HiSAXy: A fast methodology for solar wind structure identification in millions of time series

Hala Lamdouar
University of Oxford

Sairam Sundaresan
Intel Labs

Anna Jungbluth
University of Oxford

Sudeshna Boro Saikia
University of Vienna

Amanda Joy Camarata
Colorado School of Mines

Nathan Miles
University of Colorado, Boulder

Marcella Scoczynski
Federal University of Technology – Paraná

Mavis Stone
Stanford University

Anthony Sarah
Intel Labs

Andrés Muñoz-Jaramillo
Southwest Research Institute

Ayris Narock
NASA Goddard Space Flight Center
ADNET Systems Inc

Adam Szabo
NASA Goddard Space Flight Center

Abstract

We present a hybridized unsupervised clustering algorithm ***HiSAXy*** as a novel way to identify frequently occurring magnetic structures embedded in the interplanetary magnetic field (IMF) carried by the solar wind. The ***HiSAXy*** algorithm utilizes a combination of indexable Symbolic Aggregate approXimation (iSAX) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to efficiently identify clusters of patterns embedded in time series data. We utilized ***HiSAXy*** to identify small-scale structures, known as discontinuities, embedded in time series measurements of the IMF. In doing so, we demonstrate the capability of the algorithm to significantly reduce the amount of human analysis hours required to identify these structures, all the while maintaining a high degree of self similarity within a given cluster of time series data.

1 Introduction

The solar wind is a constant stream of plasma the originates from the outermost regions of the Sun's atmosphere known as the solar corona. As the solar wind plasma expands in the solar system, it carries with it a remnant of the solar magnetic field that is formally referred to as the interplanetary magnetic field (IMF). The solar wind plasma itself is highly structured and as a result, so is the IMF. The structures embedded in the solar wind span large scales in both space and time; the largest structures last on the order of days to weeks, while the smallest structures last on the order of seconds to minutes. Previous applications of artificial intelligence to the solar wind have focused on classifying the solar wind into distinct sub-types (fast or slow) based on the plasma properties, e.g., solar wind speed and number density [1, 2]. In this work, we use unsupervised learning techniques to identify and classify repeating magnetic structures in the solar wind. This task is typically done by hand in heliophysics, significantly limiting the scope and breadth of the resulting classification. Our goal is to find an optimal compromise between the specificity of our classification and how long it takes a human scientist to analyze and interpret the output of our algorithm.

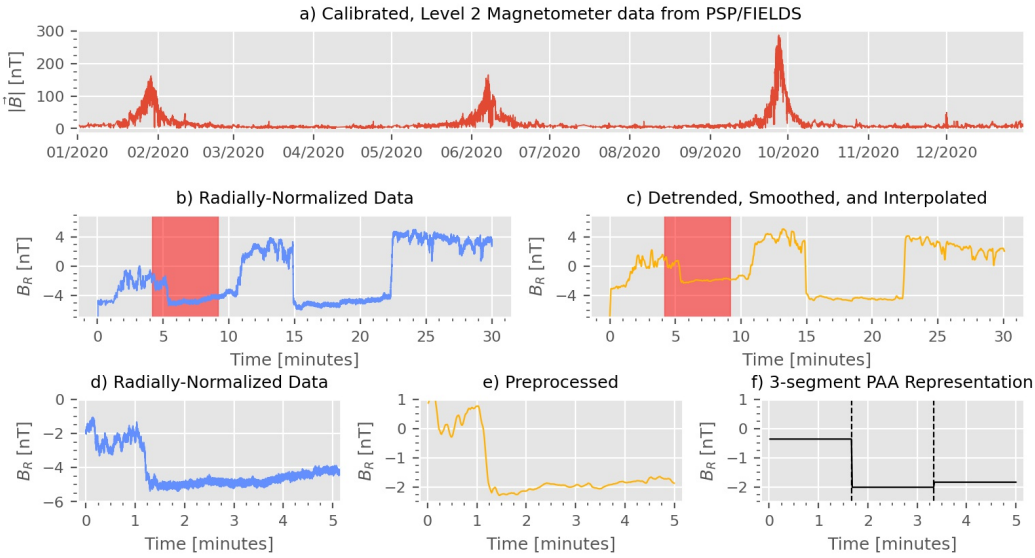

Figure 1: **a)** The original calibrated, level 2 magnetometer data from the PSP/FIELDS instrument for the year 2020. **b)** 30 minute interval of the magnetic field data after applying radial scaling. **c)** The same 30 minute interval after detrending using a rolling mean over an 1.5 hr interval, smoothing using a rolling mean over a 1 second interval, and linearly interpolating to a uniform cadence of 1 Hz. **d)** Zoom in to the 5 minute interval highlighted by the red shading in the top left. **e)** Zoom in to the 5 minute interval highlighted by the red shading in the top right. **f)** Piece-wise aggregate approximation of the 5 minute interval using three, 100 second intervals.

