

Reverse Engineering the Shape and Position of Magnetopause

H. Karimabadi, L. Valera-Guallar, H. White

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

Agenda

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

Data description

- We received simulated data for the magnetopause distance (R)
- The dataset contains 10,800 observations and 4 variables:
 - R: Magnetopause distance
 - Dp: Solar wind pressure
 - Bz: Magnetic field
 - Theta: Cone angle
- The objective is to estimate a magnetopause distance prediction model

Data processing

- We perform the following transformations to the data:
 - Logarithm of R (used as target variable)
 - Logarithm of D_p
 - Cosine of theta
 - Sine of theta
 - Indicator Bz is equal to 0
 - Indicator Bz is positive
- RDMS creates additional transformations
 - Cross-products
 - Squares

Summary statistics

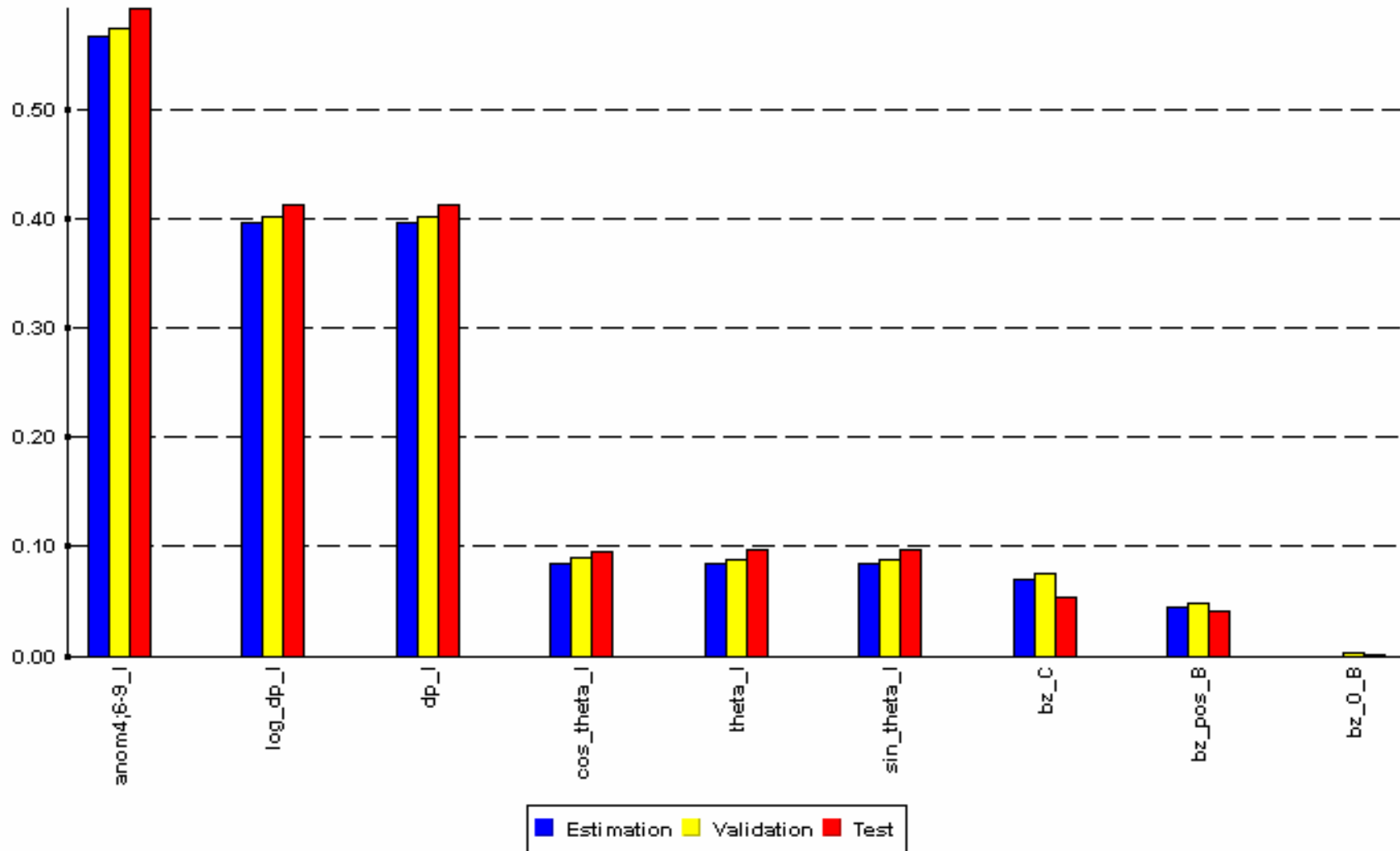
Variable	Number of Obs	Mean	Stand. Deviation	Minimum	Maximum
R	10,800	8.57	1.93	4.75	17.11
Log (R)	10,800	2.12	0.22	1.56	2.84
Bz	10,800	0.00	12.91	-20.00	20.00
Dp	10,800	12.98	13.56	1.00	50.00
θ	10,800	21.36	24.36	0.00	90.00
Log(Dp)	10,800	1.96	1.17	0.00	3.91
Cos(θ)	10,800	0.86	0.25	0.00	1.00
Sin(θ)	10,800	0.32	0.32	0.00	1.00

