(ATTEMPT AT)

Clustering on Starlink Satellite Data

Tiffany Phan, Keith Poletti, Qing Zhu

Overview

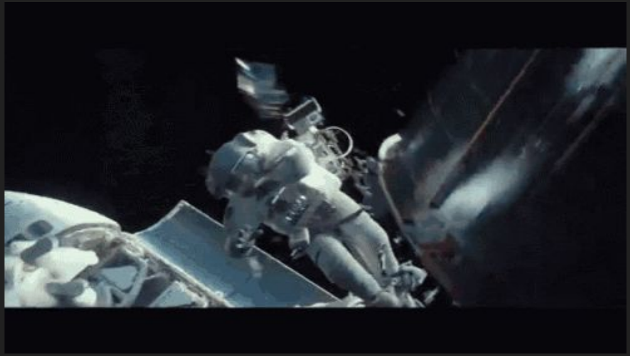
- Purpose
- Data
- Methods
  - *k*-Means
  - Time-Series *k*-Means
- Results
- Conclusions/Future Work

Starlink Satellites

- SpaceX-owned and -operated constellation of small satellites
- Primary purpose: to provide satellite Internet access across the world
- Currently over 2000 satellites in orbit with more planned
- Approximately 40-60 satellites launched every two weeks
- Launch cadence expected to increase

Starlink Satellites

- Concerns:
  - Orbit crowding and contribution to space debris
  - Space collision likelihood and the Kessler syndrome

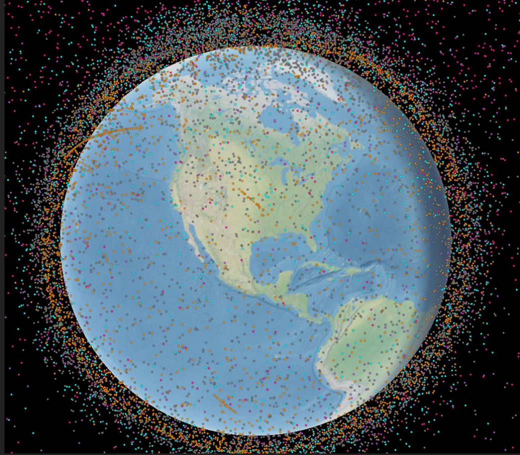


Large number of satellites

⇒ Satellite collisions & space debris

⇒ Kessler syndrome

Scope and Goals of Research

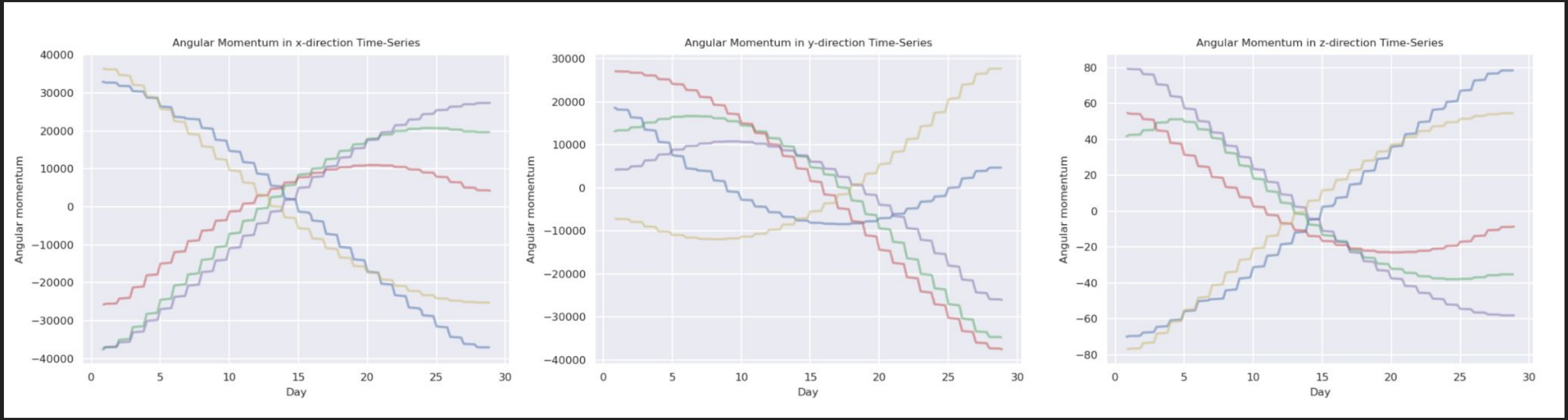

Snapshot from ASTRIAGraph

- Typical satellite behavior (modes of operation (MO)):
  - Initial deployment phase
  - Low-thrust, orbit-raising phase
  - Operations phase
  - End-of-life decay phase
  - Others?
- Goals of Research:
  - To identify the MO of a satellite in a quantifiable manner
  - To devise an algorithm to quickly and accurately determine the satellite MO
- Potential Applications:
  - Space traffic management
  - Autonomous maneuver detection

