AIAA Ground Testing Conference
24 – 27 June 2013, San Diego, California

AIAA 2013-XXXX

Analysis of Multivariate Experimental Data Using A Simplified Regression Model Search Algorithm

(Extended Abstract of Proposed Conference Paper)

N. Ulbrich*

Jacobs Technology Inc., Moffett Field, California 94035–1000

A new regression model search algorithm was developed in 2011 that may be used to analyze both general multivariate experimental data sets and wind tunnel strain–gage balance calibration data. The new algorithm is a simplified version of a more complex search algorithm that was originally developed at the NASA Ames Balance Calibration Laboratory. The new algorithm has the advantage that it needs only about one tenth of the original algorithm's CPU time for the completion of a search. In addition, extensive testing showed that the prediction accuracy of math models obtained from the simplified algorithm is similar to the prediction accuracy of math models obtained from the original algorithm. The simplified algorithm, however, cannot guarantee that search constraints related to a set of statistical quality requirements are always satisfied in the optimized regression models. Therefore, the simplified search algorithm is not intended to replace the original search algorithm. Instead, it may be used to generate an alternate optimized regression model of experimental data whenever the application of the original search algorithm either fails or requires too much CPU time. Data from a machine calibration of NASA's MK40 force balance is used to illustrate the application of the new regression model search algorithm.

Nomenclature

AF	= axial force
$N1$	= normal force at the forward normal force gage of the balance
$N2$	= normal force at the aft normal force gage of the balance
$R1$	= electrical outputs of the forward normal force gage
$R2$	= electrical outputs of the aft normal force gage
$R3$	= electrical outputs of the forward side force gage
$R4$	= electrical outputs of the aft side force gage
$R5$	= electrical outputs of the rolling moment gage
$R6$	= electrical outputs of the axial force gage
RM	= rolling moment
$S1$	= side force at the forward side force gage of the balance
$S2$	= side force at the aft side force gage of the balance

I. Introduction

Between 2005 and 2009 a regression model search algorithm was developed for the NASA Ames Balance Calibration Laboratory. The search algorithm identifies regression models of multivariate experimental data sets that meet strict statistical quality requirements and prevent "over–fitting" of the dependent variable (see, e.g., Refs. [1] to [5] for a description of different aspects of the original search algorithm). The original

* Aerodynamicist, Jacobs Technology Inc.

search algorithm was implemented in NASA's BALFIT software package (see Ref. [6]). Currently BALFIT and the search algorithm are applied on a regular basis to wind tunnel strain–gage balance calibration data sets at both NASA Ames Research Center and NASA Langley Research Center.

Experience showed that the original regression model search algorithm is not always able to successfully complete a search. The failure of the search is usually associated with the detection of a discontinuity in the algorithm's search metric. Several causes for the failure of a regression model search were identified. Most of them were related (i) to limitations of the regression model search algorithm itself or (ii) to shortcomings and imperfections in experimental data sets. Some of the limitations of the original search algorithm were removed by giving a user of the algorithm a greater choice of (i) math term group combinations, (ii) search constraint options, and (iii) search directions. In addition, improvements of BALFIT's analysis report format were made so that shortcomings in experimental data sets can more easily be spotted.

In 2010, after analyzing problems associated with some wind tunnel strain–gage balance calibration data, it was concluded that the application of the original search algorithm is also sometimes unsuccessful because tare corrections have to be applied to some balance data sets before the regression model search can be started (see Ref. [7] for a detailed description of the tare load iteration process). Unfortunately, an accurate tare correction assessment can often only be obtained if a very good estimate of the final optimized regression model of the balance calibration data is already known ("chicken–and–egg" problem). Therefore, the author attempted to simplify several elements of the original regression model search algorithm such that the CPU time needed for the search would be significantly reduced while still enforcing the three search constraints of the original search algorithm. These efforts lead to the development of the simplified regression model search algorithm that is discussed in the present paper.

The most important elements of the simplified regression model search algorithm are briefly summarized in the next section of the proposed conference paper. Then, data from the machine calibration of a wind tunnel strain–gage balance is used to assess (i) the CPU time needed by the simplified algorithm for the completion of a regression model search and (ii) the predictive capability of its optimized regression models.

