Effect of Entropy-Based Discretization on Analysis of Geomagnetic Indicies and Solar Wind Data

Andrew Zheng[1,2] Shing F. Fung[1]
[2]River Hill High School, Clarksville MD
[1]ITM Physics Laboratory, Code 675, NASA Goddard Space Flight Center, Greenbelt MD

Abstract

Binning of data is a common technique for preparing data sets for implementing machine learning algorithms or performing statistical analysis of the data. The narrower the bin, the better the data within the bin is representing the characteristics of that bin, provided that there is sufficient data within the bin for making statistical inferences. If a parameter has a relatively uniform occurrence probability distribution over a range of the parameter, then subdividing the parameter range into bins of equal width might be the simplest binning technique to be used. Alternatively, data bins can also be defined on equal frequency-basis so as to maintain the sampling size across the bins. These are known as unsupervised-binning methods. It is well known, however, that the occurrence probability distributions of solar wind and geomagnetic activity parameters are highly skewed. Given that the magnetosphere is a highly nonlinear, multi-dimensional system, there should be strong interdependencies between the parameters characterizing the states of the magnetosphere. For example, the solar wind dynamic pressure and IMF may have some influence on the resultant global geomagnetic activity levels measured by the Kp index. In this situation, supervised-binning techniques should be used. In this presentation, we will investigate the use of the entropy-based binning technique to analyze solar wind, IMF, and geomagnetic activity data for the characterization of magnetospheric states, and compare the results against those obtained previously by using unsupervised binning techniques. We will also investigate how the distributions could lead to the discovery of anomalous interactions between IMF parameters.

Introduction

- Statistical analysis of a variable (an observable) is based on the occurrence probability distribution.
- Occurrence probability distribution is often represented by a histogram.
  - Plot of frequency of occurrence as a function of the variable bins.
- Too many narrow bins are sometimes not practical for analysis.

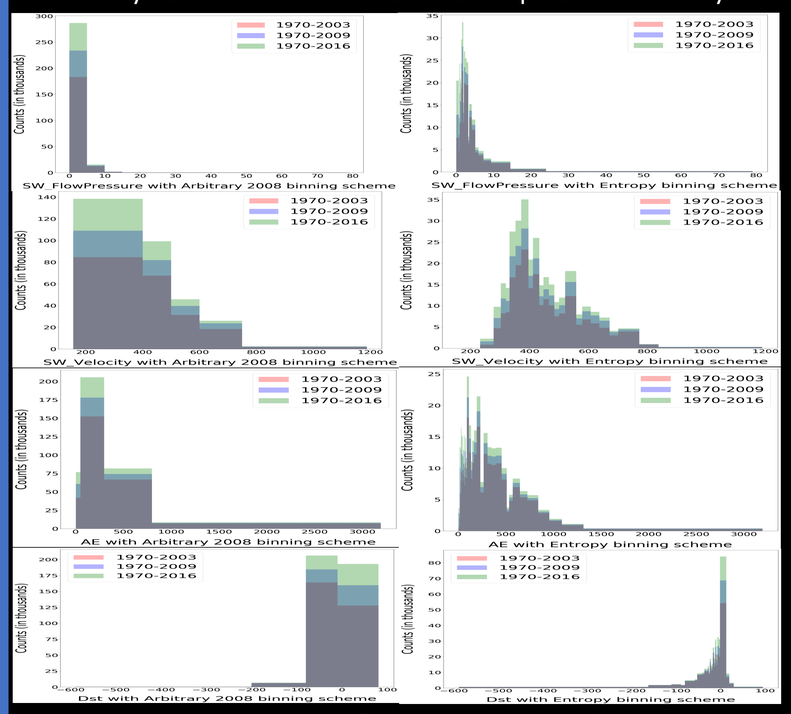


Figure 1. Psw, Vsw, AE, & Dst histograms using an arbitrary non-supervised binning scheme (left column) and supervised binning scheme (right column).

Questions to be addressed:

- How does binning scheme affect the results of statistical analysis?
- What binning scheme should be used for variables
  - With skewed-distributions?
  - Whose occurrences might depend on other variables or conditions?

Supervised vs. Unsupervised Binning Methods

- Supervised methods make use of class labels (i.e., certain correlated results) when partitioning a data set. It is based on how much effect a driver parameter may have on a set of correlated response outcomes [Dougherty et al., 1995].
- Unsupervised binning methods do not require the class information to discretize continuous attributes. It is most suited for binning independent, random variables.
- Entropy-based binning, a supervised discretization technique, is more applicable to binning parameters whose occurrences may be correlated with other parameters, and so have skewed distributions.

Entropy-Based Binning Method

- Entropy-based binning, a supervised technique, can be used to bin and analyze parameters with skewed distributions [Meurer, 2015].
- The original entropy of a skewed parameter distribution (D) is calculated according to

$$Origin\ al\ Entropy = -\sum_{k=1}^{n} p_i \log_2 p_i$$

where p_i is the driver variable probability of resulting in the i^{th} of m possible response parameter bins. (We've used Kp as responses for all driver parameters, but AE for Dst.)

- Assuming that there exists a bin boundary "a" that can split D suitably into two portions, we calculate the net entropy (information gain) of the two portions of D according to

$$Net\ Entropy\ (D)_{\leq a\ and > a} = p_{\leq a} \sum_{k=1}^{n} p_{\leq a_n} \log_2 p_{\leq a_n} + p_{>a} \sum_{k=1}^{n} p_{>a_n} \log_2 p_{>a_n}$$

- The entropy gain calculated as

$$Entropy\ Gain = Original\ Entropy - Net\ Entropy$$

Is a measure of how much information (i.e., order) is gained by having a split at "a" compared to the original distribution without the split.