Steps in the modeling process

- We divided the dataset into three samples
 - Estimation sample
 - Consists of data that are strictly used for training models
 - Contains 50% of the data (5,400 observations)
 - Validation sample
 - Consists of cross-validation data that are used to select the appropriate complexity of the candidate models
 - Contains 25% of the data (2,700 observations)
 - Test sample
 - Consists of out-of-sample data used to compare relative model performance
 - Contains 25% of the data (2,700 observations)
- We constructed the three samples 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)

Univariate R-squared Across Sub-samples



Steps in the modeling process (continued)

- We ran the Anomaly algorithm to detect outliers
 - No outliers were found
- We estimated several models
 - We started with a simple linear model, and then increased model complexity
 - The benchmark model consists of a linear model in the inputs
 - The inputs themselves are non-linear transformations of the data
 - We increased the complexity of the candidate models by adding transformations
 - Cross-products
 - Squares
 - Neural networks units
 - Anomaly variable

Steps in the modeling process (continued)

- RDMS uses the validation sample to select variables that will be included in each model
- The first variable included in the model is the one with the highest R-squared
- Additional variables are selected using an iterative process:
 - Variables with higher R-squared are considered first
 - Each candidate variable is selected based on its correlation with variables already included in the model
 - This method allows us to select only variables that provide additional predictive information, based on the Partial R-squared measure

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
 - Uses a p-value to indicate whether performance gain/loss is statistically significant
 - DRMCS 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

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

Model results

- Our benchmark consists of a simple model that includes only the original variables introduced into RDMS
- The best model is a non-linear model that includes 25 variables:
 - Original variables introduced into RDMS
 - Anomaly variable, constructed from the following variables:
 - Dp
 - Theta
 - Cosine of theta
 - Sine of theta
 - Logarithm of Dp
 - Cross products
 - Square of original variables

Model results: Inputs from the benchmark model

Variable	Partial R-squared	Coefficient
Log(Dp)	0.9475	-0.0516
Cosine of theta	0.0003	0.0503
Bz	0.0002	-0.6414

Model results: Inputs from the best candidate model

- RDMS selected the following top ten explanatory variables

Variable	Partial R-squared	Coefficient
Anomaly variable	0.9998	0.0046
Log(Dp)	0.7096	-0.0039
Dp	0.8019	-0.0232
Anomaly*(Bz > 0)	0.5650	-0.0035
Cosine of theta	0.4869	-0.5656
(Sin of theta)*theta	0.8915	0.0294
(Cosine of theta) ²	0.8828	-0.0004
(Bz > 0)*theta	0.4664	-0.0026
Bz	0.3533	0.0130
Bz > 0	0.7981	-0.0146

Model results: Inputs from the best candidate model (continued)

Variable	Partial R-squared	Coefficient
Anomaly*Dp	0.7318	-0.5684
Anomaly*(Bz = 0)	0.1980	0.0065
(Bz > 0)*Dp	0.8924	0.3097
(Bz > 0)*Bz	0.8199	0.0638
Anomaly* Anomaly	0.8786	0.0006
(Bz = 0)*theta	0.4224	0.0083
Anomaly*theta	0.8708	-0.0089
(Bz = 0)*(Bz > 0)	0.6956	-0.0106
(Bz = 0)*Dp	0.8577	0.3040
Cosine of theta*Bz	0.7848	-0.0354