II. Simplified Search Algorithm

The simplified regression model search algorithm was derived from a more complex algorithm that was developed for the NASA Ames Balance Calibration Laboratory between 2005 and 2010 (see Refs. [1] to [4] for details about the original more complex algorithm). The development of the simplified search algorithm became necessary because some wind tunnel strain–gage balance calibration data sets require tare corrections before the original more complex search algorithm can be applied. Accurate tare corrections, however, can often only be obtained if already a good "guess" of the final optimized regression model is available before the beginning of the search (chicken–and–egg problem). This "guess", in order to be useful, has to be determined within a fraction of the CPU time that is needed for the original more complex search process. The simplified search algorithm fulfills all these requirements.

The simplified search algorithm has matured since it was first developed in 2011. It is now included in NASA's BALFIT software package as a new "Math Model Type" choice that an analyst can make (i.e., the "Math Model Type" choice "Suggested Math Model"). The simplified search algorithm uses many of the statistical quality metrics and search constraints that the original search algorithm applies (see Ref. [5] for a general discussion of different metrics that may be used to evaluate a regression model of experimental data). However, the metrics and constraints are applied in a simplified fashion in order to greatly reduce the total number of math models that have to be tested during the regression model search.

Figure 1 summarizes basic elements of the simplified regression model search algorithm. The search starts by first selecting a combination of math term groups (function classes) that help define an upper bound for the regression model search. The math term group combination is usually selected by using subject matter knowledge about the given experimental data set. The chosen combination, e.g., may consist of linear terms, cross–product terms, square terms, and absolute value terms. Not every term of the chosen math term groups may be supported by the given experimental data set. Therefore, the upper bound, i.e., the *Permitted Math Model*, needs to be determined for the regression model search such that only non–singular solutions of the related least squares problem exist. This is accomplished by using a numerical technique called Singular Value Decomposition (SVD). In the next step, the variance inflation factors of the terms of the *Permitted Math Model* are inspected to make sure that the upper bound does not have unwanted near–linear dependencies. Terms are typically omitted in the *Permitted Math Model* that exceed the literature recommended variance inflation factor threshold of 10. It is important to point out that the simplified search algorithm removes terms such that the *Hierarchy Rule* is still satisfied. In other words, each term of the *Permitted Math Model* will still have all lower order terms after the omission of terms that violate the variance inflation factor threshold. In the next step, the revised upper bound for the simplified search is applied to the experimental data set and the p–values of the t–statistic of each regression model coefficient are computed. Now, "insignificant" terms, i.e., terms that have a p–value of greater than 0.0001 are removed in order to obtain the final optimized regression model. Again, the term removal is performed such that the *Hierarchy Rule* is maintained in the final optimized regression model.

A more detailed comparison of the simplified and more complex regression model search algorithms will be included in the final manuscript of the proposed conference paper. This discussion will also compare the CPU time needed for the regression model search for different math term group combinations and other search parameter that the simplified and more complex search algorithm have in common. In the next section of the extended abstract a data example is discussed to illustrate different aspects of the application of the simplified regression model search algorithm.

III. Discussion of Example

Data from the calibration of the NASA Ames MK40 force balance was chosen to illustrate steps needed for the application of the simplified regression model search algorithm. The MK40 is a wind tunnel strain–gage balance that was originally manufactured by the Task Corporation. It is a six–component force balance that measures five forces and one moment ($N1$, $N2$, $S1$, $S2$, AF, RM). The balance has a diameter of 2.5 inches and a total length of 17.31 inches. Table 1 shows the load capacity of each load component.

Table 1: Load capacities of the NASA Ames 2.5in MK40 balance.

	$N1$, *lbs*	$N2$, *lbs*	$S1$, *lbs*	$S2$, *lbs*	RM, *in–lbs*	AF, *lbs*
CAPACITY	3500	3500	2500	2500	8000	400

The balance calibration data was obtained in 2008 in Triumph Aerospace's balance calibration machine. The supplied machine calibration data was already corrected for the weight of the balance shell and other calibration fixtures. Therefore, the tare load iteration process was omitted during the analysis of the data. Table 2 below summarizes important characteristics of the balance and its calibration data set.