- The trial bin boundary "a" at which the entropy gain is largest yields the correct bin boundary.
- The process is repeated for each sub-intervals to determine additional bin boundaries. And it is stopped when the entropy gains become "small."

Statistical Analysis of Magnetospheric States

- Fung and Shao [2008] showed that magnetospheric state can be prescribed by corresponding pairs of magnetospheric driver and response states.
- Correspondence between magnetospheric drivers and responses are established by noting the time delays, τ, between the various driver (i.e., Psw, Btot, Bz, Vsw) and response (i.e., Kp, AE, Dst) state parameters
- τ is given by the time lag at which the weighted global average standard deviation, <σ>, of the correlation between the response and driver parameters over its m data bins, i.e.,

$$\langle \sigma \rangle = \sqrt{\sum_{j=1}^{m} {\sigma_j}^2 n_j / \sum_{j=1}^{m} n_j}$$

is a minimum, where n_j is the sample number in the j^{th} data bin of the driver parameter.

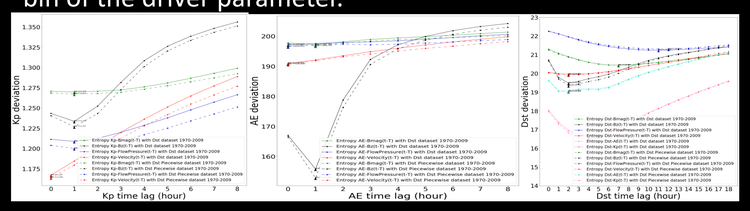


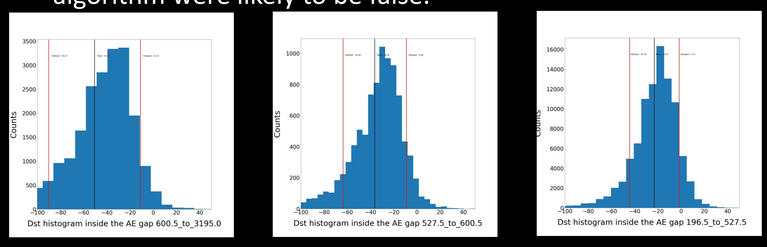
Figure 2. Time lag curves made with entropy-based binning scheme

Rs \ Ds	Psw	Bmag	Bz	Vsw	Kp	AE
Kp	0-2 (0-3)	0-2 (0)	1 (0-2)	0 (0)	-	-
AE	0-2 (0-8)	0-2 (0-2)	1 (1)	0 (0)	-	-
Dst	8-14 (2-14)	5-10 (4-8)	2-3 (2-3)	1-3 (0-4)	3-5 (2-5)	1-3 (1-2)

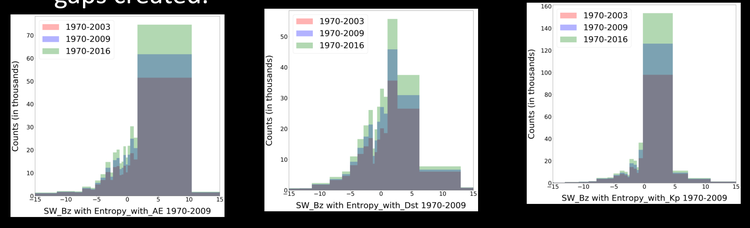
Figure 3. Time Lag table with unsupervised binning method results (parenthesis) compared to the entropy-based binning method results (not parenthesis)

Count Gaps in supervised histograms

- Looking at the AE histogram made with an entropy-based binning scheme, there are two distinct count gaps shown.
- In order to examine the count gaps in more detail, the AE dataset was split up into 5 regimes, where the two gaps would be their own regime and the three sections of data around the two gaps would also be their own regimes.
- If the gaps were true, it could possibly entail that the two represent a quick phase change to the three more "stable" regimes.
- Histograms were created for each regime to see if there are any differences between the parameters
- Specific count tables were also made to examine the five AE regimes more clearly.
- In the end, there was nothing worthy to been seen concerning five regimes and the two regimes created by the algorithm were likely to be false.



- Because Bz directly affects the three response parameters, supervised binning schemes can be made to validate the gaps created.



- Because there are gaps present in parameter Bz, there could be something overlooked within the data.

Acknowledgements

I would like to thank my mentor Dr. Fung for letting him work with him throughout the school year. I would also like to thank Mrs. Bender and Mrs. Whetzel for making this opportunity possible.

References

Dougherty et al., Supervised and unsupervised discretization of continuous features, in Machine Learning: Proc. Twelfth International Conference, 1995.

Fung, S. F. and X. Shao, Specification of multiple geomagnetic responses to variable solar wind and IMF input, Ann. Geophys., 26, 639–652, 2008.

Fung, S. F., J. A. Tepper, and X. Cai (2016), Magnetospheric state of sawtooth events, J. Geophys. Res. Space Physics, 121, doi:10.1002/2016JA022693.

Meurer, Kevin, A Simple Guide to Binning Data Using an Entropy Measure (Nov 19, 2015) (kevinmeurer.com/a-simple-guide-to-entropy-based-discretization/).