Effect of Various Supervised Discretization Techniques on Time-Lag Analysis

Andrew Zheng[1] and Shing Fung[2]

[1]River Hill High School, Clarksville, MD

[2]ITM Physics Laboratory, Code 675, NASA Goddard Space Flight Center, Greenbelt MD

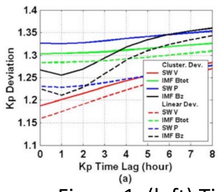
Introduction

- 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
- Time-lag is the time required for a response parameter to respond to a driver parameter
- τ is given by the time 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.,

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

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

- In order to determine the appropriate time-lag, however, there must be an appropriate binning scheme that is representative of each parameter.



(a)

Responses \ Drivers	V_{sw}	B_{tot}	IMF B_z	P_{sw}	K_p	AE
K_p	0	0	0 - 2	0 - 3	-	-
D_{st}	0 - 4	4 - 8	2 - 3	2 - 14	2 - 5	1 - 2
AE	0	0 - 2	1	0 - 8	-	-

Figure 1. (left) Time-lag curve of response parameter Kp and driver parameters (right) time-lag table and timeline of every pair of response and driver parameters

Binning considerations

- 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.

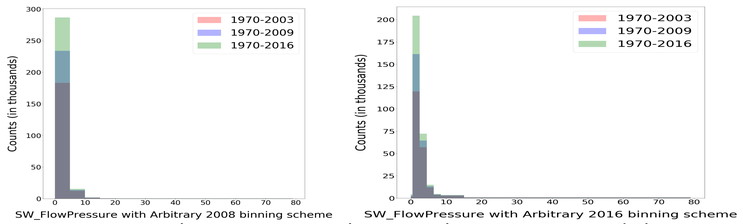


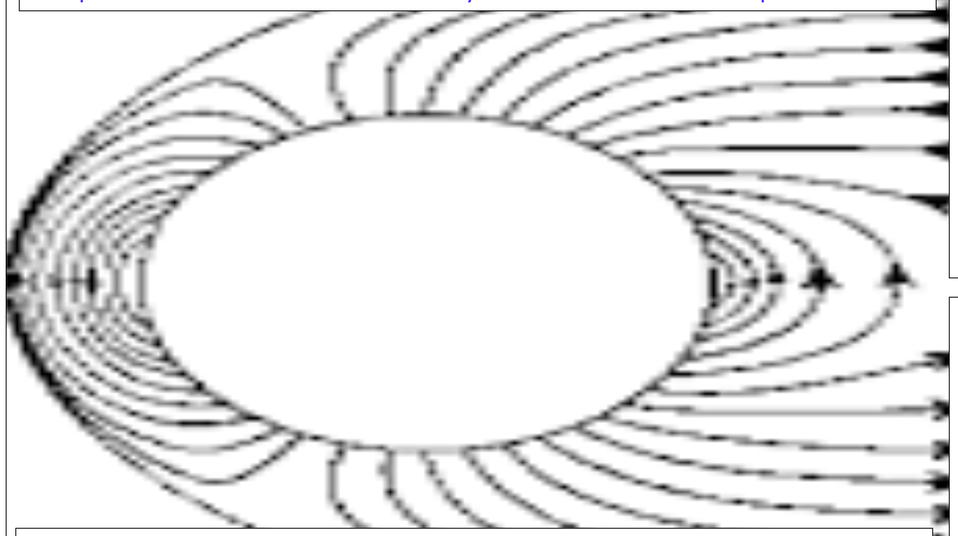
Figure 2. Psw histograms using binning schemes in Fung and Shao [2008] (left column) and Fung et al., [2016] (right column).

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.
- A supervised discretization technique is more applicable to binning parameters whose occurrences may be correlated with other parameters.



Previous Work

- Last year, the Entropy-Based Binning Technique was used as a comparison to an arbitrary binning scheme
- Entropy-based binning, a supervised technique, can be used to bin and analyze parameters with skewed distributions [Meurer, 2015].
- The idea of entropy-based binning is to reduce the entropy – thermodynamics (disorder within the data) – with every bin split you make.

$$Original\ Entropy(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

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

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

- The time-lag curves created from the Entropy-Based Binning Technique matched well with the arbitrary schemes
- However, when the scheme was created, there appeared to be discretization noise within the scheme
- In order to avoid discretization noise, another supervised binning technique was needed to reduce the number of bins

ChiMerge Algorithm

- A supervised binning scheme which uses a chi-squared statistic to determine whether bin boundaries should be kept or removed
- If the bin boundary is determined to be the most insignificant out of all the bin boundaries, then the bin boundary is eliminated.
- We will use the ChiMerge Algorithm to reduce our Entropy-based binning technique-created binning schemes

Results

- The result time-lags for each driver-response parameter pair for the binning schemes produced with an entropy-based binning scheme and a ChiMerge-refined entropy-based binning scheme

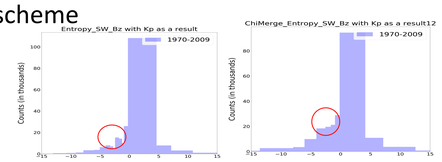
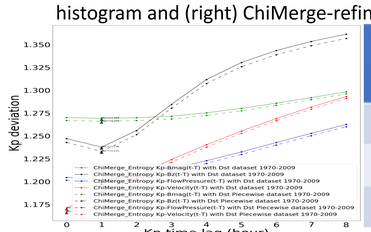


Figure 4. Histograms of (left) entropy-based binning technique produced histogram and (right) ChiMerge-refined entropy-based binning scheme



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

Figure 5. (left) Picture of a Kp time-lag curve (right)time-lag table of ChiMerge-refined entropy-based binning scheme and (in parenthesis) Entropy-based binning scheme.

Possible Limitations

- The time-lag values achieved through these calculations are only characteristic values; they don't signify the exact time-lag value between two parameters at a specific time.
- In order to use these supervised binning techniques, we need a categorical variable. However, our dataset only has continuous data, so arbitrary bins were needed to be created in order to turn a result parameter into a categorical one.

No.	1: sepallength Numeric	2: sepalwidth Numeric	3: petallength Numeric	4: petalwidth Numeric	5: **class** Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa

Figure 6. A dataset with a true categorical variable is shown, dissimilar to the solar wind dataset

Acknowledgements

I would like to thank my mentor Dr. Shing Fung for working with me this summer. I would also like to thank Mr. Jeremy Davis and Mrs. Brittany Whetzel for making this internship 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.

Kerber, Randy, ChiMerge: Discretization of Numeric Attributes. Lockheed Artificial Intelligence Center, 1992.