Table 2: Balance and calibration data set characteristics of the MK40 balance.

BALANCE NAME	MK40
BALANCE TYPE	FORCE BALANCE (TASK DESIGN)
DIAMETER	2.5 [in]
CALIBRATION DATE	JULY 2008
CALIBRATION METHOD	MACHINE CALIBRATION
TOTAL NUMBER OF CALIBRATION POINTS	1863
LOAD FORMAT	TARE CORRECTED LOADS
GAGE OUTPUT FORMAT	GAGE OUTPUT DIFFERENCES

First, in order to start the regression model search, it was necessary to define an upper bound, i.e., the Permitted Math Model. The balance is a multi–piece balance (TASK design). Therefore, as the data is to be processed using the *Iterative Method* recommended in Ref. [7], it was decided to include absolute value terms in the math term group combination for the regression models of the six gage outputs of the balance.

The following groups were chosen for the regression model of the gage outputs

$$F,\ |F|,\ F \cdot F,\ F \cdot |F|,\ F \cdot G$$

where F and G are symbols representing the load components of the balance. The balance calibration data set was obtained in a calibration machine that supports all load combinations that can be constructed from the five chosen math term groups. Therefore, after including the intercept in the regression models of the gage outputs, the initial upper bound for the search has a total of 40 terms for each component. The corresponding regression models of the gage outputs $R1$ to $R6$ are given in Fig. 2.

Specific knowledge about the characteristics of a TASK balance may be used to reduce the number of regression model terms in the initial upper bound that is given in Fig. 2. In a recent paper it was shown that not all gages of a TASK balance have the bi–directional characteristic that traditionally justifies the selection of absolute value terms for a TASK balance (see Ref. [8]). Figure 3a shows, for example, the bi–directionality of the MK40 balance. It can clearly be seen that the rolling moment gage $R5$ and the axial force gage $R6$ are not bi–directional (cf. last two plots in Fig. 3a). Consequently, there is no justification for the use of the terms $|AF|$ and $|RM|$ in the regression models of the gage outputs of the MK40 balance. Then, after removing the related terms from the initial upper bound, a new 36–term upper bound of the math model is obtained that is shown in Fig. 3b.

An initial analysis of the variance inflation factors of the 36–term upper bound for the regression model of the forward normal force gage output $R1$ is shown in Fig. 3c. A near–linear dependency between the absolute value terms of the normal/side force components and corresponding quadratic terms appears to exist as terms of those two function classes have variance inflation factors that exceed the literature recommended threshold of 10 (see red boxes in Fig. 3c). The connection between the absolute value and quadratic terms comes from the fact that both the absolute value function and the quadratic function are even. The variance inflation factor analysis for the MK40 data set indicates that either the absolute value terms or the quadratic terms of the normal/side force components can be omitted in the regression models. A TASK balance is, by design, a very linear device. Therefore, it was decided to omit the quadratic terms of the normal and side force components. The corresponding second revision of the upper bound for the regression model search is shown in Fig. 3d.

Now, the original and the simplified search algorithms can be applied to the MK40 data set by using the math model shown in Fig. 3d as the upper bound for the regression models of the gage outputs. Figure 4a shows the optimized models of the gage outputs that were obtained after applying the original search algorithm. It took approximately 28 minutes of CPU time to complete the search. Figure 4b shows the calibration load residuals of the six load components that are obtained after applying the data reduction matrix derived from the optimized models of Fig. 4a to the gage outputs. Figure 5a shows the optimized models of the gage outputs that were obtained after applying the simplified search algorithm. It took approximately 3 minutes of CPU time to complete the search. Figure 5b shows the calibration load residuals of the six load components that are obtained after applying the data reduction matrix derived from the optimized models of Fig. 5a to the gage outputs.

It is interesting at this point to directly compare the standard deviation of the load residuals that can be obtained for the five different regression model sets discussed above. Table 3 below lists the standard deviation as a percentage of the capacity for the six load components.

Table 3: Standard deviation of the load residuals in percent of the load capacity.