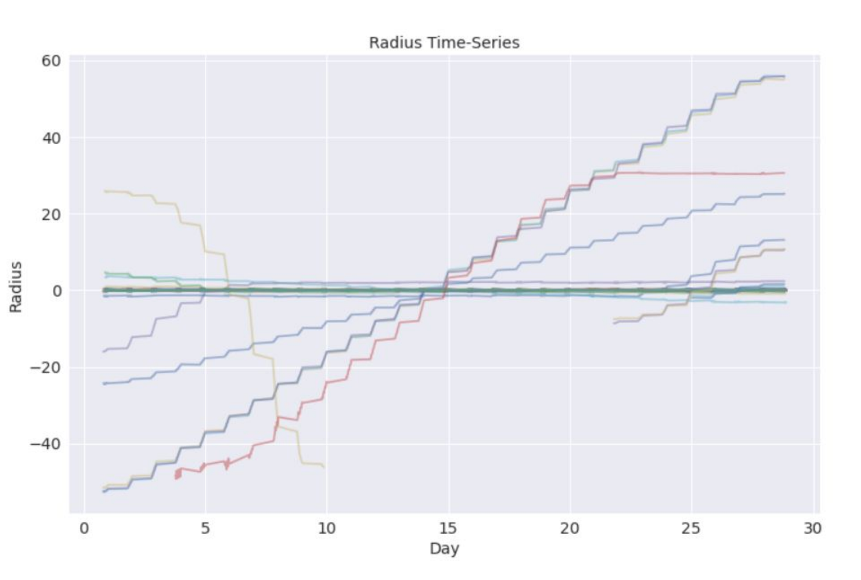
Scope and Goals of Research

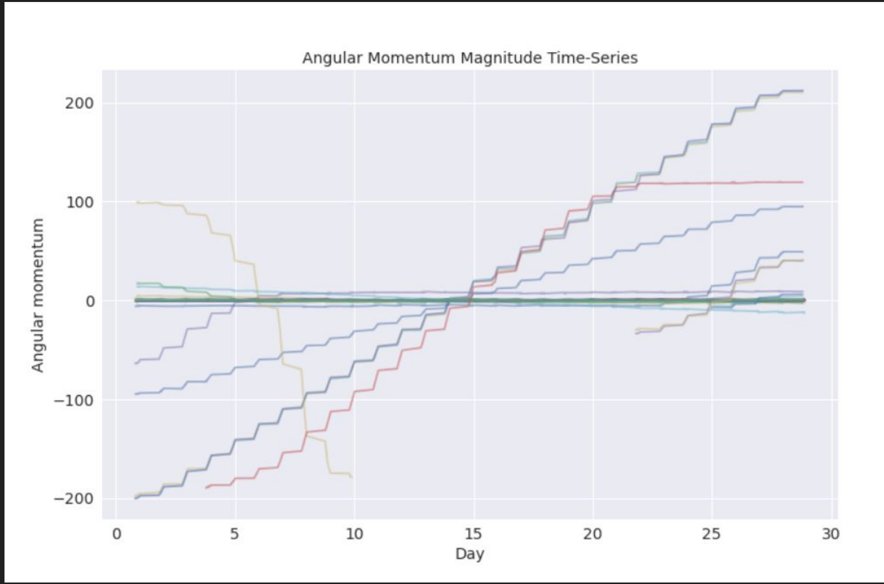
- Obstacles:
  - Ambitious goals
  - Processing of time-series data of about 2000 satellites
- Narrowing it down
  - Restrict goal to identifying MO
  - Restrict time series to one month (February 2022)
  - Cluster data and hope for the best

Sample Time-Series Plots

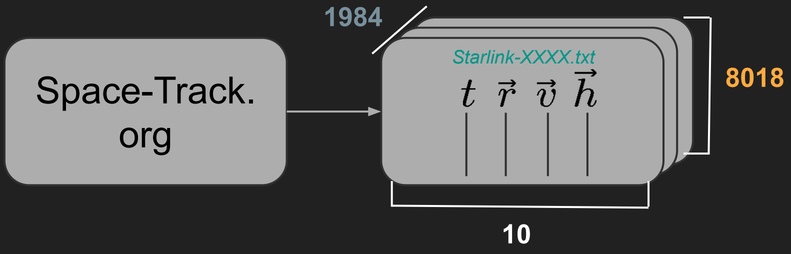


Sample Time-Series Plots



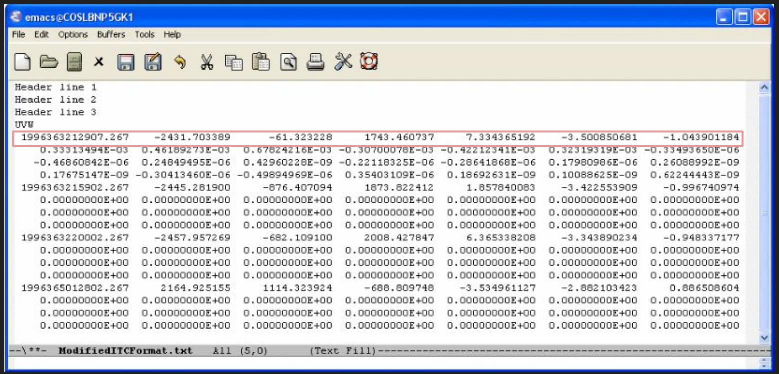


Starlink Data



- Satellite orbit data comes from space-track.org in single-day files
- First row is the ephemerides containing:
  - Julian-time
  - Position vector $\vec{r} = (r_x, r_y, r_z)$
  - Velocity vector $\vec{v} = (v_x, v_y, v_z)$
- Use the position and velocity to calculate angular momentum

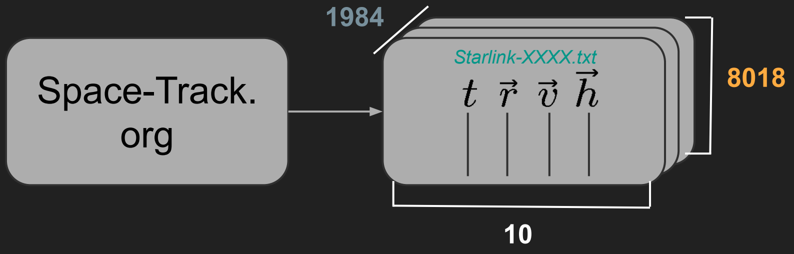
$$\vec{h} = \vec{r} \times \vec{v}$$




```
Header line 1
Header line 2
Header line 3
UVW
 1996363212907.267    -2431.703389     -61.323228    1743.460737     7.334365192    -3.500850681    -1.043901184
    0.33313494E-03  0.46189273E-03  0.67824216E-03 -0.30700078E-03 -0.42212341E-03  0.32319319E-03 -0.33493650E-06
   -0.46860842E-06  0.24849495E-06  0.42960228E-09 -0.22118325E-06 -0.28641868E-06  0.17980986E-06  0.26088992E-09
    0.17675147E-09 -0.30413460E-06 -0.49894969E-06  0.35403109E-06  0.18692631E-09  0.10088625E-09  0.62244443E-09
 1996363215902.267    -2445.281900    -876.407094    1873.822412     1.857840083    -3.422553909    -0.996740974
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
 1996363220002.267    -2457.957269    -682.109100    2008.427847     6.365338208    -3.343890234    -0.948337177
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
 1996365012802.267     2164.925155    1114.323924    -688.809748    -3.534961127    -2.882103423     0.886508604
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
    0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
```

Source: https://www.space-track.org/documents/Spaceflight_Safety_Handbook_for_Operators.pdf

Starlink Data




FEBRUARY 8, 2022

GEOMAGNETIC STORM AND RECENTLY DEPLOYED STARLINK SATELLITES

Preliminary analysis show the increased drag at the low altitudes prevented the satellites from leaving safe-mode to begin orbit raising maneuvers, and up to 40 of the satellites will reenter or already have reentered the Earth's atmosphere. The deorbiting satellites pose zero collision risk with

Source: https://www.spacex.com/updates/#sustainability

- Compile each Starlink's data for February
    - Why February?
        - Special events to help identify MOs
- Less data if recently launched or crashed
    - Pad with zeros or NaN's
- Combined final data matrix is:

Satellite x Time x Ephemerides

k-Means Clustering

- *k*-means

  Find a partition $C_1 \cup C_2 \cup \cdots \cup C_k = P$ and a set of means $\mu_1, \mu_2, \cdots, \mu_k \in \mathbb{R}^d$ such that the following objective is minimized:

$$\min_{C_1, C_2, \cdots, C_k} \min_{\mu_1, \cdots, \mu_k \in \mathbb{R}^d} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in C_i} \left|\left|\boldsymbol{x}_j - \mu_i\right|\right|^2$$