Unsupervised supervised machine learning techniques can help classify inputs in the absence of known labels. This is accomplished by searching for clusters in a feature space comprised of the inputs. In this project we use build on the Hierarchical density-based spatial clustering of applications with noise (HDBSCAN)[3], which is an extension of DBSCAN[4]. HDBSCAN works by finding kernels of density in phase space, which are gradually grown to encompass nearby points enabling the merging of adjacent density based clusters. This is accomplished by creating a hierarchical tree where small individual peaks of density gradually merge into larger clusters until all points have been aggregated together.

The novelty of our approach is to sort all time series segments using the indexable Symbolic Aggregate approXimation (iSAX) algorithm[5], prior to the application of HDBSCAN. iSAX is a novel multi-resolution symbolic representation of time series that enables the scaling and indexing of massive datasets, with millions of time series, requiring a single operation per time-series, and enabling ultra-fast retrieval and pattern matching [6]. It is constructed by encoding time series segments into words of specified length and of a specified cardinally (SAX)[7], and sorting these words into a binary search tree. This tree grows organically, increasing the cardinality of the SAX representation. We demonstrate that, while an iSAX tree is a superb framework for pattern matching and retrieval, it is suboptimal for unsupervised pattern identification. Our hybridized approach, ***HiSAXy***, combines the strengths of iSAX (speed and scalability) and HDBSCAN (robust clustering) to produce a highly efficient unsupervised clustering algorithm that saves an incredible amount of human effort in pattern discovery and interpretation.

2 Datasets

We analyzed magnetometer data obtained by the FIELDS [8] instrument on board the Parker Solar Probe (PSP) from October 2, 2018 to December 31, 2020. This data can be downloaded through the Space Physics Data Facility's Coordinated Data Analysis Website [1] (CDAWeb). In total, we analyzed 150,488 5-minute segments of time series data containing $\sim$2.2 billion in-situ measurements of the three vector components of the IMF embedded in the solar wind. For each 5-minute segment, we

