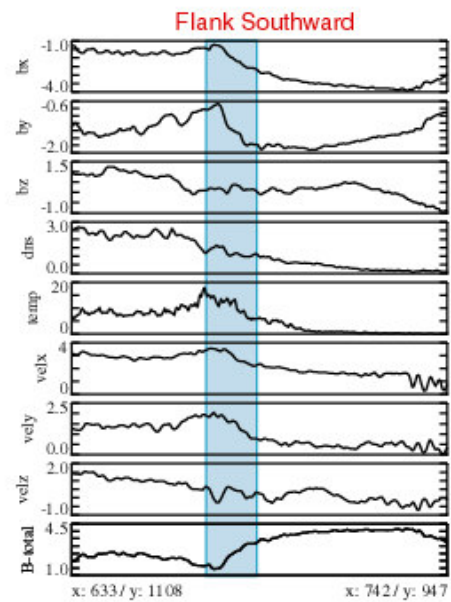
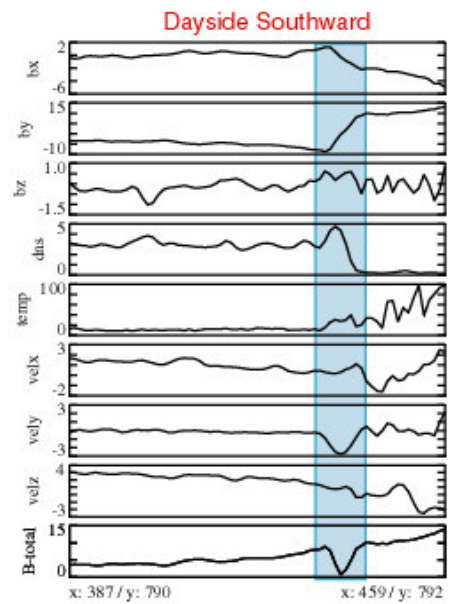
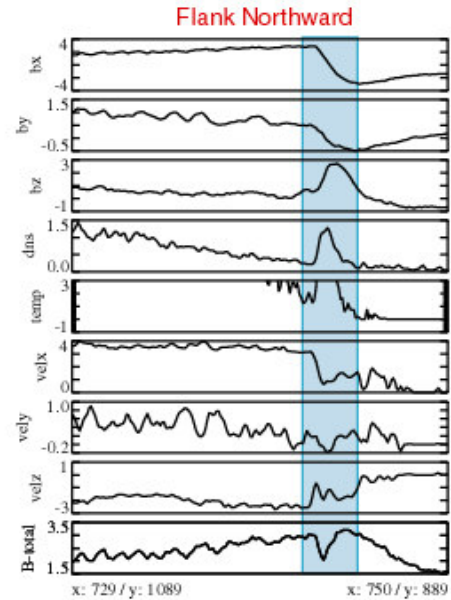
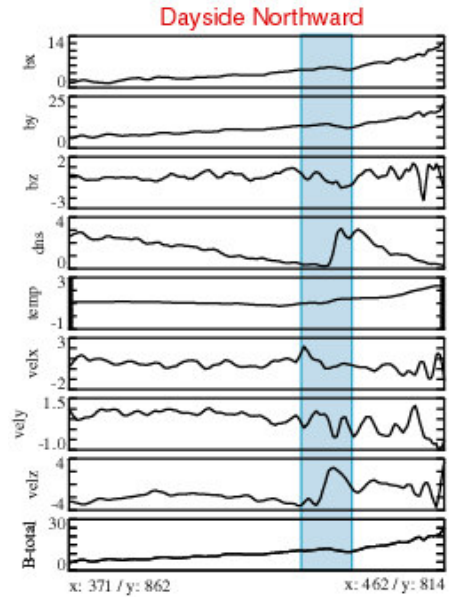


Automated Identification of Magnetopause Crossings

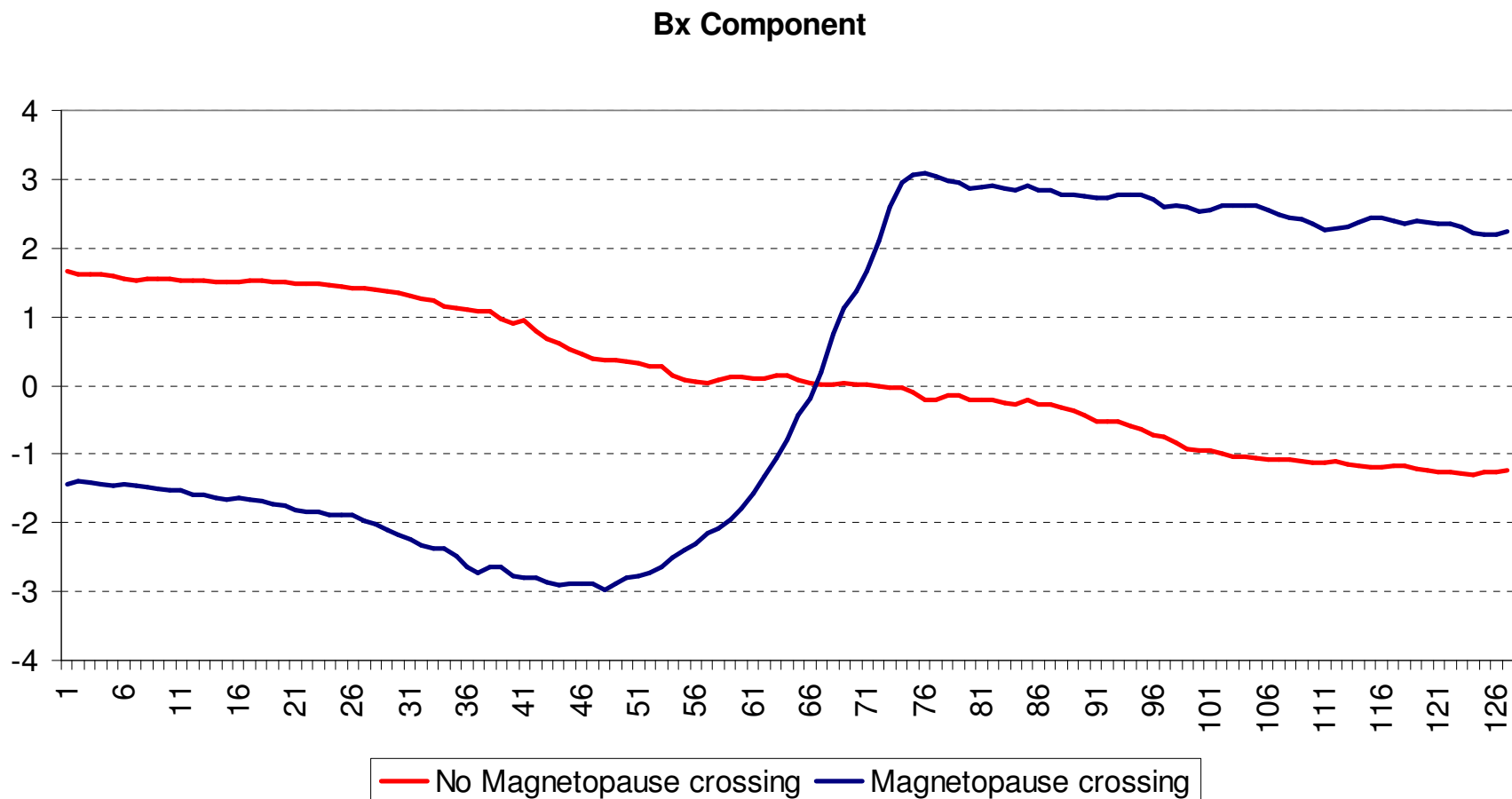
SciberQuest, Inc.