- Lloyd's algorithm
- *k*-means++ initializes *k*-means by choosing a random initial mean based on the probability proportional to the cost function:

$$\mathrm{cost}(T) = \sum_{\mu \in T} \sum_{\boldsymbol{x} \in C_\mu} \|\boldsymbol{x} - \mu\|^2$$

- Scikit-learn Python library

How to Cluster for *k*-Means (i.e., How to Group Satellites)

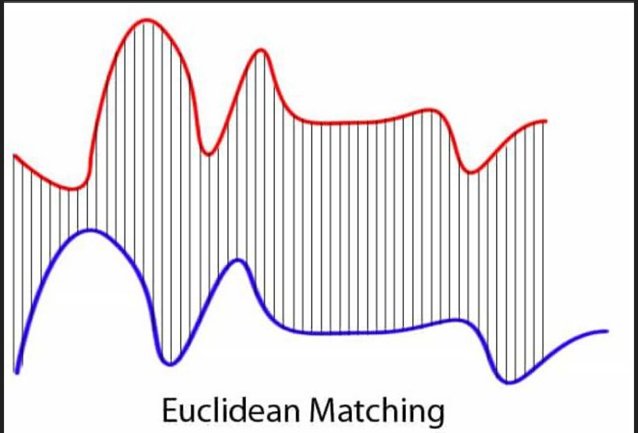
- Options to run *k*-means and compare results on:
  - Vector components (x, y, z) of ***r*** and ***h***
  - Magnitude |***r***| and |***h***|
  - Flattened data matrix of ***r*** and ***h***, each of size 1984 x 24054
- Normalized the data

$$\text{normed data} = \frac{\text{data} - \min\limits_{\text{satellites}}(\text{data})}{\max(\text{data}) - \min\limits_{\text{satellites}}(\text{data})}$$

- Less data if recently launched or crashed
  - Append zeros to ephemerides if crashed
  - Prepend zeros to ephemerides if launched

Problems with *k*-Means Clustering

- Euclidean metric not suitable for time-series data
  - Cannot identify time shifts
- Cannot cluster tensors
  - Each time point can contain multiple quantities
- Requires all time-series to have the same number of time points



Time-Series *k*-Means Clustering

- Dynamic time warping (DTW) metric
- DTW barycenter averaging (DBA) to calculate the cluster centers
- tslearn Python library
- Advantages:
  - Invariant to time-shift
  - Can cluster tensors
  - Each time-series is allowed to have arbitrary number of time points
- Disadvantages:
  - Runs much slower than *k*-means (1 minute vs 30 minutes)

Time-Series *k*-Means Algorithm

The algorithm is similar to *k*-means.

1. Initialize the cluster centers.
2. Calculate the DTW distance between each time-series and the cluster centers.
3. Assign each time-series to a cluster.
4. DBA to find the cluster centers.
5. Repeat until convergence.

Dynamic Time Warping

- Creates all possible contiguous mappings from the indices of one time-series to another
- Calculates and sum the distance between each matched time point in a mapping
- Takes the minimum distance out of the possible mappings

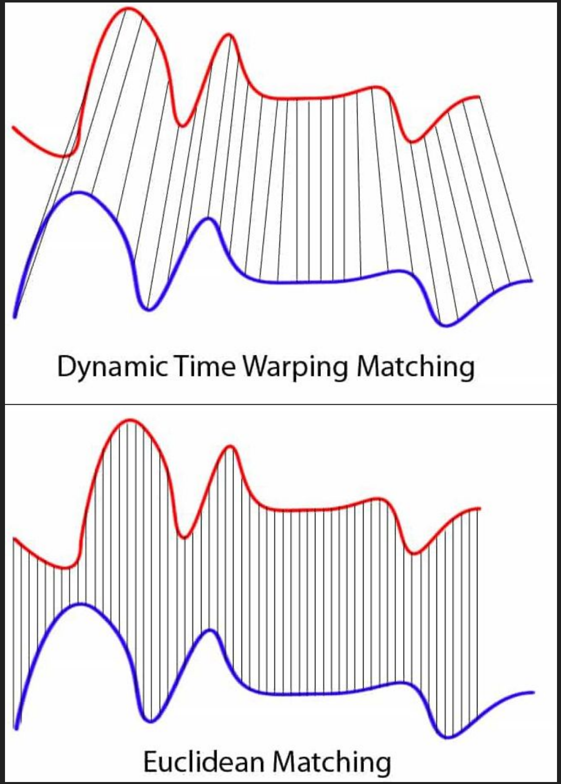
Let $\boldsymbol{x} = (x_0, \cdots, x_n)$ and $\boldsymbol{y} = (y_0, \cdots, y_m)$ be two time series, where $x_i, y_i \in \mathbb{R}^d$

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j) \in \pi} d(x_i, y_j)^2}$$

where $\pi = [\pi_0, \cdots, \pi_K]$ is a path that satisfies the following properties:

- it is a list of index pairs $\pi_k = (i_k, j_k)$ with $0 \le i_k < n$ and $0 \le j_k < m$
- $\pi_0 = (0, 0)$ and $\pi_K = (n-1, m-1)$
- for all $k > 0$, $\pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
    - $i_{k-1} \le i_k \le i_{k-1} + 1$
    - $j_{k-1} \le j_k \le j_{k-1} + 1$

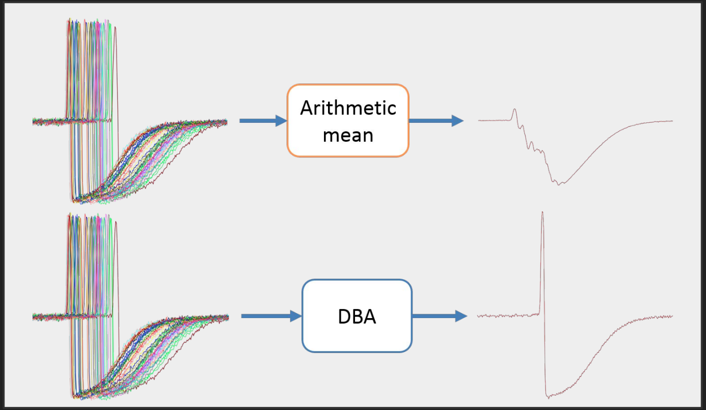
Euclidean vs DTW



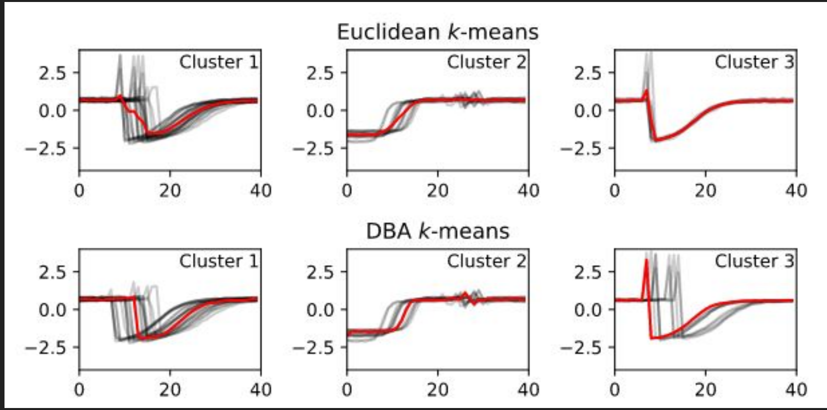
Source: https://databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html

DTW Barycenter Averaging

- Iteratively refines an average time-series using an expectation-maximization scheme:
  a. Find the best alignment of the set of time-series data to the fixed average using DTW.
  b. Update the average time-series using this alignment.



