$
**3 – Maximization:** Compute parameters $\theta'_i$ that maximize the likelihood $p(X_q \mid \theta_i)$ i.e. probability that data point $X_q$ is sampled from distribution $N_i$. Then update $\theta_i$ --> $\theta'_i$.
**4 – Convergence:** Repeat steps 2 + 3 until $\theta_i$ converges.

$$p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}$$




This algorithm produces a vector of probabilities for each data point, indicating confidence in cluster assignment
It should be noted that this algorithm depends on randomized initial conditions; it’s recommended to make multiple runs and preserve the best fit.

## III. SOURCE DATA

The source data is from the THEMIS Mission (Time History of Events and Macroscale Interactions during Substorms) at a variety of resolutions: ESA-reduced (~3 secs), ESA-full (~90 secs), and the on-board computed moments MOM (~3 secs).

- ESA-reduced is drawn from primarily, with gaps supplemented by ESA-full and MOM
- Runs from March 2007 to Dec 2020
- Reduced to 1-min resolution.

The input parameters for the clustering model are (8 in total):

- Bx, By, Bz (nT – GSE)
- Ion Vx, Vy, Vz (km/s – GSE)
- Log-scale ion density (#/cc)
- Log-scale ion temperature (eV)

Want to constrain data to dayside and avoid close-in magnetosphere, so only keep data where…

- $|\mathbf{B}| <= 200$ nT
- $5\ R_E < R < 30\ R_E$
- $X_{GSE} >= 0$

Overall, leaves 9.3M pts across 5 spacecraft (THEMIS-A,B,C,D,E); or 4.6M each for training and testing after a 50/50 test-train split.

## IV. CLUSTERING RESULTS

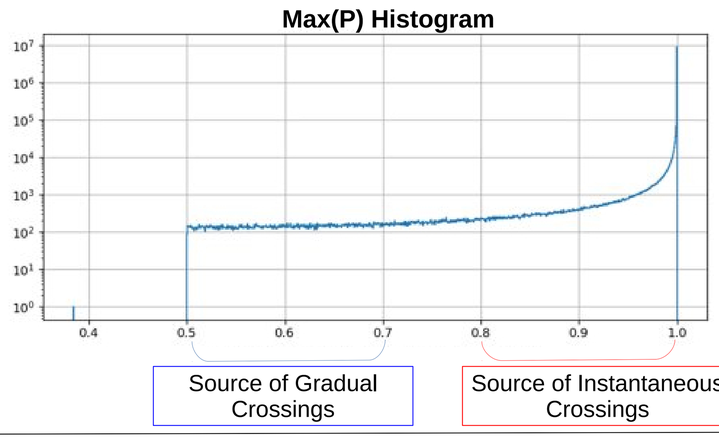
Once clustered, we can analyze two sets of data:

- The inter-cluster information (how the data was clustered as a whole - see at right)
- The intra-cluster information (what are the properties of a *particular* cluster – see below)

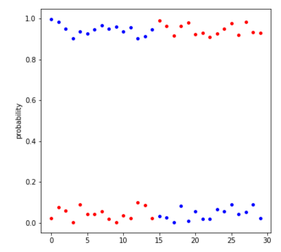
**Max(P) Example**

$$P = \begin{matrix} C1 & C2 & C3 \\ \{[0.93, & 0.07, & 0.00]\} \\ \{[0.66, & 0.30, & 0.04]\} \\ \{[0.41, & 0.50, & 0.09]\} \end{matrix} \Rightarrow \text{Max(P)} = \begin{matrix} \{[0.93]\} \\ \{[0.66]\} \\ \{[0.50]\} \end{matrix}$$