- Data comparison
- Data processing
- Summary statistics
- Modeling process
- Model description
- Model results

Examples of Magnetopause Crossings



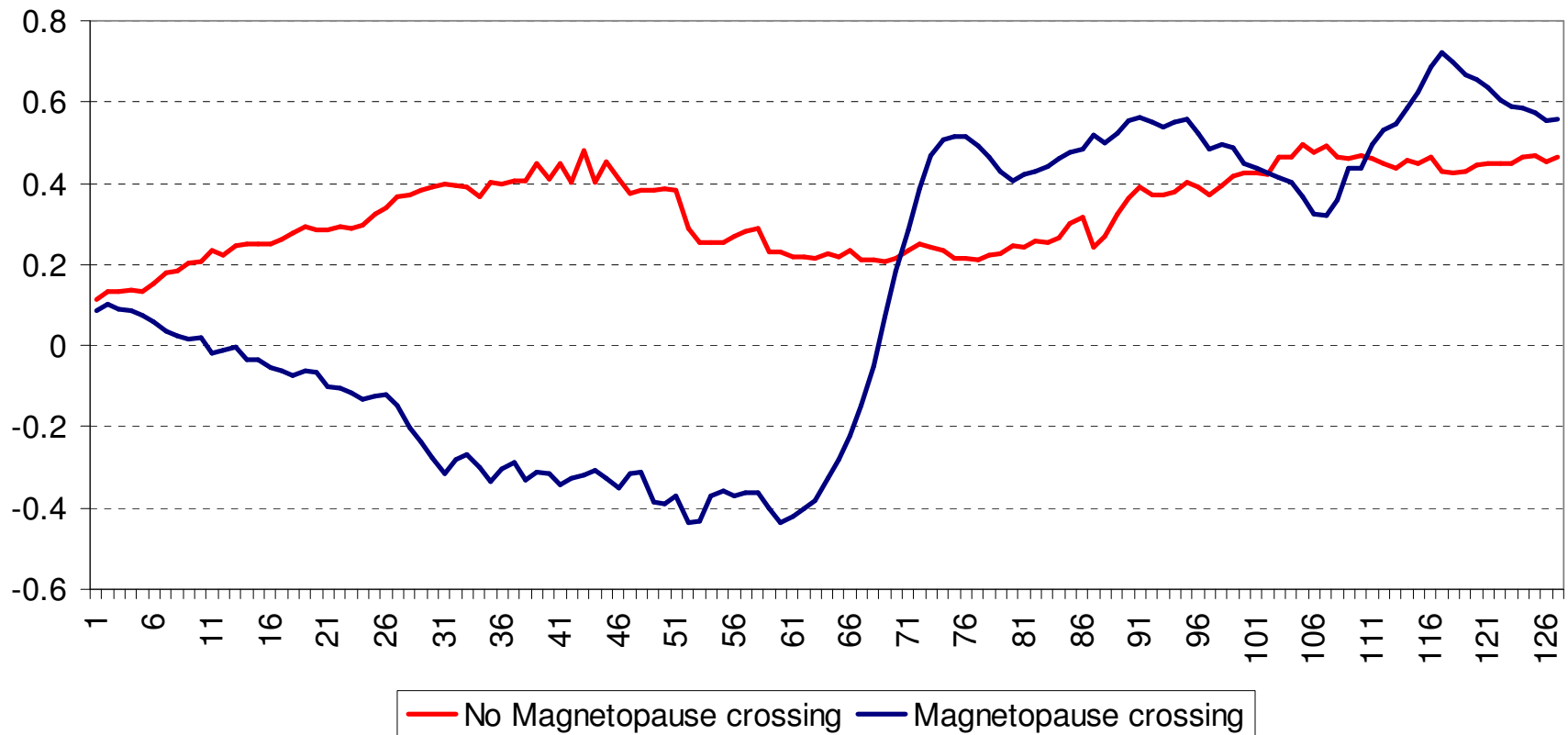
Comparison of Bx: Magnetopause crossing vs. Non-magnetopause crossing