Source: https://github.com/fpetitjean/DBA



k-Means

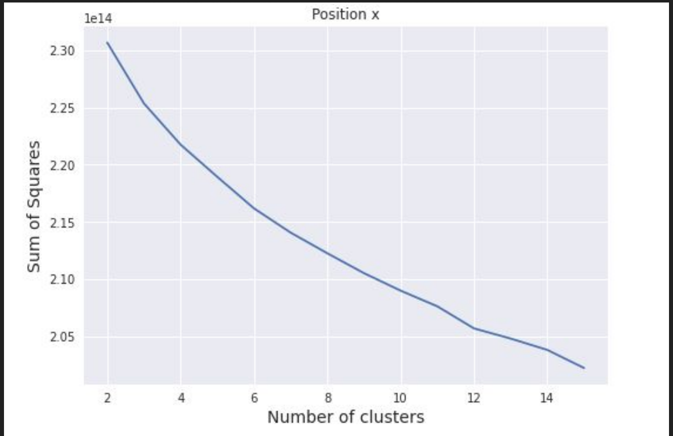
- Selecting the optimal cluster number
  - Elbow method for sum of squares
  - Silhouette score
- Performance metrics:
  - Adjusted Mutual Information score (compared to human-labeled data)
    - *AMI* = 1: same clusters
    - *AMI* = 0: different clusters
  - Silhouette score
    - *s* = 1: means good clusters
    - *s* = -1: means bad clusters

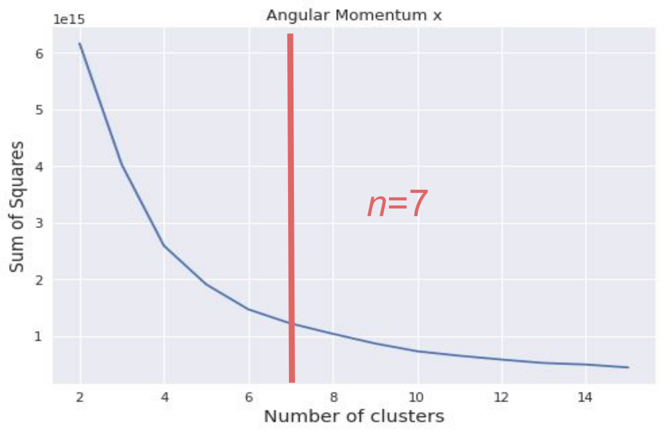
$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{\operatorname{avg}\{H(U),H(V)\} - E(MI(U,V))}$$

$$s = \frac{b-a}{\max(a,b)}$$

k-Means Results

- Elbow plot optimal cluster:
  - Position vectors never converged
    ⇒ Ruled out for further analysis
  - Radius : 5
  - Angular momentum vectors: 7
  - Angular momentum magnitude: 5
  - Angular momentum flattened: 6



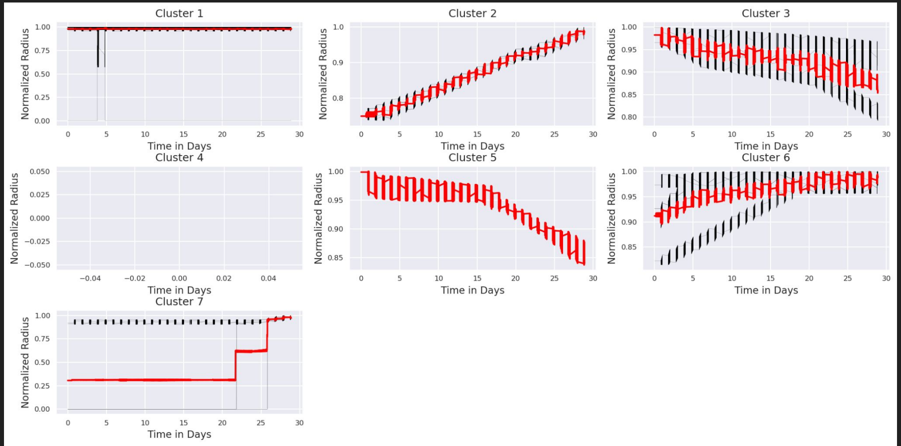


k-Means Results

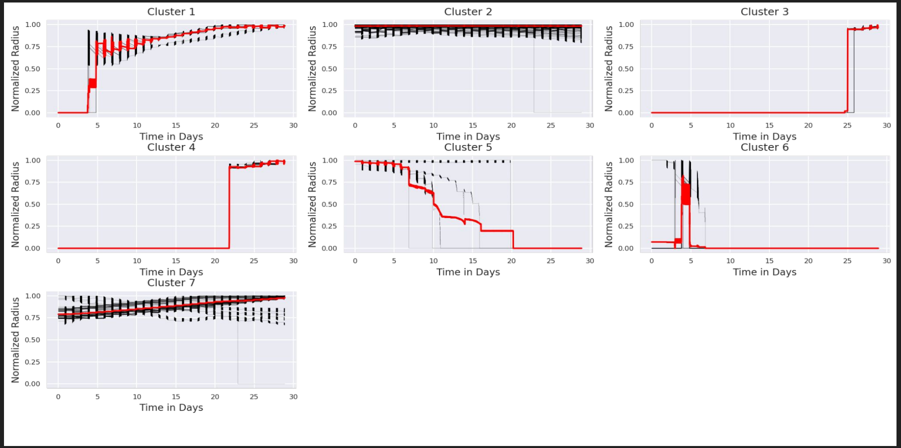
n clusters = 7

Parameter	Silhouette Coefficient	AMI (Human Truth)
h_x	0.636	-0.042
h_y	0.626	-0.010
h_z	0.981	0.242
$\|h\|$	0.842	0.495
h flat	0.644	-0.027
$\|r\|$	0.831	0.460