Model results: Inputs from the best candidate model (continued)

Variable	Partial R-squared	Coefficient
Sine of theta*Bz	0.8712	-0.0144
Cosine of theta*Dp	0.8139	-0.0989
Log(Dp)*theta	0.8918	-0.0026
Anomaly*Bz	0.8542	0.0091
Log(Dp)*Bz	0.8829	-0.9819

Evaluating model performance

- We consider three different error measures to compare performance across models
 - Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{predicted}_i - \text{actual}_i)^2}$$

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |\text{predicted}_i - \text{actual}_i|$$

- Mean Relative Error (MRE):

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{\text{predicted}_i - \text{actual}_i}{\text{actual}_i}$$

Comparing performance

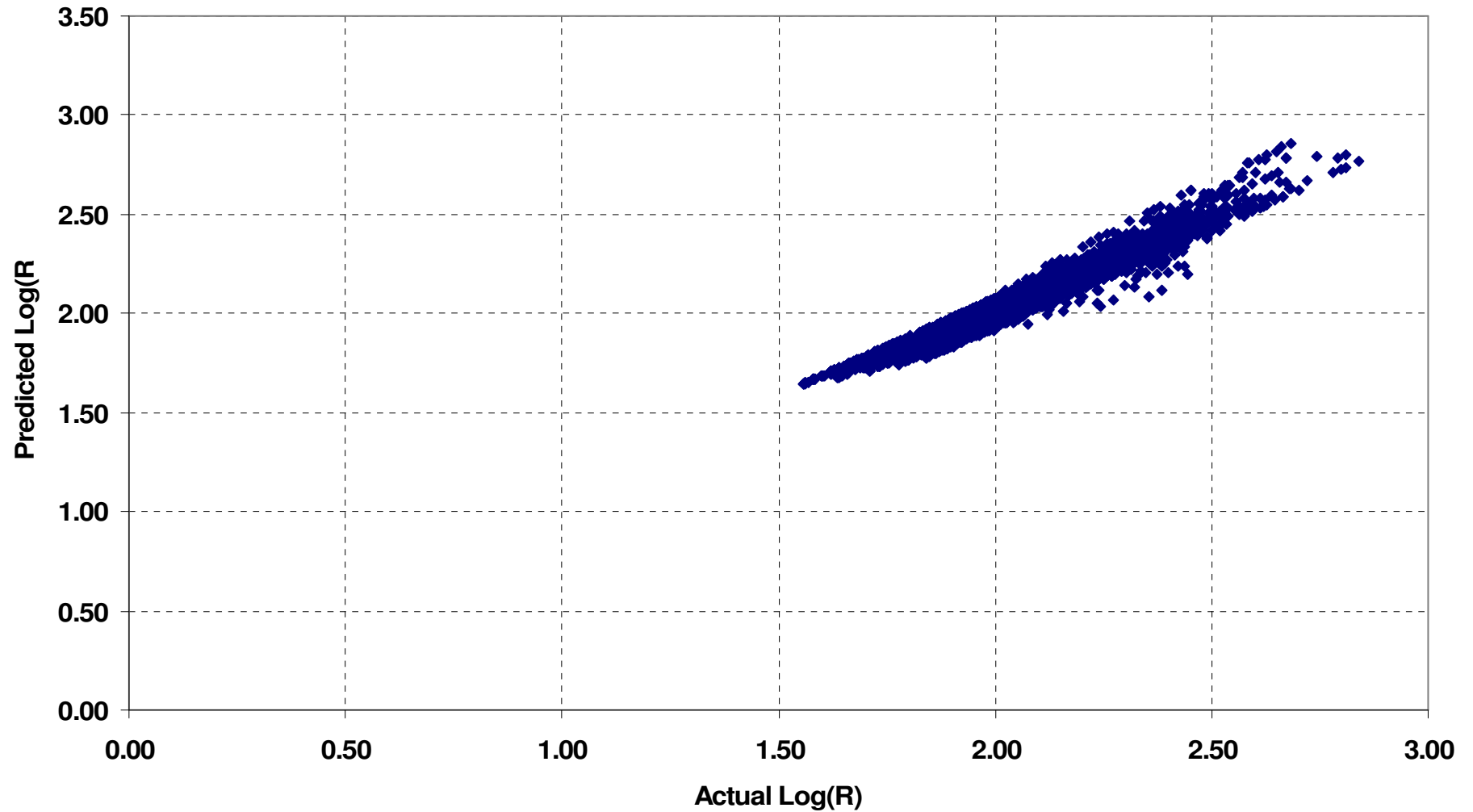
- The following table compares the performance of the benchmark model and our best candidate model for the out-of-sample data