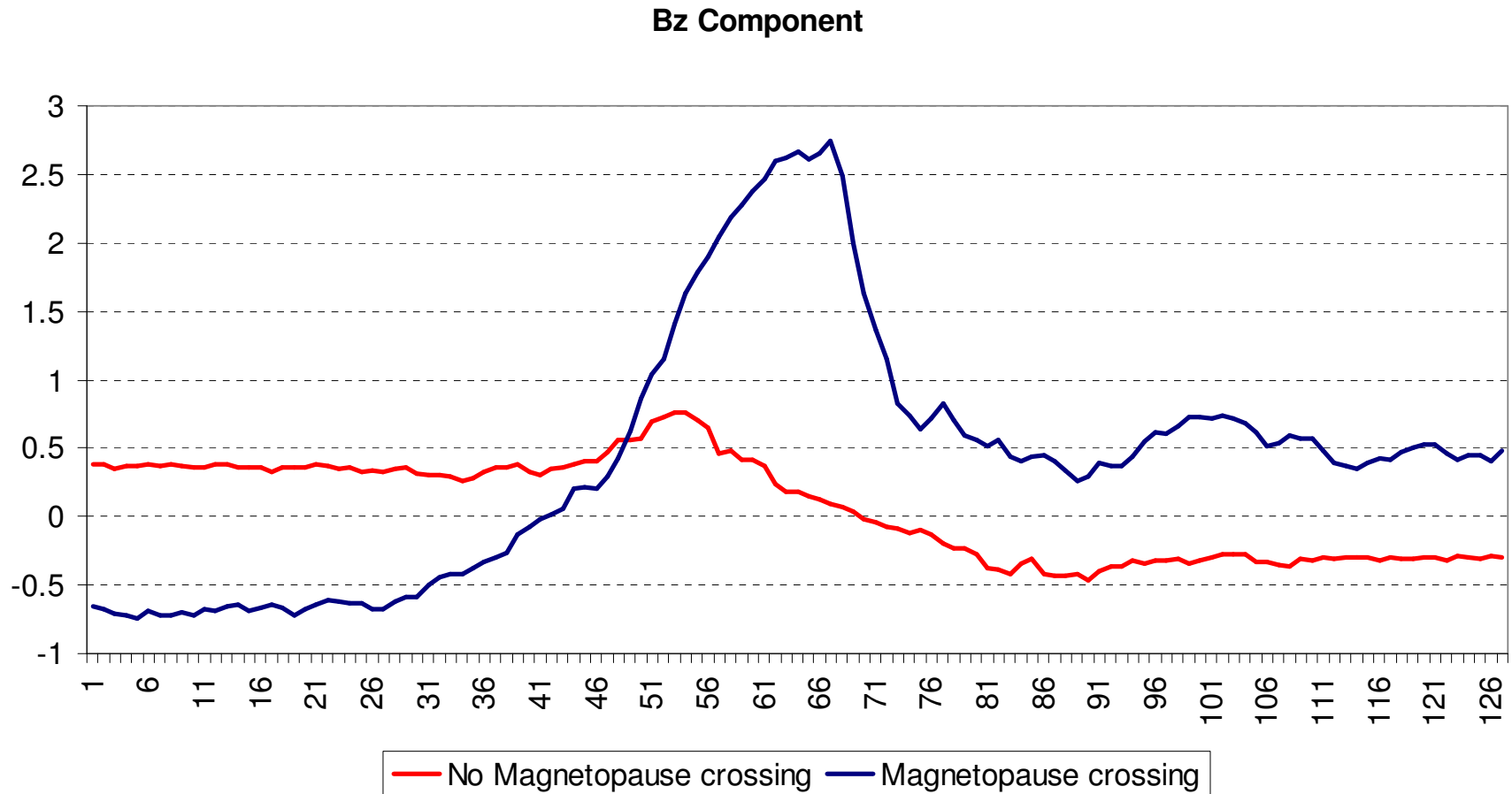
Two observations selected at random, permit us to inspect the characteristics of key components when in the presence of, and in the absence of, a magnetopause