Human-Labeled Clusters



Radius Clustering

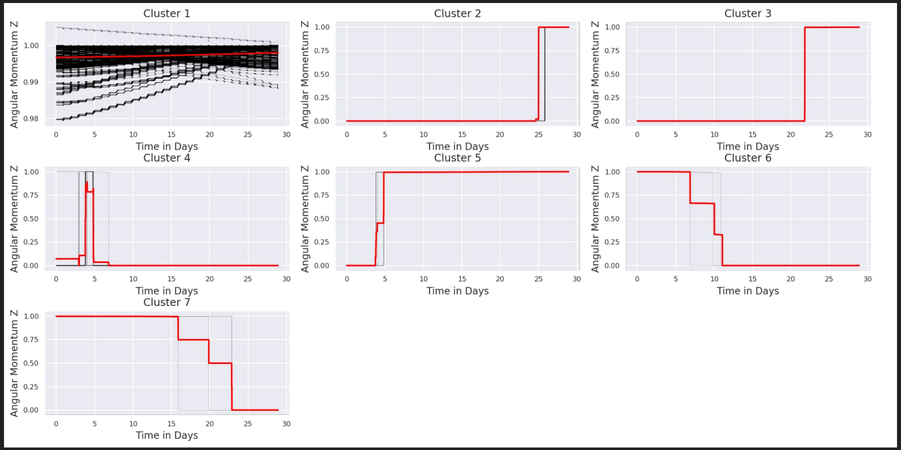


k-Means Results

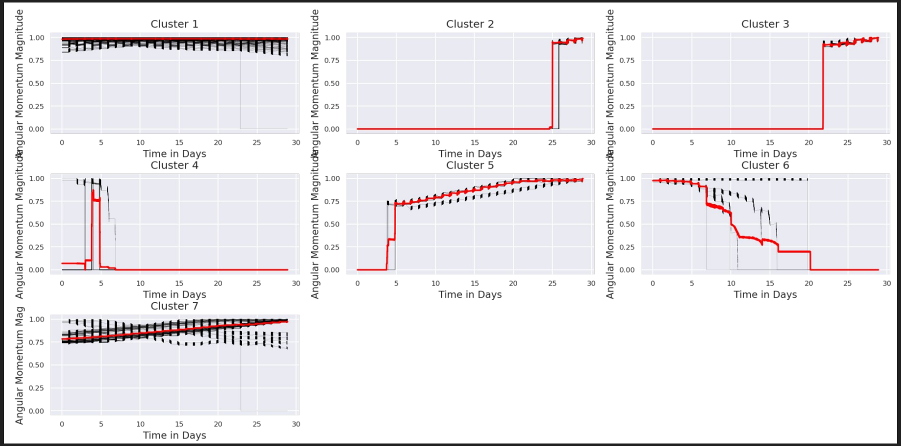
n clusters = 7

Parameter	Silhouette Coefficient	AMI (Human Truth)
h_x	0.636	-0.042
h_y	0.626	-0.010
h_z	0.981	0.242
$\|h\|$	0.842	0.495
h flat	0.644	-0.027
$\|r\|$	0.831	0.460

Angular Momentum *z*



Angular Momentum Magnitude

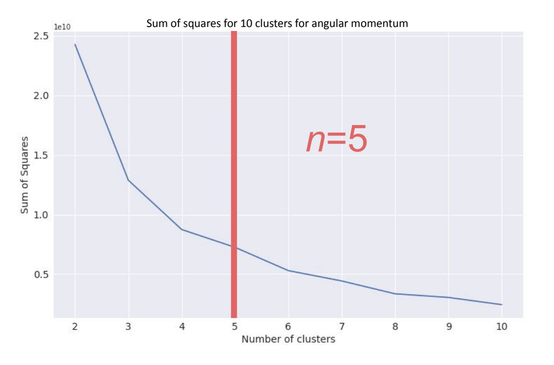


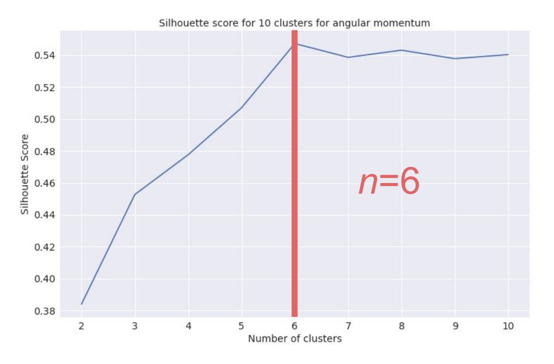
Time-Series *k*-Means Method

- Increase the time interval between data points to 2 hours: 1984 x 335 x 10
- Cluster on ***h***, |***r***|, and |***h***|
- Center each time-series to zero
- Select the optimal cluster number
  - Elbow method for sum of squares
  - Silhouette score
- Performance metrics:
  - Adjusted Mutual Information score
  - Silhouette score