Performance Measure	Benchmark Model	Best Model	Error Reduction
RMSE	0.002729	0.000013	99.51%
MAE	0.042116	0.002356	94.41%
MRE	0.000171	0.000034	80.37%

- The best model performs dramatically better than the benchmark model
 - MAE indicates an error reduction of 94%

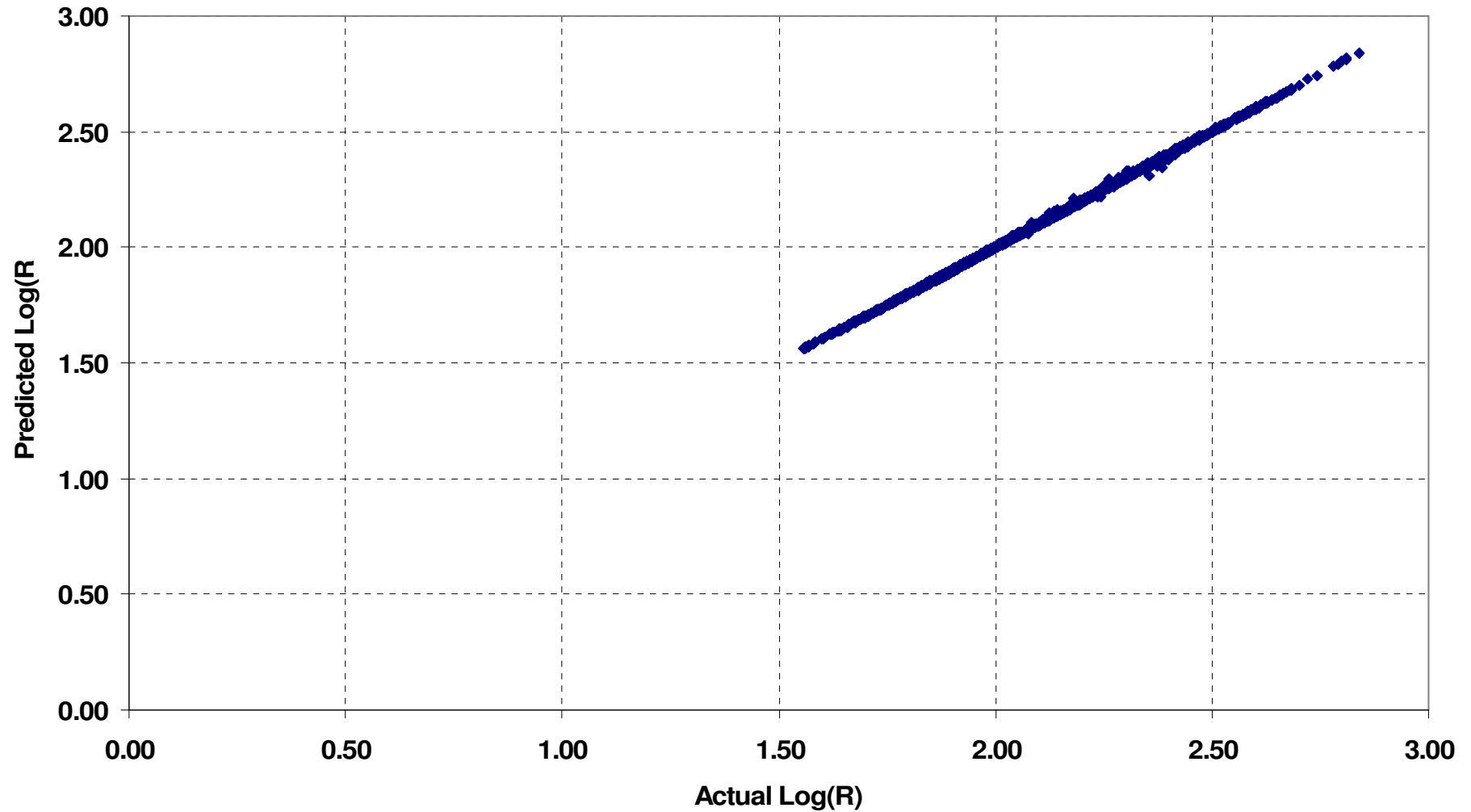
Comparing performance – Benchmark predicted values