Comparison of By: Magnetopause crossing vs. Non-magnetopause crossing

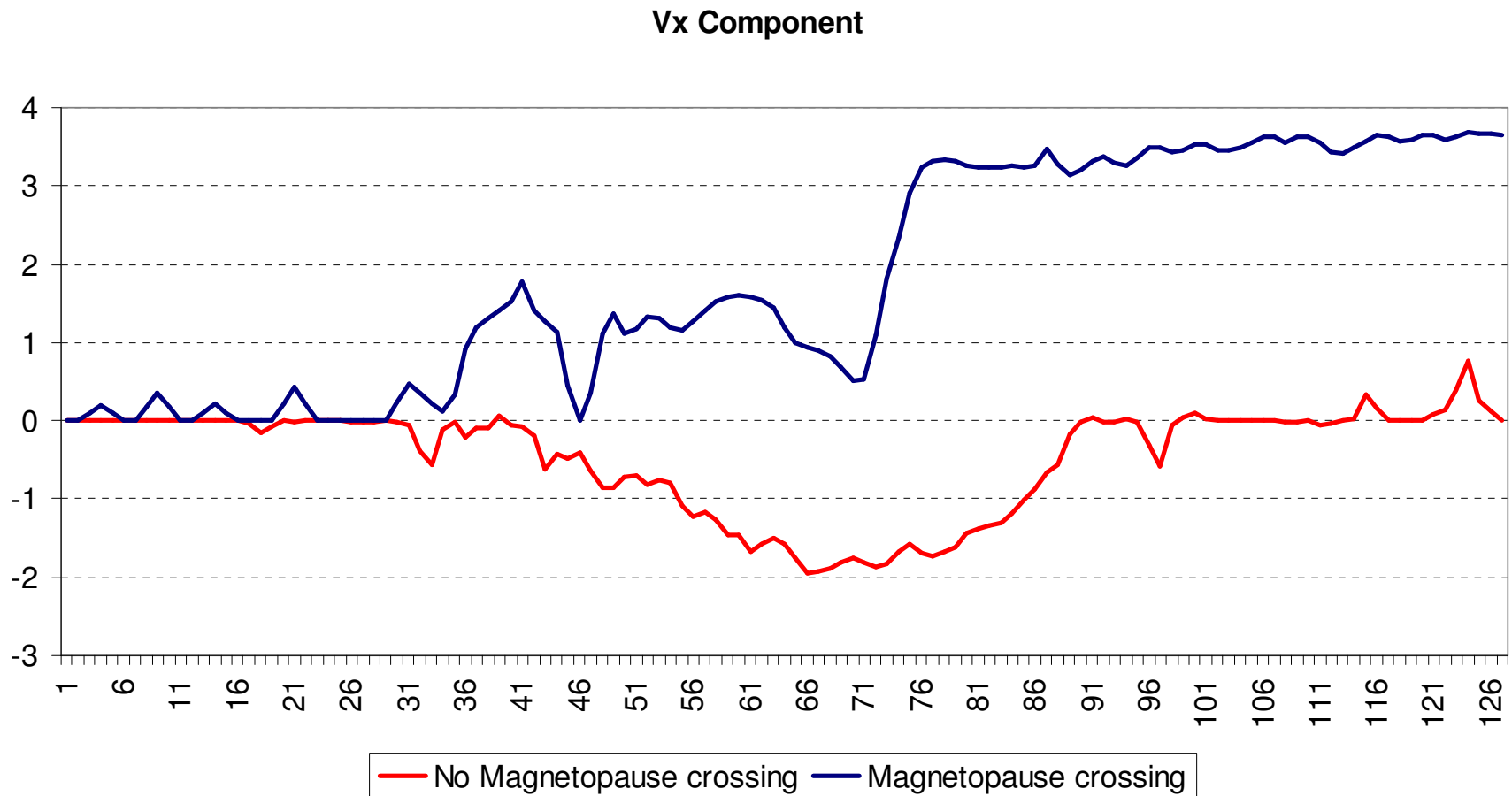
By Component



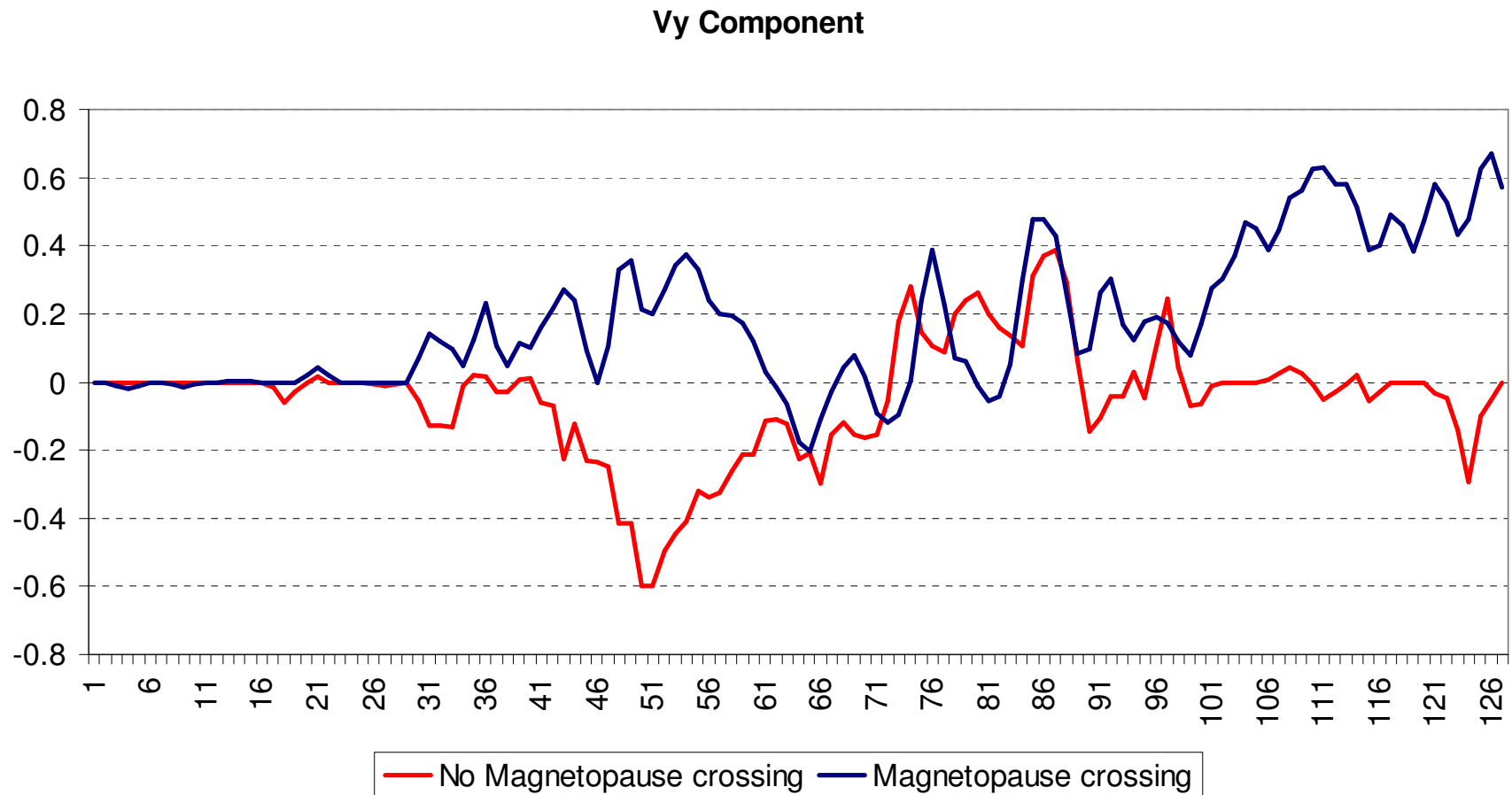
Comparison of Bz: Magnetopause crossing vs. Non-magnetopause crossing