MATH MODEL	$N1$	$N2$	$S1$	$S2$	RM	AF
Fig. 2	0.1008 %	0.0967 %	0.2115 %	0.1762 %	0.1780 %	0.1354 %
Fig. 3b	0.1101 %	0.1032 %	0.2402 %	0.1822 %	0.2011 %	0.1709 %
Fig. 3d	0.1207 %	0.1063 %	0.2456 %	0.1931 %	0.2016 %	0.1760 %
Fig. 4a	0.1216 %	0.1065 %	0.2473 %	0.1986 %	0.2055 %	0.1808 %
Fig. 5a	0.1229 %	0.1075 %	0.2481 %	0.1942 %	0.2065 %	0.1782 %

Overall, the standard deviation of the load residuals compare very well for the five regression model sets. The comparison of the standard deviations also illustrates that differences between the predictive capability of the regression models of the original seach algorithm, i.e., Fig. 4a, and the regression models of the simplfied seach algorithm, i.e., Fig. 5a, are very small (less than 0.005 % of load capacity).

IV. Conclusions

A new simplified regression model search algorithm was described. The new algorithm was derived from a more complex search algorithm that was originally developed at the NASA Ames Balance Calibration Laboratory. The new simplified search algorithm tries to rapidly identify a regression model for a given experimental data set that meets strict statistical quality requirements and simultaneously prevents "over–fitting" of the data. The simplified search algorithm has the advantage that it requires only about one tenth of the original algorithm's CPU time for the completion of a regression model search. However, the simplified search algorithm cannot guarantee that the final optimized regression model will always meet the chosen statistical quality requirements. Therefore, the simplified search algorithm is intended to be used for the generation of an alternate optimized regression model of the given data whenever the application of the original search algorithm either fails or requires too much CPU time. In addition, the simplified algorithm may be used to get better initial estimates of tare corrected loads of wind tunnel strain–gage balance calibration data. These improved estimates are needed as input whenever the original search algorithm is applied to strain–gage balance calibration data that is not yet corrected for the weight of the calibration body and other calibration hardware components.

Machine calibration data of NASA's MK40 six–component force balance is used to illustrate the application of the simplified search algorithm to a strain–gage balance data set. First, it is explained how a suitable upper bound for the regression model search can be defined for the chosen balance calibration data set. New information about the bi–directionality characteristics of the individual gages of a TASK balance is used for this purpose. Therefore, all absolute value terms of the rolling moment and axial force and the quadratic terms of the normal and side force components are omitted in the regression models of the gage outputs. Then, both the original and the simplified algorithm are applied to the calibration data set using the upper bound to limit the regression model search. This analysis demonstrated that the prediction accuracy, i.e., the standard deviation of the load residuals, shows excellent agreement whenever results for the optimized models of the original search algorithm are compared with results for the optimized models of the simplified search algorithm. The simplified algorithm, however, needed only about one tenth of the original algorithm's CPU time for the regression model search.

References

[1]Ulbrich, N. and Volden, T., "Strain–Gage Balance Calibration Analysis Using Automatically Selected Math Models," AIAA 2005–4084, paper presented at the 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, Arizona, July 2005.

[2]Ulbrich, N. and Volden, T., "Application of a New Calibration Analysis Process to the MK–III–C Balance," AIAA 2006–0517, paper presented at the 44th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2006.

[3]Ulbrich, N. and Volden, T., "Regression Analysis of Experimental Data Using an Improved Math Model Search Algorithm," AIAA 2008–0833, paper presented at the 46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2008.

[4]Ulbrich, N., "Regression Model Optimization for the Analysis of Experimental Data," AIAA 2009–1344, paper presented at the 47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 2009.

[5]Ulbrich, N. and Volden, T., "Regression Model Term Selection for the Analysis of Strain–Gage Balance Calibration Data," AIAA 2010–4545, paper presented at the 27th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, Chicago, Illinois, June/July 2010.

[6]Ulbrich, N. and Volden, T., "Development of a User Interface for a Regression Analysis Software Tool," AIAA 2010–0932, paper presented at the 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 2010.

[7]AIAA/GTTC Internal Balance Technology Working Group, "Recommended Practice, Calibration and Use of Internal Strain–Gage Balances with Application to Wind Tunnel Testing," AIAA R–091–2003, American Institute of Aeronautics and Astronautics, Reston, Virginia, 2003.