Predicted vs. Actual Values of Magnetopause Distance
Benchmark Model



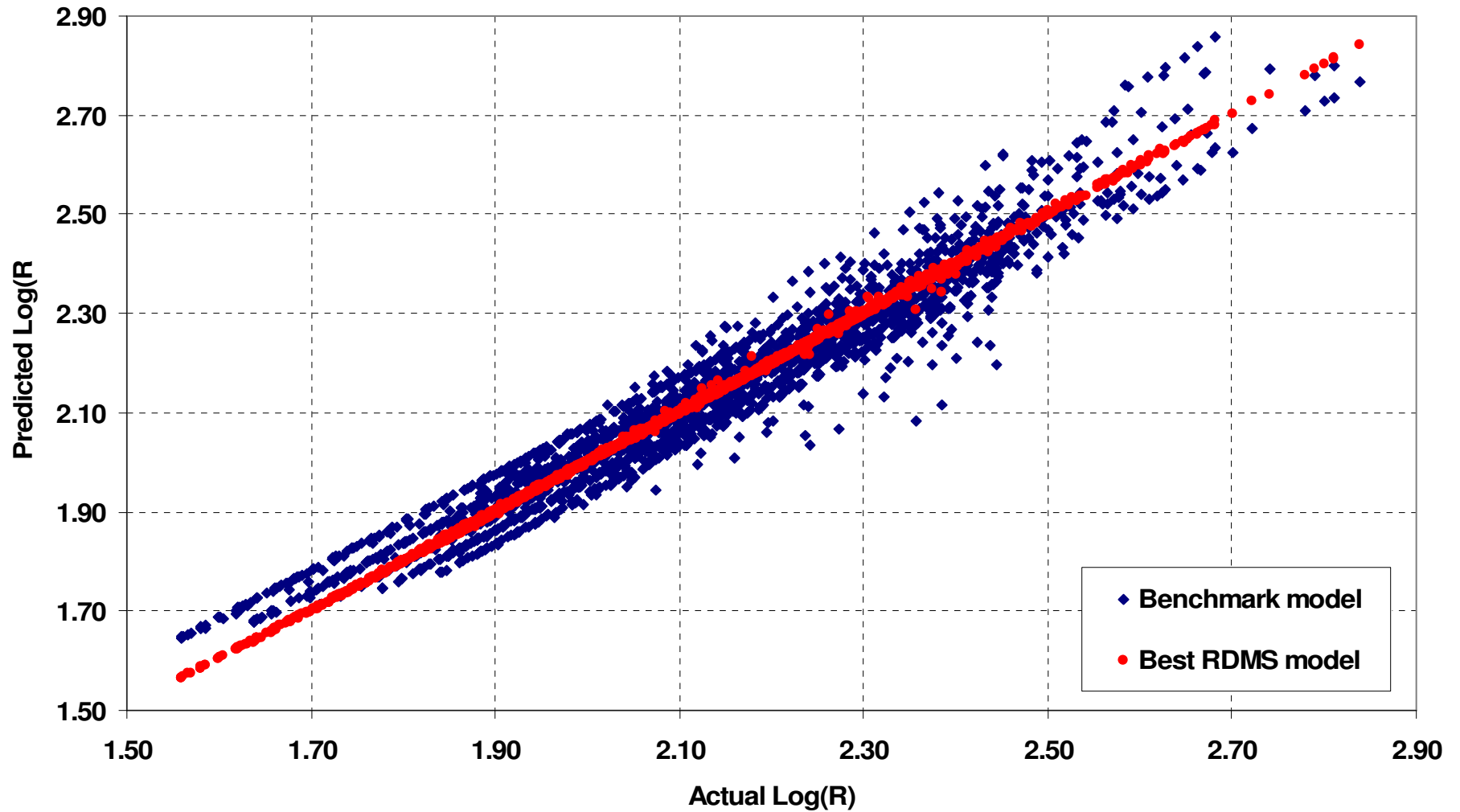
Comparing performance – Best model predicted values