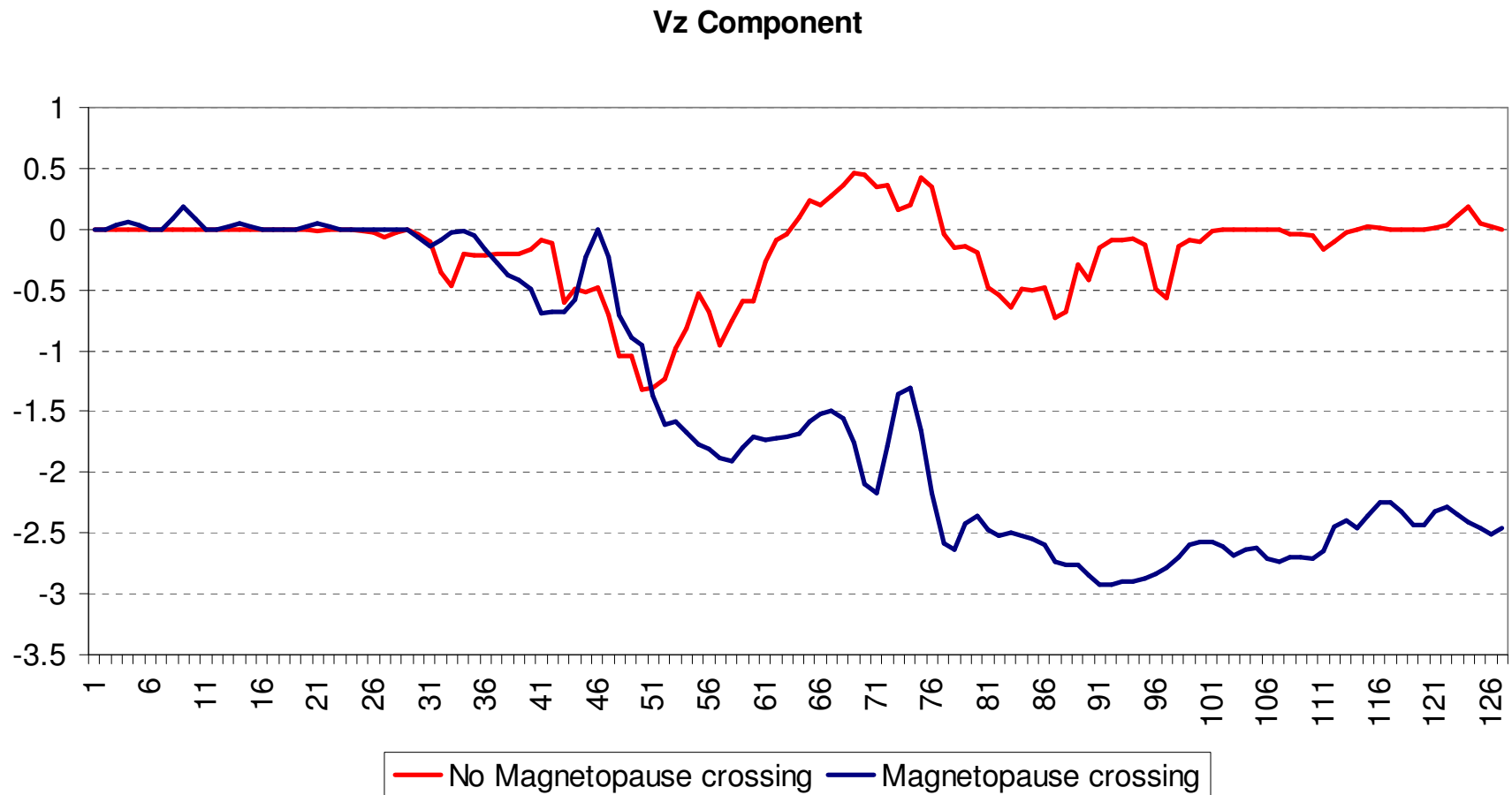
Comparison of Vx: Magnetopause crossing vs. Non-magnetopause crossing



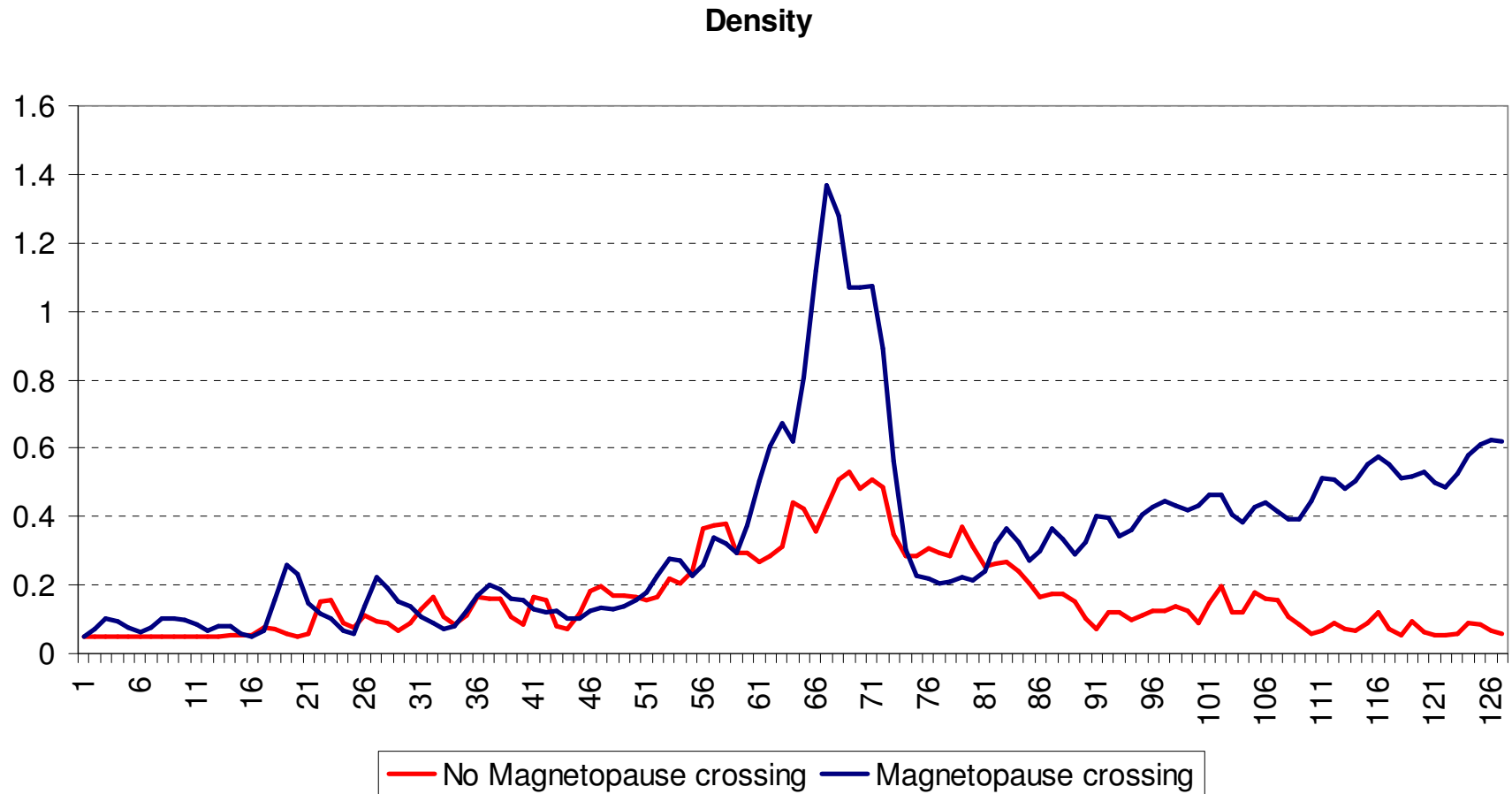
Comparison of Vy: Magnetopause crossing vs. Non-magnetopause crossing