[1] https://cdaweb.gsfc.nasa.gov/index.html/

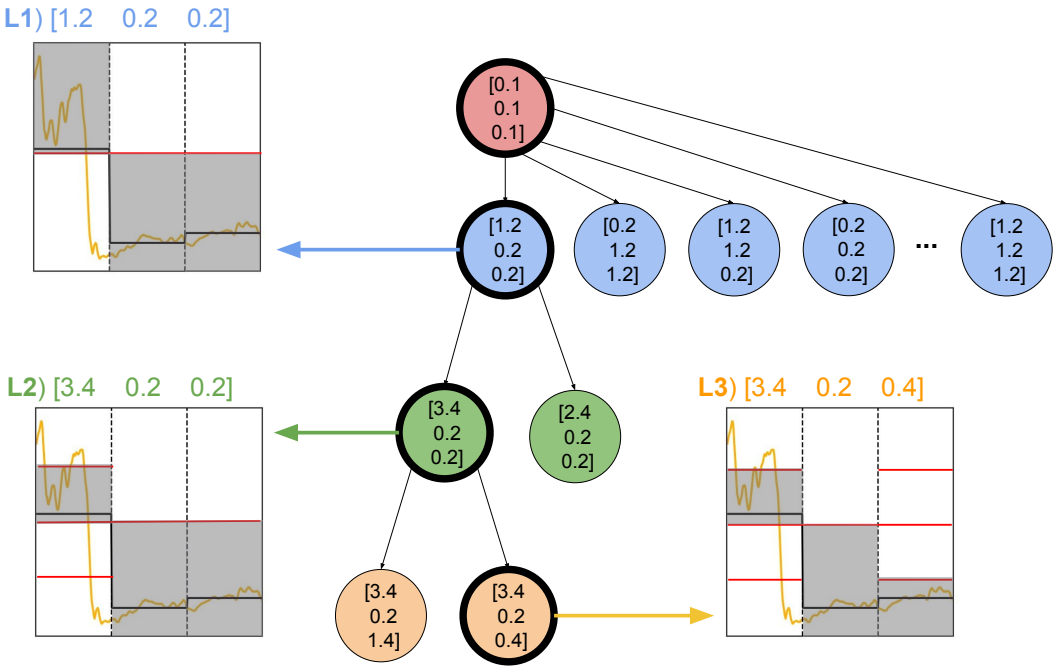

Figure 2: iSAX tree for indexing sequences using 3-letter words. Tree depth is represented using different colors. iSAX increases the depth of the tree by escalating the cardinality of one of the letters in each node-split. A black highlight denotes the tree node containing the sample time series at each tree level. An iSAX letter is composed of two numbers separated by a dot: range (***r***) and cardinality (***c***) – ***r.c***. The sample time series is represented by the iSAX word [1.2 0.2 0.2] (**L1**), [3.4 0.2 0.2] (**L2**), and [3.4 0.2 0.4] (**L3**). Shaded areas indicate the range of each letter for the sample time series in a given cardinality and can be use to approximate the euclidean distance between any two time series in the tree, enabling the clustering of tree nodes using HDBSCAN. The sample sequence is the same shown in Figs. 1-e & f. Tick marks and labels have been removed for simplicity.

apply the following preprocessing steps. First, we apply the appropriate radial-scaling laws to the components of the magnetic field to remove the effects of PSP's orbit on the baseline magnetic field strength evident in Fig. 1a). Next, we detrend the data with an 1.5hr window to remove any local trends on times scales that are 18 times longer than discontinuities. After that, we smooth with a rolling mean over a one-second window and linearly interpolate to a one-second cadence to create the cleaned times series. The end result of preprocessing can be seen in Figure 1e.

3 Methodology

HiSAXy first indexes all the time series segments using iSAX. iSAX creates a tree structure, illustrated in Figure 2, with a root node (red), which contains pointers to children nodes that can be internal nodes (pointing to children of their own) or terminal nodes (holding onto indexed segments). Each node is characterized by an iSAX word that encodes well defined ranges in magnetic fields and times (see shaded regions in Fig. 2). The first tree level (blue in Fig. 2) contains nodes with iSAX words at a minimum cardinality of two. This amounts to initially sorting time series depending on whether each letter in their PAA representation is below or above the mean value. Each terminal node can hold a maximum number of segments determined by the user. Whenever that number is exceeded, terminal nodes spawn two children that each contain a portion of the segments and become internal nodes themselves. The new nodes inherit the iSAX word of their parent, with the exception of a single letter which is escalated in cardinality, enabling a more nuanced description of the segments it holds. The example of Fig. 2 shows how this escalation leads from the starting iSAX word with the minimum cardinality of 2 (L1: [1.2 0.2 0.2]) into an iSAX word with one letter of cardinality 4 at level 2 (L2 [3.4 0.2 0.2]), and two letters of cardinality 4 at level 3 (L3 [3.4 0.2 0.4]).