Predicted vs. Actual Values of Magnetopause Distance
Preferred Model



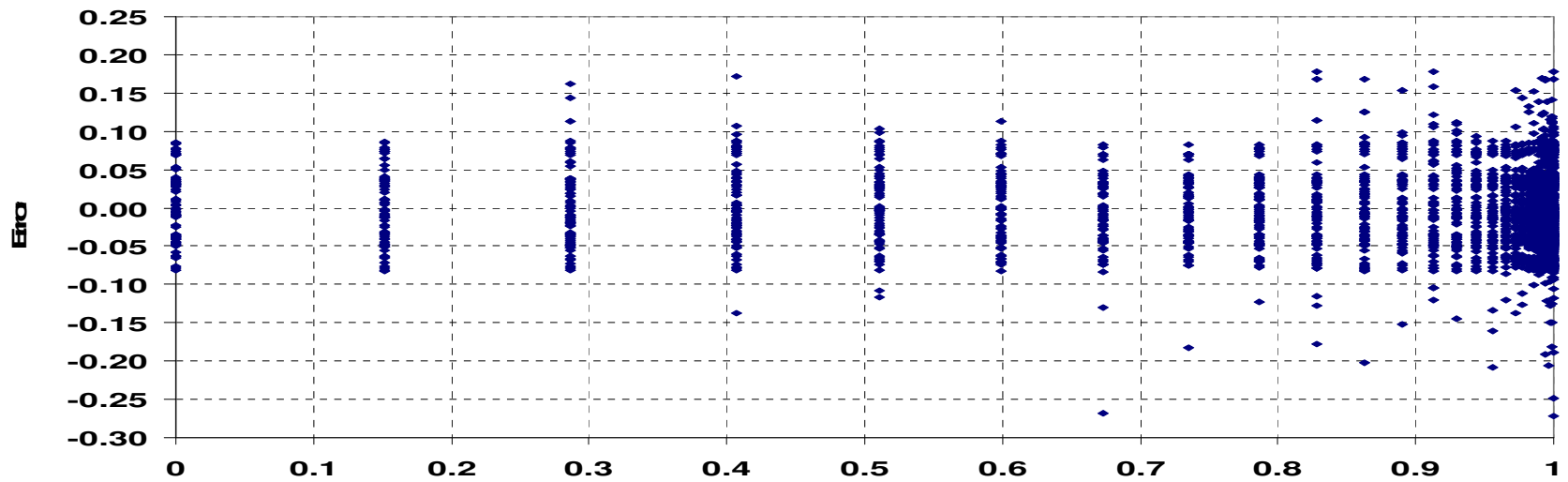
Comparing performance – Best model predicted values

Predicted vs. Actual Values of Magnetopause Distance
Benchmark and Preferred Model