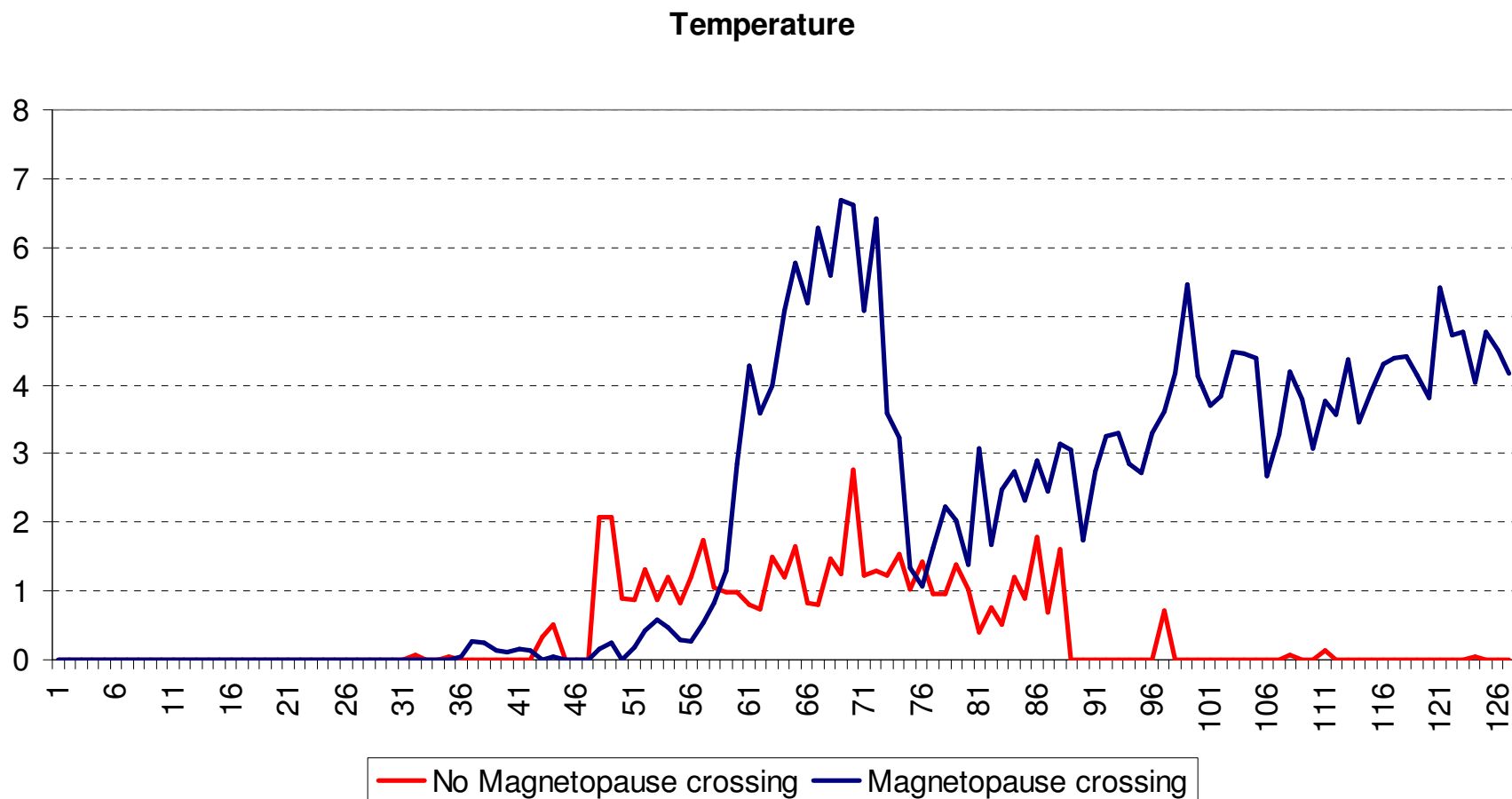
Comparison of Vz: Magnetopause crossing vs. Non-magnetopause crossing



Comparison of density: Magnetopause crossing vs. Non-magnetopause crossing



Comparison of temperature: Magnetopause crossing vs. Non-magnetopause crossing



Data processing

- We added 3 additional components to each observation

- Magnitude of B: $|B| = \sqrt{B_x^2 + B_y^2 + B_z^2}$

- Magnitude of V: $|V| = \sqrt{V_x^2 + V_y^2 + V_z^2}$

- Angle between B and V: $\cos \alpha = \frac{B \cdot V}{|B||V|} = \frac{B_x V_x + B_y V_y + B_z V_z}{\sqrt{B_x^2 + B_y^2 + B_z^2} \sqrt{V_x^2 + V_y^2 + V_z^2}}$

- We obtained characteristics of the different components

- Summary statistics

- Mean
 - Standard deviation
 - Median
 - Interquantile range
 - Minimum and maximum

- Number of times a component value crosses zero

Summary statistics: Mean

Variable	Magnetopause				No Magnetopause			
	Mean	Stand. Dev.	Min.	Max.	Mean	Stand. Dev.	Min.	Max.
Bx	-0.33	2.58	-8.34	8.41	0.54	0.48	-0.06	1.76
By	2.85	4.85	-0.54	18.88	0.88	0.78	0.06	4.43
Bz	0.44	0.38	-0.32	1.43	0.53	0.50	-0.21	1.51
Vx	1.28	0.68	-0.04	2.38	2.59	2.12	-0.75	4.94
Vy	0.07	0.21	-0.53	0.52	0.43	0.41	-0.07	1.63
Vz	-1.16	0.21	-1.72	-0.45	-1.11	0.99	-4.20	0.11
Density	0.61	0.48	0.25	1.86	0.98	0.73	0.13	3.41
Temperature	9.39	13.84	1.29	66.11	3.47	2.86	0.24	13.65
Cos α	-0.08	0.27	-0.61	0.49	0.16	0.17	-0.35	0.69
Magnitude B	5.55	4.77	1.19	18.98	1.61	0.62	0.89	4.61
Magnitude V	2.08	0.35	1.43	2.81	3.28	1.96	0.24	5.39