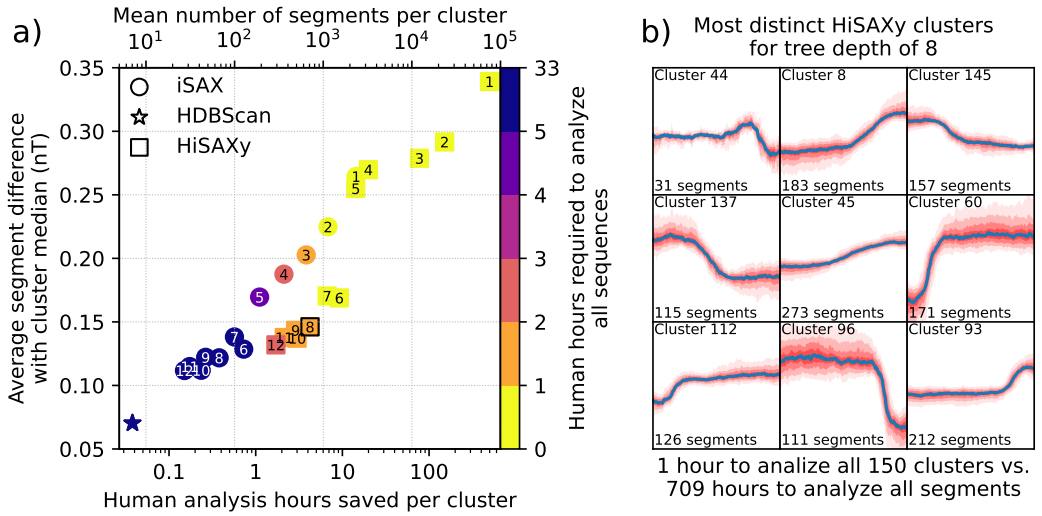

Figure 3: **a)** iSAX (circles) can be used as an approximate clustering algorithm by extracting all the nodes at a specified tree depth (numbers inside markers) and treat them as clusters. A deeper (larger) tree level has more specific clusters with smaller average differences between members (bottom of y-axis). However, each cluster also has fewer segments so human analysis still requires significant effort (left of x-axis). Coupling iSAX with HDBSCAN (HiSAXy; squares) combines similar nodes into clusters, reducing the number of clusters that need to be analyzed, without significantly impacting cluster specificity. Resulting in significant saving of human effort. HDBSCAN alone (star) produces even more clusters than iSAX alone, which is undesirable for our application. **b)** We find that combining iSAX and HDBSCAN for nodes of tree level 8 is a very good compromise between specificity and human effort. This approaches saves more than 700 hours of human effort compared to working with the original segments.

At any specified tree level, all time series segments indexed underneath a node are more similar to each other than those under any other node at the same level. However, tree escalation is dependent on the order in which the segments are provided and is strongly dependent on the user specified threshold. This is problematic when using it as a clustering mechanism. We address this by clustering all nodes at a specified level using HDBSCAN. We do this by using the euclidean distance between nodes given the expected value of each letter in their iSAX word.

HiSAXy is built on the iSAX implementation of Lucas Foulon [9], with the addition of a framework that enables the retrieval of the origin of each time series segment in the database. For HDBSCAN, we use the open source python package HDBSCAN [3]. Our codebase can be found in a github repository to be released after the anonymized review. Our experiments were run on the Google Cloud Platform using the Compute Engine with a c2-standard-8 VM (8vCPUs, 32GB RAM) and 1 TB of SSD disk space.

4 Experiments and Results

In this work, we compare the results of clustering the time series data using iSAX, HDBSCAN, and ***HiSAXy*** using different tree level depths. Focus on identifying an optimal combination of specificity and human effort that enables scientific discovery at an unprecedented scale in heliophysics. To asses cluster quality, we use a measure of self similarity defined as follows,

$$\alpha = \frac{1}{T}\sum_{t=0}^{T} s(t) = \frac{1}{T}\sum_{t=0}^{T}\left(\frac{1}{N_s}\sum_{i=1}^{N_s}|B_{\text{med}}(t) - B_i(t)|\right), \tag{1}$$

where N_s is the number of segments in a cluster, $B_i(t)$ is the i^{th} segment of magnetic ftield time-series data, and $B_{\text{med}}(t)$ is the median of all the segments at each time step, t. To assess human effort, we visually inspected ***HiSAXy*** clusters to label clusters containing magnetic discontinuities. We registered an average rate of ~ 150 clusters per hour per human. Using this classification rate,

we asses the time saved in analyzing clusters of similar curves generated by the three algorithms. In Figure 3, we show the results of the comparison which demonstrates the unique ability of ***HiSAXy*** to generate an optimal number of segments per cluster (thereby increasing the time saved) without sacrificing specificity.

5 Conclusions, Limitations, and Future Work

In this paper, we introduced a hybridized unsupervised learning approach, ***HiSAXy*** , for classifying structures embedded in the solar wind. When compared to iSAX and HDBSCAN individually, we found that ***HiSAXy*** outperforms both by identifying larger clusters without sacrificing the intracluster self similarity. At the present moment, the fundamental limitation of ***HiSAXy*** is the implicit assumption that the time series data follow a Gaussian distribution. However, this is not a hard constraint and can be modified as needed. In the future, we plan on extending our work by applying ***HiSAXy*** to study a variety solar wind properties across a range of timescales.