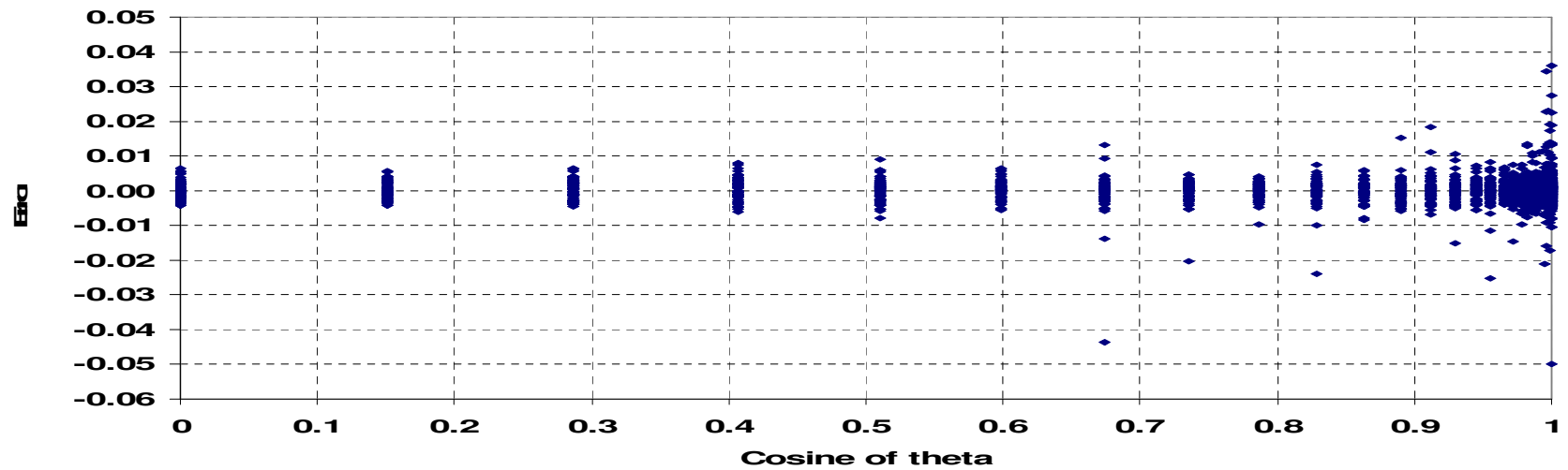


Prediction error as a function of dependent variables – Cosine of theta

Benchmark model

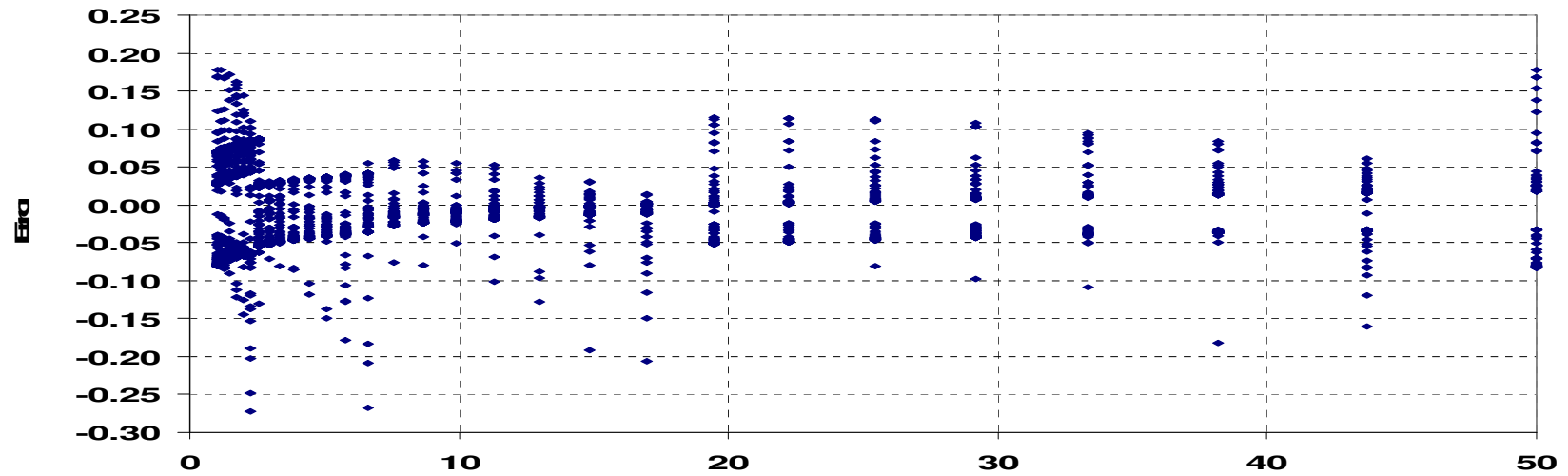


Preferred model

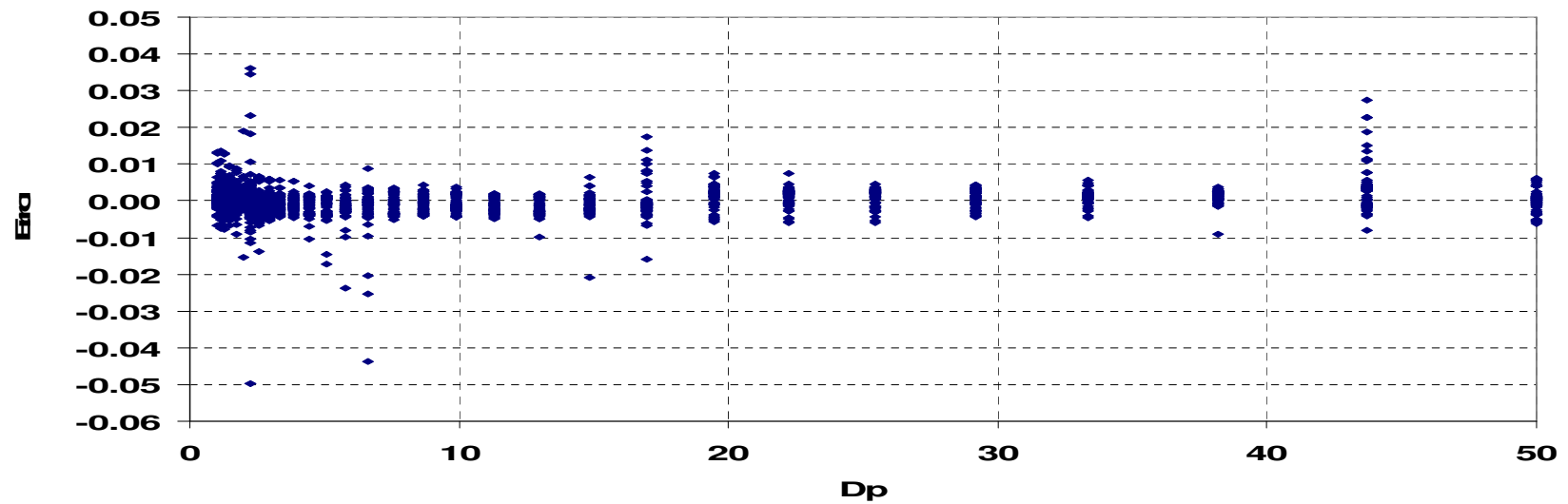