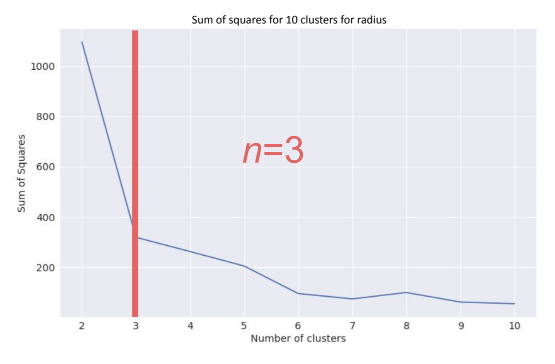
Select the Optimal Cluster Number

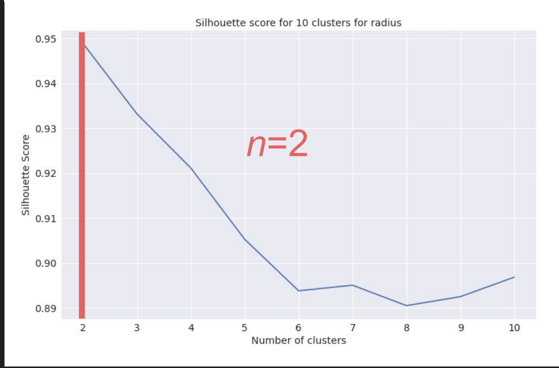
Angular momentum



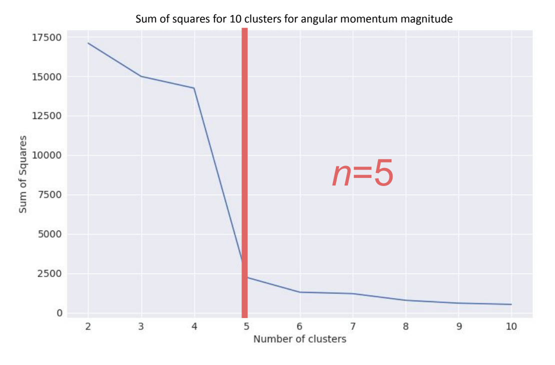


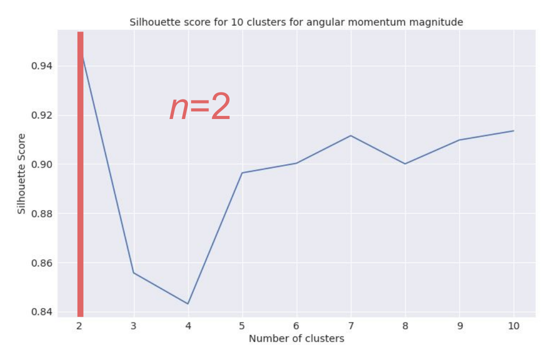
Radius





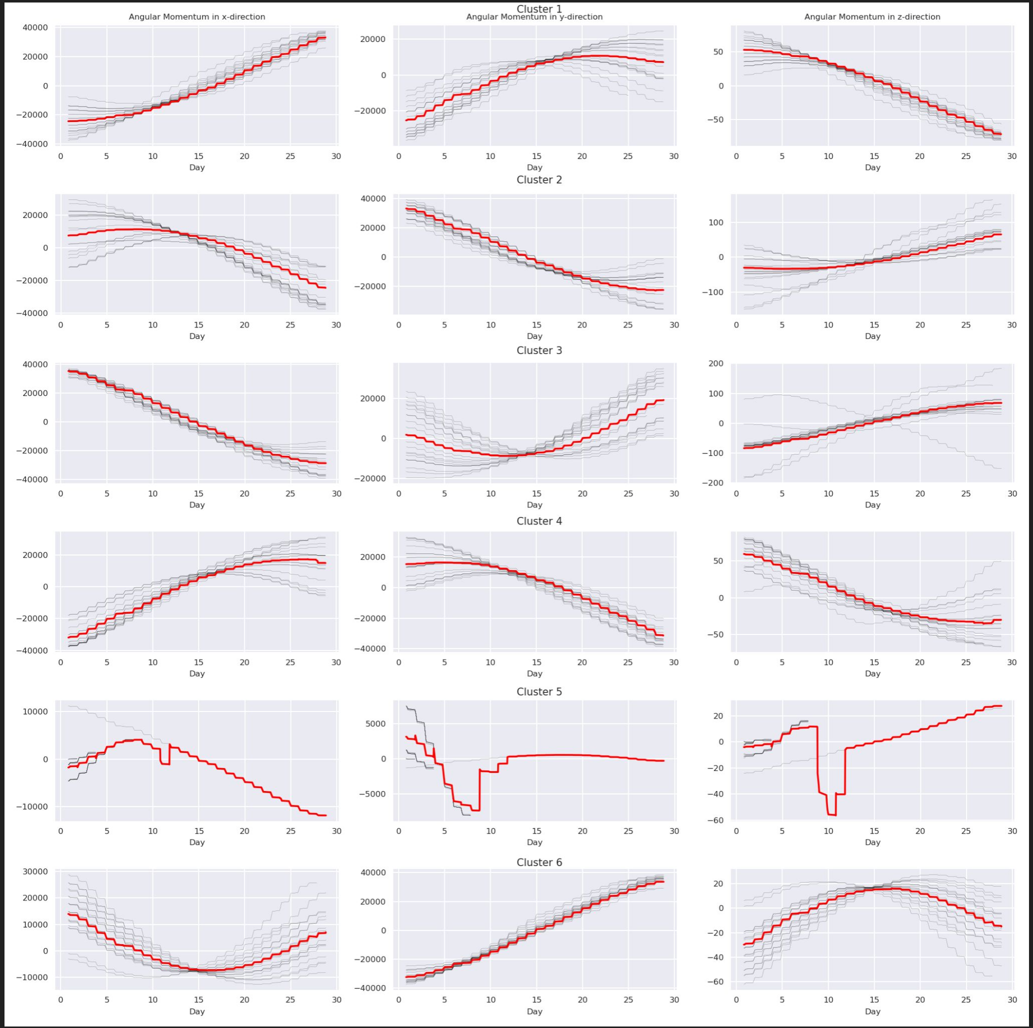
Angular momentum magnitude





Angular Momentum Result

- Optimal cluster number: 6
- Difficult to interpret angular momentum

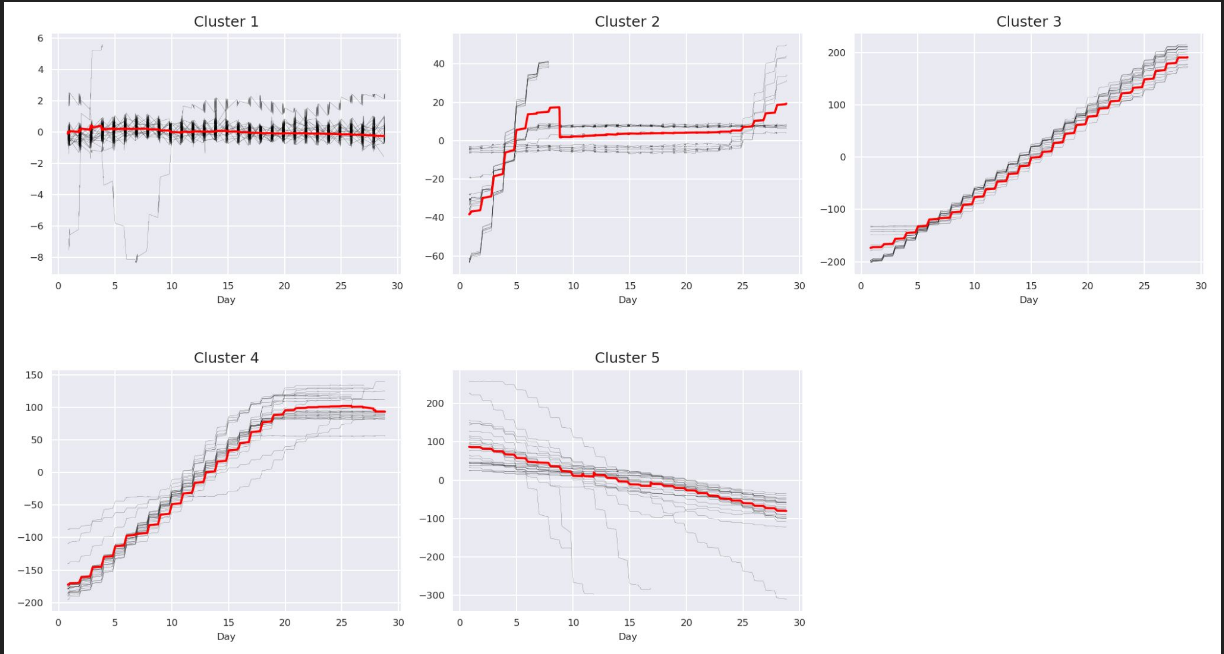


Radius Result

- Optimal cluster number: 3
- Doesn’t capture all of the satellite trajectories

Angular Momentum Magnitude Result

- Optimal cluster number: 5
- Good compromise between angular momentum and radius