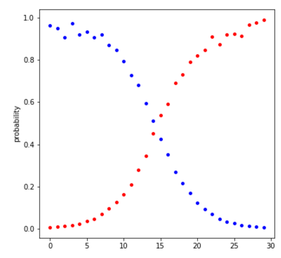
**Max(P) Histogram**



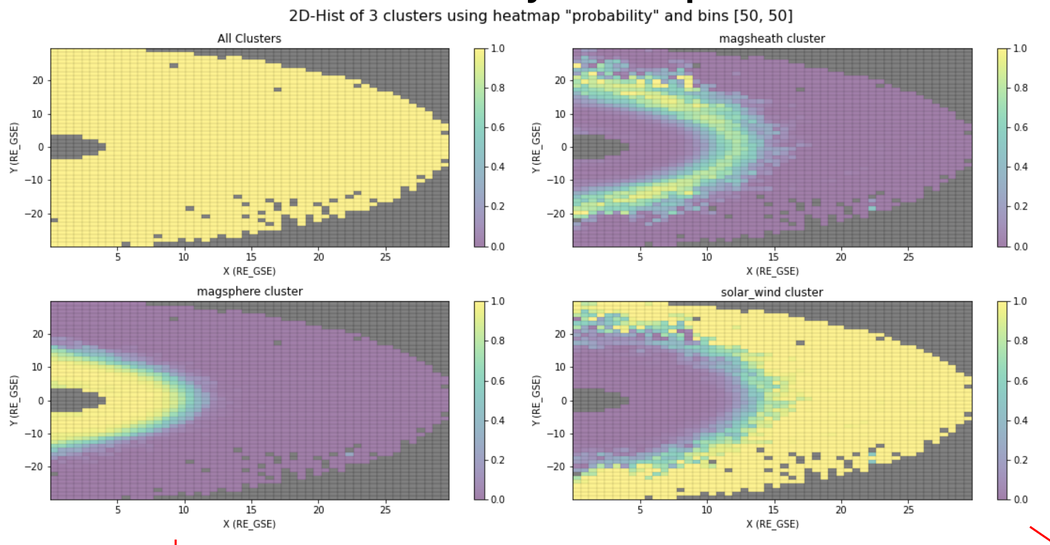

Instantaneous Crossing Example



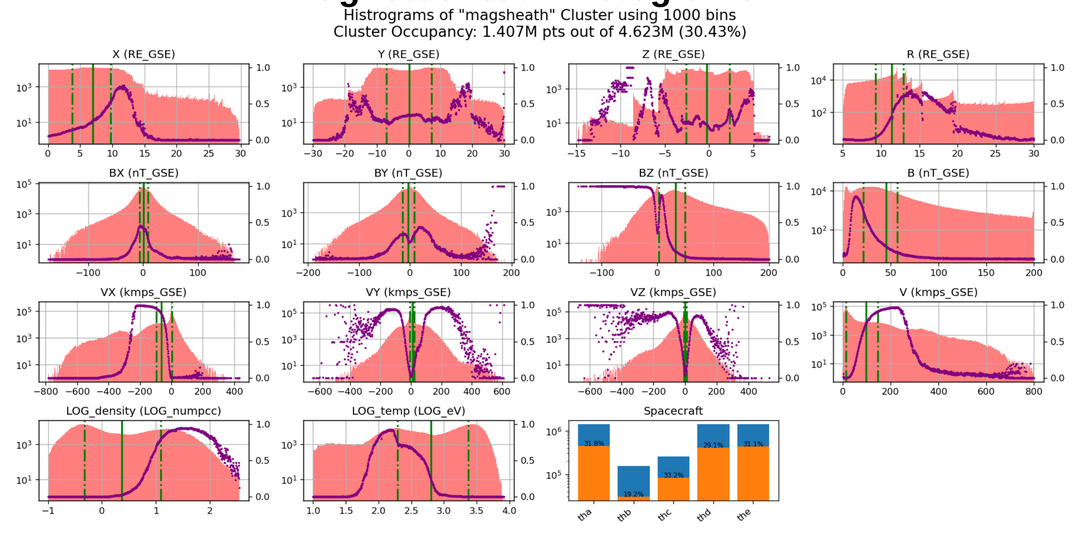
Gradual Crossing Example



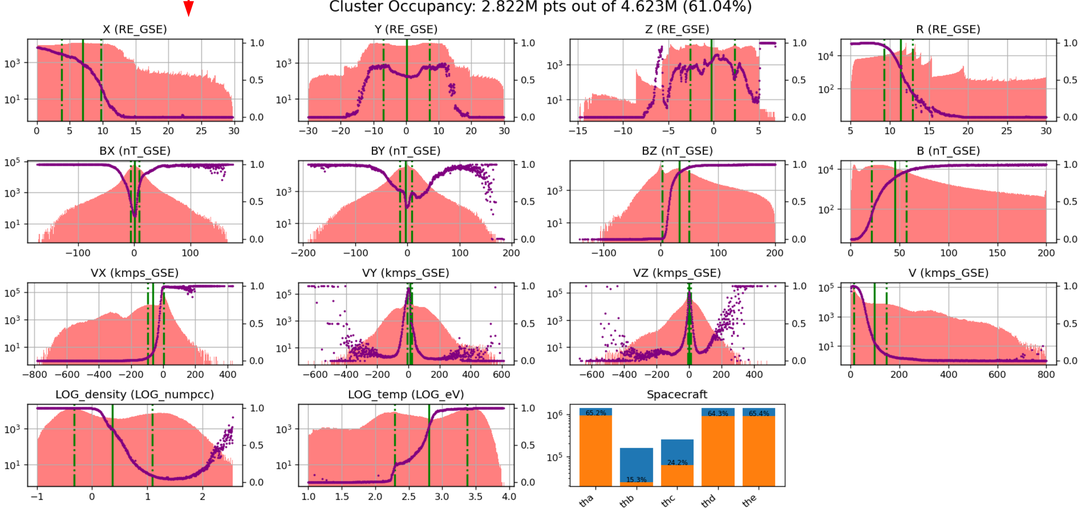
## Probability Heatmap



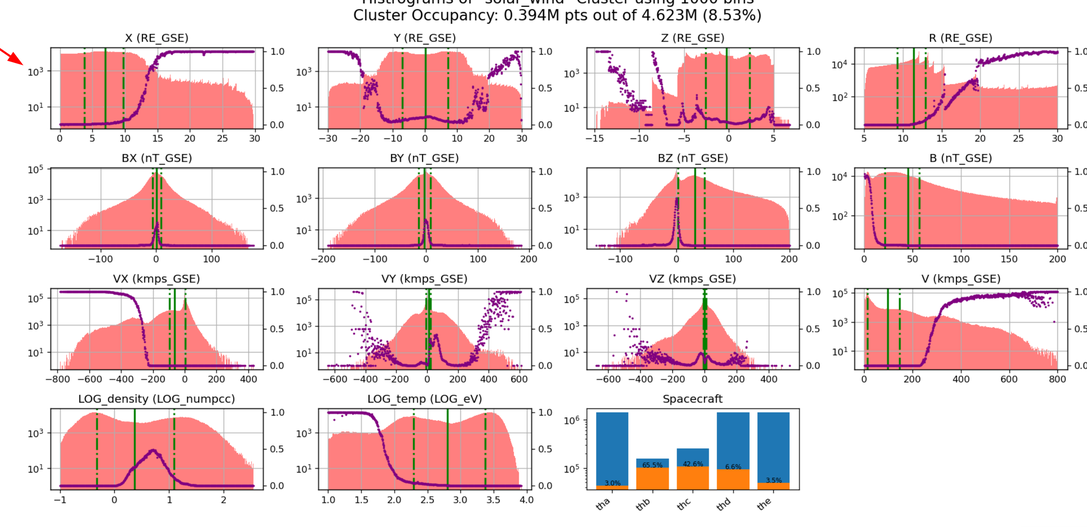

## Magnetosheath Histograms



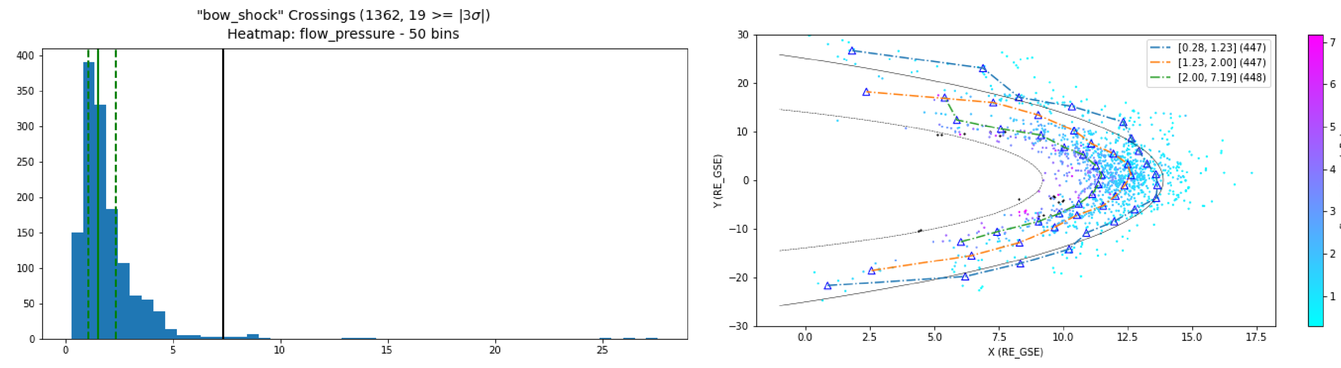

## Magnetosphere Histograms




## Solar Wind Histograms




Boundary crossings can be inferred from the probability time series between two adjacent plasma regions (see the instantaneous and gradual crossing examples above).
Specifically, crossings between the magnetosheath and the solar wind can be taken as bow shock crossings. These crossings can be paired with the at-crossing latest solar wind measurements of the spacecraft so that they can be analyzed graphically in relation to their upstream solar wind parameters.




## V. Summary / Future Work

Using the gaussian mixture model, an unsupervised clustering algorithm, we’ve clustered magnetospheric dayside THEMIS data into three regions: magnetosphere, magnetosheath, and solar wind. Analysis of inter- and intra-cluster information indicates a successful clustering. Bow shock crossings are found from the magnetosheath / solar wind separation, and a flow pressure heatmap of the crossings acts as a loose, qualitative form of verification.

We intend to incorporate more data from other spacecraft that frequent the dayside, such as MMS and Cluster, in the future, and we will use the large dataset of crossings they produce as the basis for a future bow shock model.

## VI. CITATIONS

1: Fabian Pedregosa *et al*,. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830, 2011

2: Hugo Breuillard *et al,* Automatic classification of plasma regions in near-earth space with supervised machine learning: Application to magnetospheric multi scale 2016–2019 observations. Frontiers in Astronomy and Space Sciences, 7, Sep 2020

3: Vyacheslav Olshevsky *et al.* Automated classification of plasma regions using 3d particle energy distributions. Journal of Geophysical Research: Space Physics, 126(10):e2021JA029620, 2021.

4: D. da Silva *et al,* Automatic region identification over the mms orbit by partitioning n-t space, 2020

5: M. E. Innocenti *et al*, Unsupervised classification of simulated magnetospheric regions. Annales Geophysicae, 39(5):861–881, 2021