[8]Ulbrich, N., "Detection of Bi–Directionality in Strain–Gage Balance Calibration Data," AIAA 2012–3320, paper presented at the 28th AIAA Aerodynamic Measurement Technology, Ground Testing, and Flight Testing Conference, New Orleans, Louisiana, June 2012.

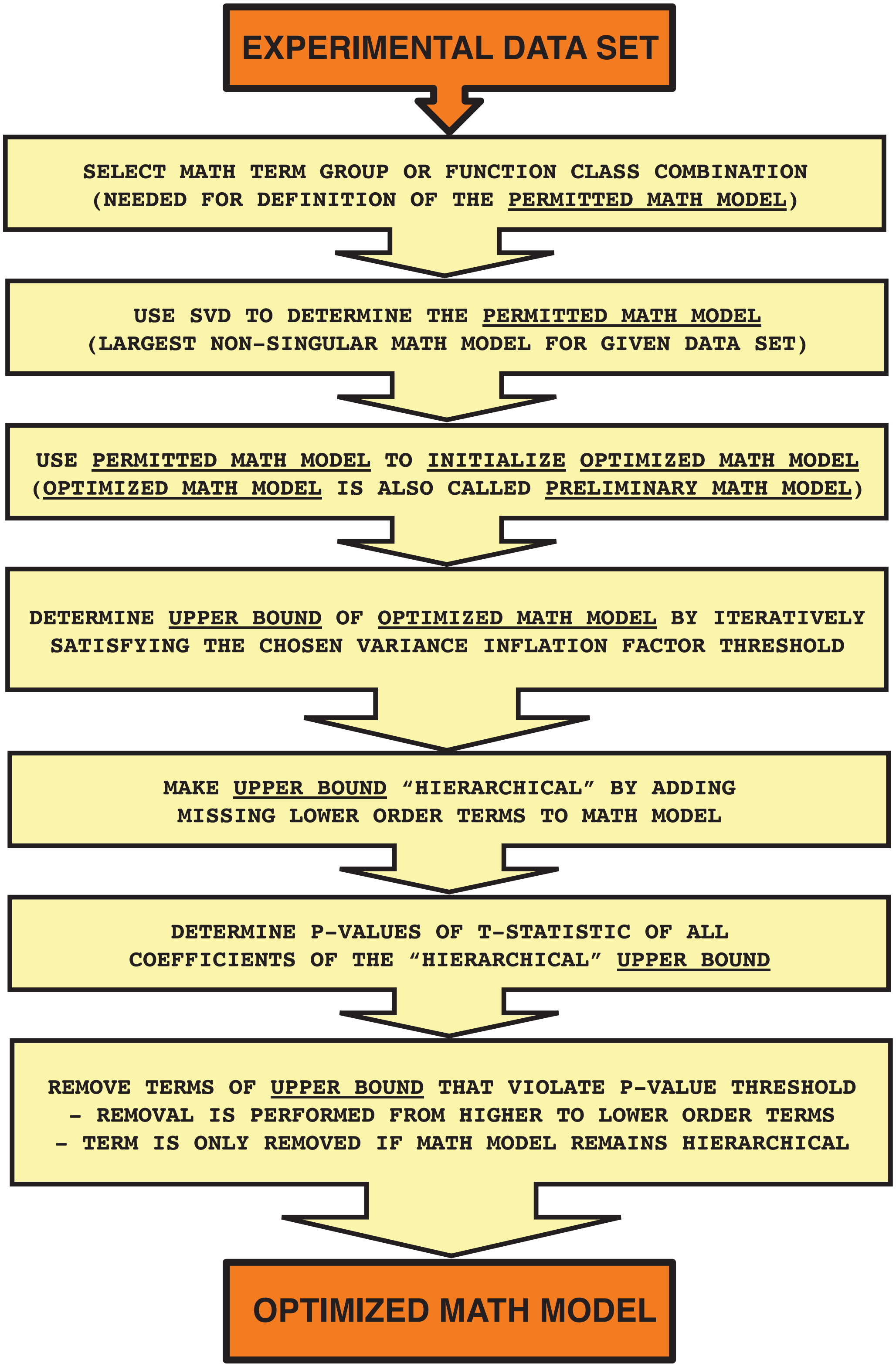

Fig. 1 Basic elements of the simplified regression model search algorithm.