Performance Metrics

For n = 7 clusters

k-Means

Parameter	Silhouette Coefficient	AMI (Human Truth)
h_z	0.981	0.242
$\|\boldsymbol{h}\|$	0.842	0.495
$\|\boldsymbol{r}\|$	0.831	0.460

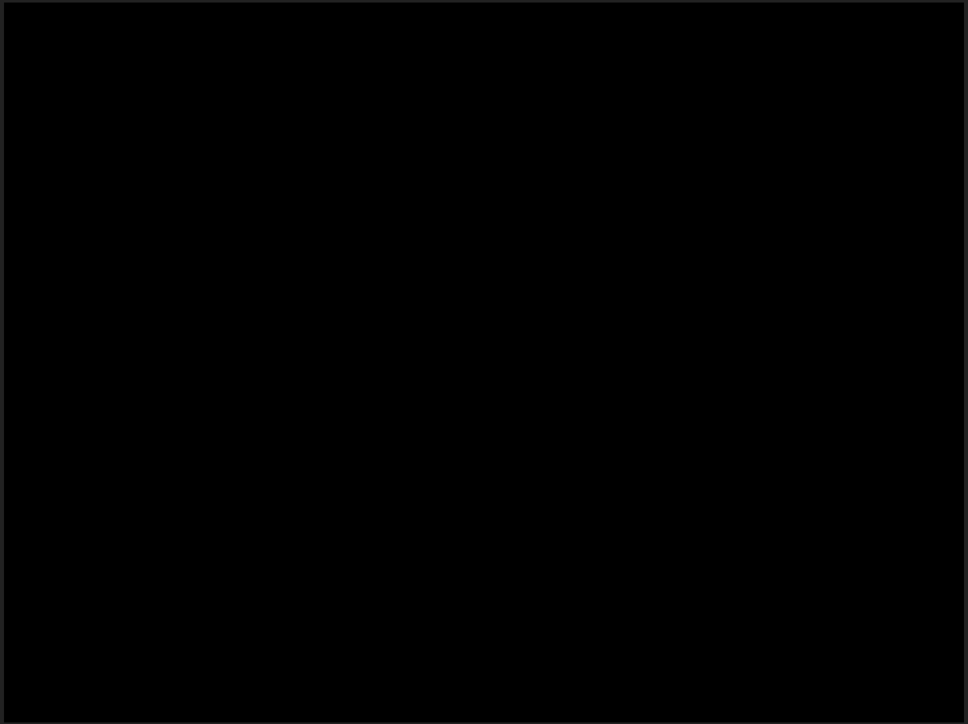
Time-Series *k*-Means

Parameter	Silhouette Coefficient	AMI (Human Truth)
$\boldsymbol{h}$	0.539	0.059
$\|\boldsymbol{h}\|$	0.895	0.703
$\|\boldsymbol{r}\|$	0.912	0.723

Conclusion & Future Work

- Lessons learned
  - Determine optimal clusters numbers
  - Density-based clustering
  - Padding data
  - Normalization
  - Trade-offs

- Future work
  - Other types of clustering techniques
  - Develop metrics to simplify data
  - Non-ML Approach:
    - Work with space community
    - Develop registry of maneuvers
    - Lobby congress