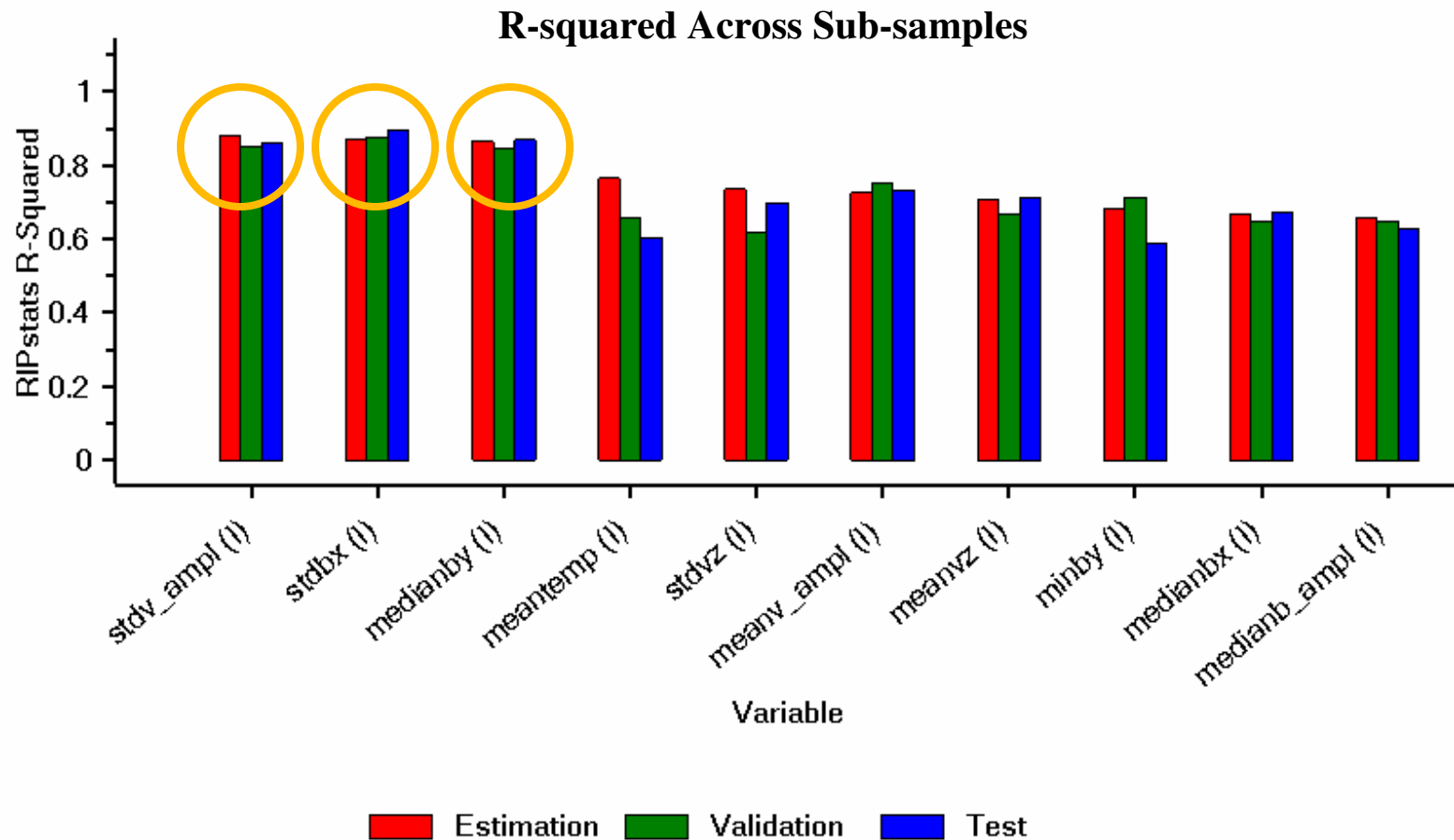
Summary statistics: Standard deviation

Variable	Magnetopause				No Magnetopause			
	Mean	Stand. Dev.	Min.	Max.	Mean	Stand. Dev.	Min.	Max.
Bx	3.83	2.66	1.09	13.76	0.58	0.36	0.02	1.30
By	1.99	8.78	0.074	19.50	0.28	0.36	0.03	1.75
Bz	0.75	0.19	0.30	1.49	0.37	0.21	0.04	0.99
Vx	1.17	0.37	0.29	1.64	0.51	0.42	0.03	2.04
Vy	0.32	0.15	0.09	1.35	0.30	0.21	0.05	1.07
Vz	1.28	0.19	0.86	2.38	0.45	0.47	0.04	2.01
Density	0.45	0.27	0.16	1.08	0.33	0.30	0.07	1.43
Temperature	13.07	22.53	1.52	105.06	1.51	1.66	0.13	8.09
Cos α	0.62	0.20	0.17	0.85	0.38	0.29	0.02	0.93
Magnitude B	2.29	3.87	0.22	19.53	0.41	0.35	0.02	1.80
Magnitude V	1.52	0.33	0.67	1.99	0.41	0.19	0.03	1.27

Steps in the modeling process

- We divided the dataset in three samples
 - Estimation sample
 - It consists of data that are strictly used for the training of the models
 - It contains 50% of the data (2,500 observations)
 - Validation sample
 - It consists of data that are used for cross-validation purposes in order to select the appropriate complexity of the candidate models
 - It contains 25% of the data (1,250 observations)
 - Test sample
 - It consists of pure out-of-sample data used for comparing the performance of the models
 - It contains 25% of the data (1,250 observations)
- The three sample were constructed in such a way that the predictive power of each variable is similar across the samples
- This sampling strategy avoids overfitting the model

Steps in the modeling process (continued)



The comparable R-squared performance of these variables across samples illustrates the strategy of controlling for model overfit

Steps in the modeling process (continued)

- We ran the Anomaly algorithm to detect outliers
 - No outliers were found
- We fit several models, starting from the simplest linear model and then increasing in complexity
 - The benchmark model consists of a linear model in the inputs (which themselves are non-linear transformations of the data)
 - We increased the complexity of the candidate models by adding transformations
 - Cross-products
 - Squares
 - Betas
 - Neural networks units

Steps in the modeling process (continued)

- The model estimation algorithm makes use of the validation sample to select the variables that will be included in each one of the models
- The first variable included in the model is the one presenting the highest R-squared
- Additional variables are selected using an iterative process:
 - Variables with higher R-squared are considered first
 - Each variable considered is selected based on its correlation with the already included ones
 - This method allows to select only variables with additional predictive information

Steps in the modeling process (continued)

- Data Mining Reality Check (DMRC)TM selects the best predictive model
 - This patented algorithm tests whether a candidate model is superior to the benchmark, using a p-value that indicates whether performance gain/loss is statistically significant
 - This test establishes the likelihood that the best candidate model could have occurred by chance; the higher the p-value the more likely the result occurred by chance
 - These p-values are valid