Prediction error as a function of dependent variables – Dp

Benchmark model

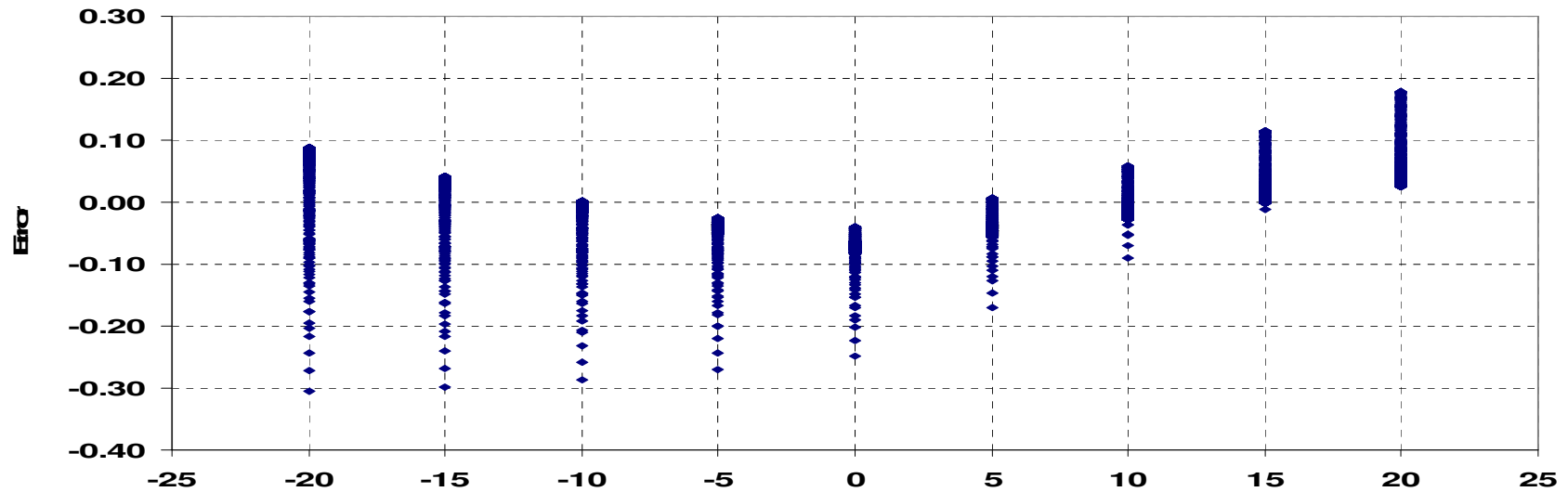


Preferred model



Prediction error as a function of dependent variables – Bz

Benchmark model



Preferred model