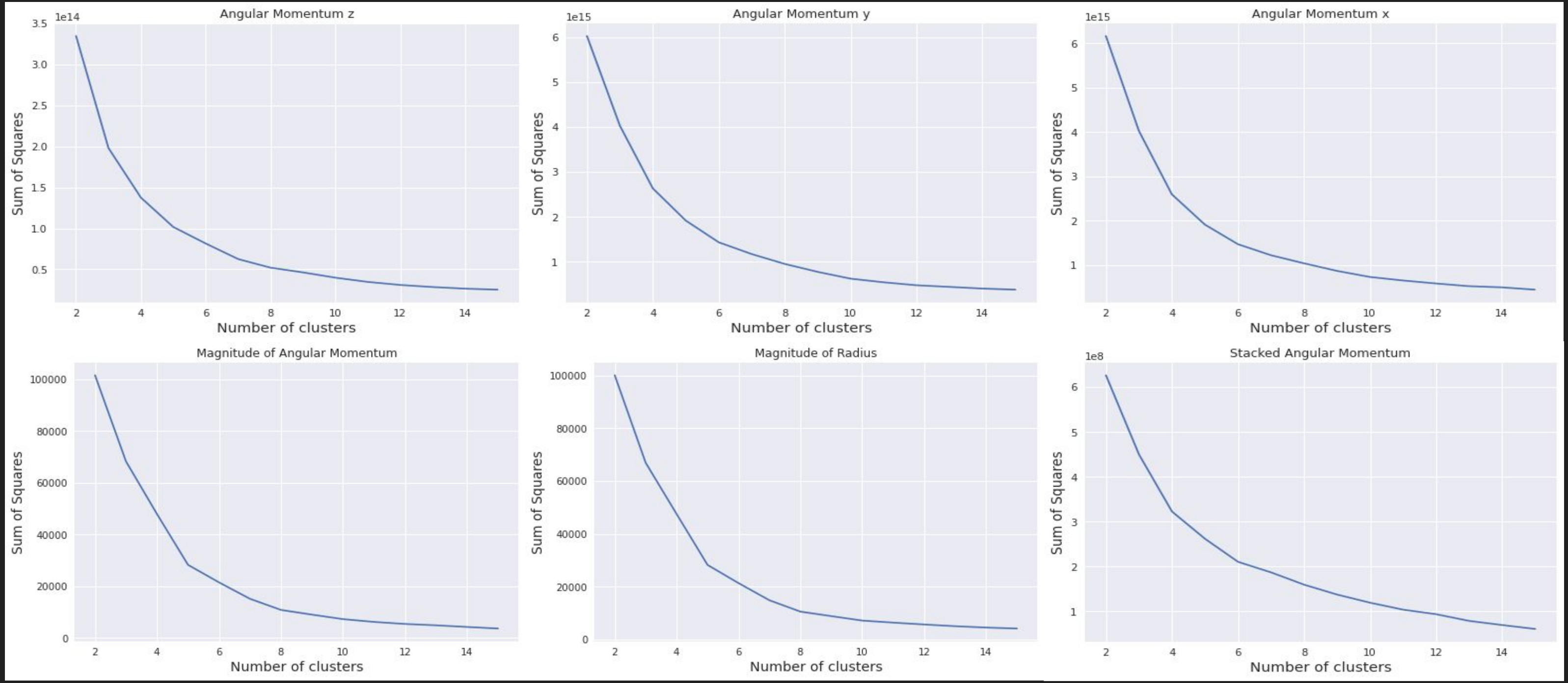
Reference

1. F. Petitjean, A. Ketterlin, P. Gançarski, “A global averaging method for dynamic time warping, with applications to clustering,” *Pattern Recognition*, vol. 44, p. 678-693, September, 2010. [Online]. Available: https://lig-membres.imag.fr/bisson/cours/M2INFO-AIW-ML/papers/PetitJean11.pdf. [Accessed: Apr. 23, 2022].
2. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, October, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf. [Accessed: Apr. 2, 2022].
3. R. Tavenard et al., “Tslearn, A Machine Learning Toolkit for Time Series Data,” *Journal of Machine Learning Research*, vol. 21, no. 118, p. 1-6, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-091.html. [Accessed Apr. 2, 2022].
4. R. Tavenard., “An introduction to Dynamic Time Warping,” n.d. [Online]. Available: https://rtavenar.github.io/blog/dtw.html [Accessed: Apr. 22, 2022].

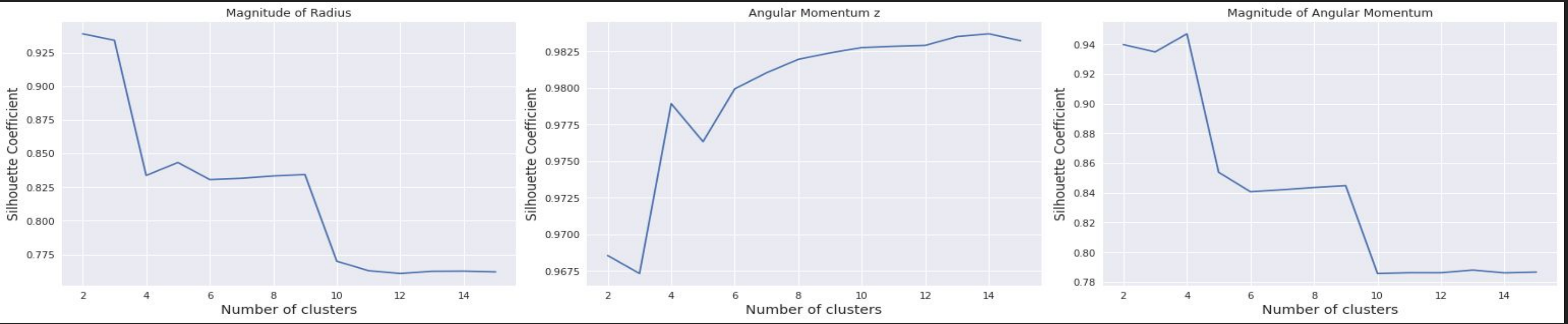
Questions?

Backup Slides

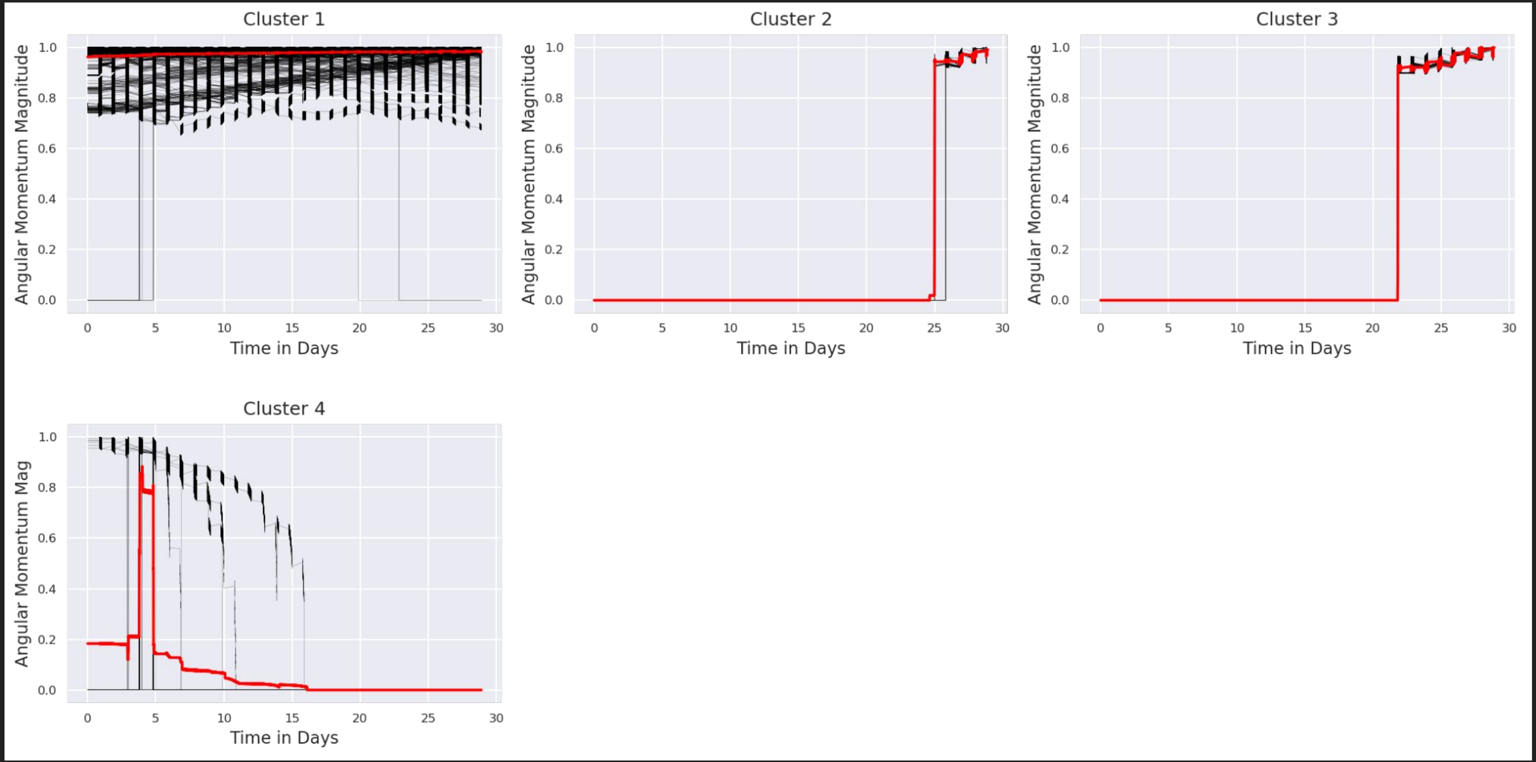
k-Means Elbows



k-Means Silhouettes



Angular Momentum Magnitude *k*-means Optimal Clusters



Radius *k*-Means Optimal Clusters