All candidate models perform worse than the benchmark model

Model	Performance (Log-likelihood)	Performance Relative to Benchmark	P-value
Benchmark	-0.008		
Candidate 1	-0.011	-0.003	0.52
Candidate 2	-0.009	-0.001	0.49
Candidate 3	-0.011	-0.003	0.59
Candidate 4	-0.313	-0.305	0.44

DMRC is cited in "A Reality Check for Data Snooping"
Econometrica, vol. 68, num. 5, September 2000, pp 1027-1126

Evaluating model performance using a ROC curve

- Once we obtain the predicted probability of observing a magnetopause crossing, we assess the accuracy of the model with a Receiver Operating Characteristic (ROC) curve
- Now we can classify each observation in one of the two possible categories as a function of different risk-reward trade offs
 - A magnetopause crossing appears
 - A magnetopause crossing does not appear
- Calculating false positive and true positive rates

$$\text{False Positive Rate} = \frac{\text{Total number of (Predicted = 1 and Y=0)}}{\text{Total number of (Y=0)}}$$

$$\text{True Positive Rate} = \frac{\text{Total number of (Predicted = 1 and Y=1)}}{\text{Total number of (Y=1)}}$$

- A *ROC curve* is a graphical representation of the trade-off between the false negative and false positive rates for all possible combinations between 0 and 1

Interpreting the ROC curve

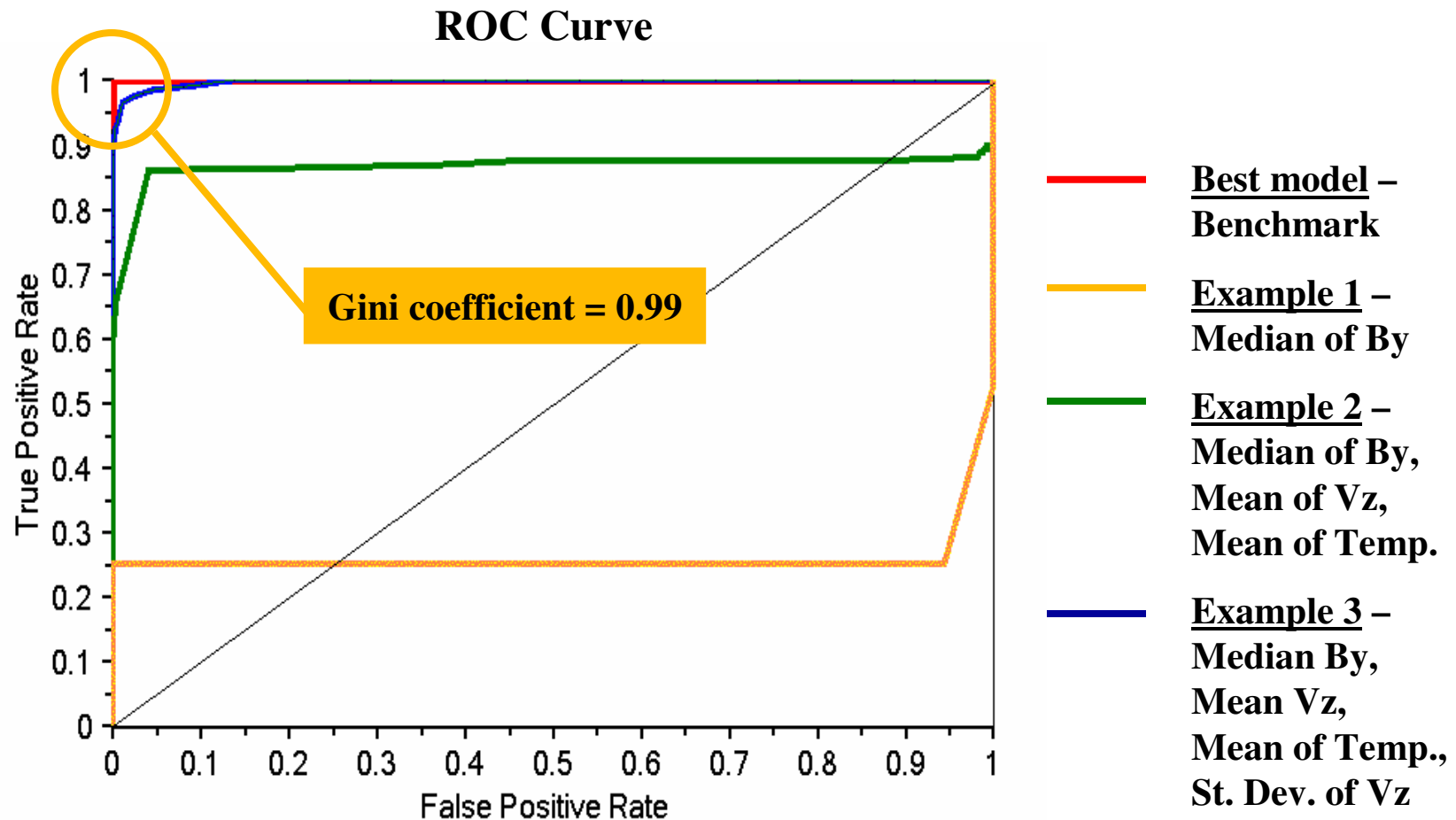
- Along the 45-degree line, the false positive rate is equal to the true positive rate, and therefore such a model is equivalent to flipping a coin, as it predicts correctly 50% of the time
- The *Gini coefficient* is a measure of the area between the ROC curve and the 45-degree line as a percentage of the area of the upper triangle of the graph
- The Gini coefficient provides another measure of model performance
 - A Gini coefficient equal to 1 indicates that the model is predicting perfectly
 - A Gini coefficient close to 0 indicates that the model is predicting no better than random chance

Model results: Inputs from the best candidate model

- The best model is the benchmark
- RDMS selected the following ten explanatory variables

Variable	R-squared	Variable	R-squared
St. Dev. of Magnitude of V	0.88	Mean of Magnitude of V	0.72
St. Dev. of Vz	0.73	Mean of Temperature	0.76
St. Dev. of Bx	0.87	Minimum of By	0.68
Median of Bx	0.66	Median of By	0.86
Median of Magnitude of V	0.65	Mean of Vz	0.71

Comparing the relative performance of the models



Model estimation performance: execution time

- Using RDMS, model estimation results are available in $1/20^{\text{th}}$ – $1/50^{\text{th}}$ of the time of a traditional artificial neural network (ANN)
 - Total modeling processing time = 20 minutes
 - Total model estimation time = 7 seconds