Broader Impact

The ability to automate the identification of aperiodic structures embedded in the solar wind has been a long-standing goal for the heliophysics community. Current methodologies rely on heuristic techniques that can result in discrepancies in the identification of even the largest structures like ICMEs [10]. As the archive of solar wind data continues to grow with data from new missions like PSP and Solar Orbiter, so does the burden on helio- and space physicists. By utilizing the ***HiSAXy*** framework, researchers can focus their efforts on analyzing the most frequently occurring magnetic structures at arbitrary timescales (5-minute intervals, 30-minute intervals, 12-hour intervals, etc...). Furthermore, our work is applicable to not only time series measurements of other solar wind properties like number density, velocity, and temperature, but also any form of time series data.

Acknowledgments and Disclosure of Funding

This work was conducted at the Frontier Development Laboratory (FDL) USA 2021. The FDL USA is a public / private research partnership between NASA, the SETI Institute and private sector partners including Google Cloud, Intel, IBM, Lockheed Martin, and NVIDIA. These partners provide the data, expertise, training, and compute resources necessary for rapid experimentation and iteration in data-intensive areas.

References

[1] Verena Heidrich-Meisner and Robert F. Wimmer-Schweingruber. Chapter 16 - solar wind classification via k-means clustering algorithm. In Enrico Camporeale, Simon Wing, and Jay R. Johnson, editors, *Machine Learning Techniques for Space Weather*, pages 397–424. Elsevier, 2018.

[2] Jorge Amaya, Romain Dupuis, Maria Elena Innocenti, and Giovanni Lapenta. Visualizing and interpreting unsupervised solar wind classifications. *Frontiers in Astronomy and Space Sciences*, 7:66, 2020.

[3] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[5] Jin Shieh and Eamonn Keogh. isax: Indexing and mining terabyte sized time series. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 623–631, New York, NY, USA, 2008. Association for Computing Machinery.

[6] Lucas Foulon, Serge Fenet, Christophe Rigotti, and Denis Jouvin. Scoring message stream anomalies in railway communication systems. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 769–776, 2019.

[7] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, page 2–11, New York, NY, USA, 2003. Association for Computing Machinery.

[8] S. D. Bale, K. Goetz, P. R. Harvey, P. Turin, J. W. Bonnell, T. Dudok de Wit, R. E. Ergun, R. J. MacDowall, M. Pulupa, M. Andre, M. Bolton, J. L. Bougeret, T. A. Bowen, D. Burgess, C. A. Cattell, B. D. G. Chandran, C. C. Chaston, C. H. K. Chen, M. K. Choi, J. E. Connerney, S. Cranmer, M. Diaz-Aguado, W. Donakowski, J. F. Drake, W. M. Farrell, P. Fergeau, J. Fermin, J. Fischer, N. Fox, D. Glaser, M. Goldstein, D. Gordon, E. Hanson, S. E. Harris, L. M. Hayes, J. J. Hinze, J. V. Hollweg, T. S. Horbury, R. A. Howard, V. Hoxie, G. Jannet, M. Karlsson, J. C. Kasper, P. J. Kellogg, M. Kien, J. A. Klimchuk, V. V. Krasnoselskikh, S. Krucker, J. J. Lynch, M. Maksimovic, D. M. Malaspina, S. Marker, P. Martin, J. Martinez-Oliveros, J. McCauley, D. J. McComas, T. McDonald, N. Meyer-Vernet, M. Moncuquet, S. J. Monson, F. S. Mozer, S. D. Murphy, J. Odom, R. Oliverson, J. Olson, E. N. Parker, D. Pankow, T. Phan, E. Quataert, T. Quinn, S. W. Ruplin, C. Salem, D. Seitz, D. A. Sheppard, A. Siy, K. Stevens, D. Summers, A. Szabo, M. Timofeeva, A. Vaivads, M. Velli, A. Yehle, D. Werthimer, and J. R. Wygant. The FIELDS Instrument Suite for Solar Probe Plus. Measuring the Coronal Plasma and Magnetic Field, Plasma Waves and Turbulence, and Radio Signatures of Solar Transients. *Space Sci. Rev.*, 204(1-4):49–82, December 2016.

[9] Lucas Foulon. *Détection d'anomalies dans les flux de données par structure d'indexation et approximation. Application à l'analyse en continu des flux de messages du système d'information de la SNCF. (Anomaly detection in data streams using indexing and approximation. Application to continuous analysis of message streams within the SNCF information system)*. PhD thesis, INSA de Lyon, France, 2020.

[10] CT Russell and AA Shinde. On defining interplanetary coronal mass ejections from fluid parameters. *Solar Physics*, 229(2):323–344, 2005.