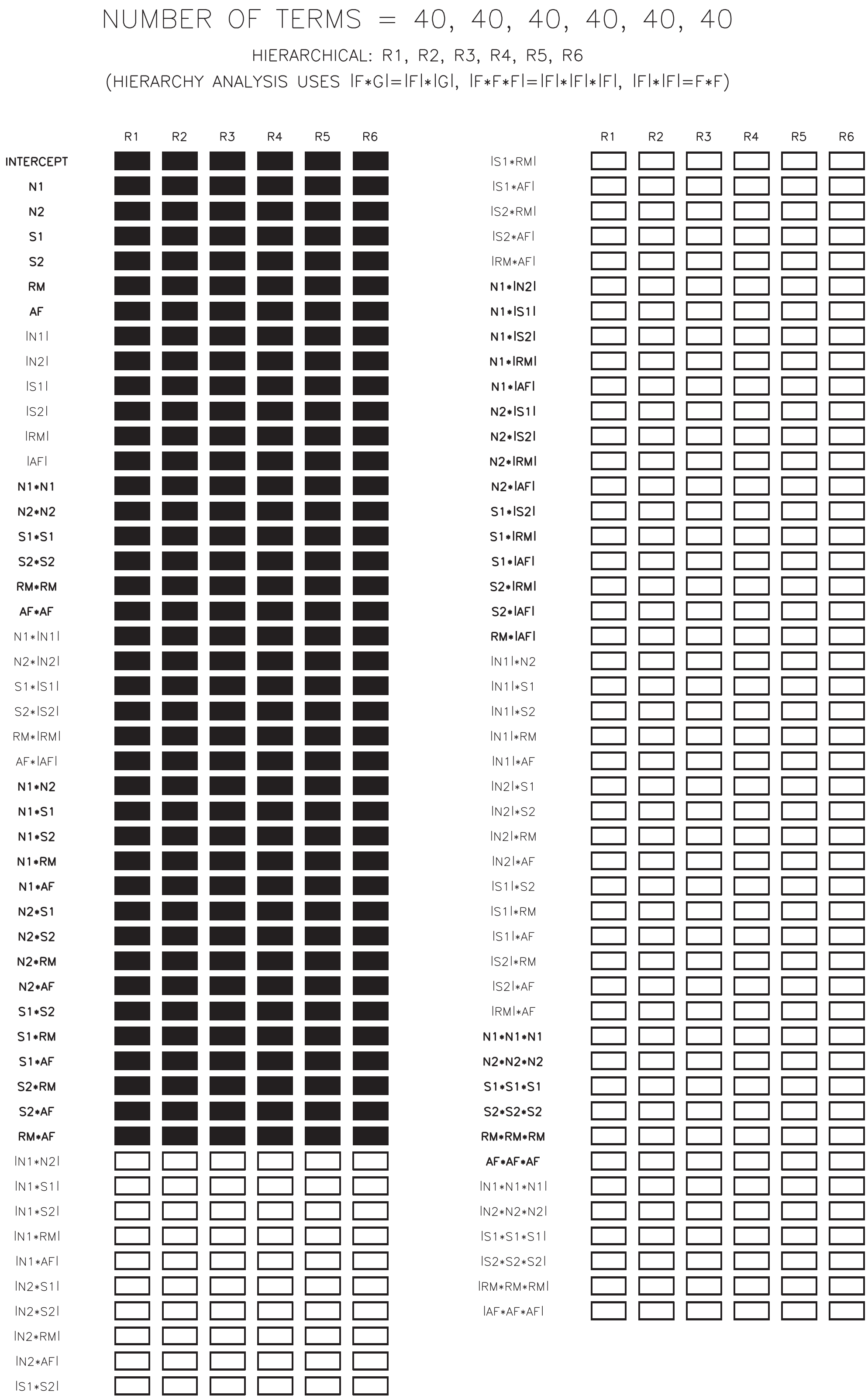

Fig. 2 Initial permitted math model of the MK40 calibration data set.

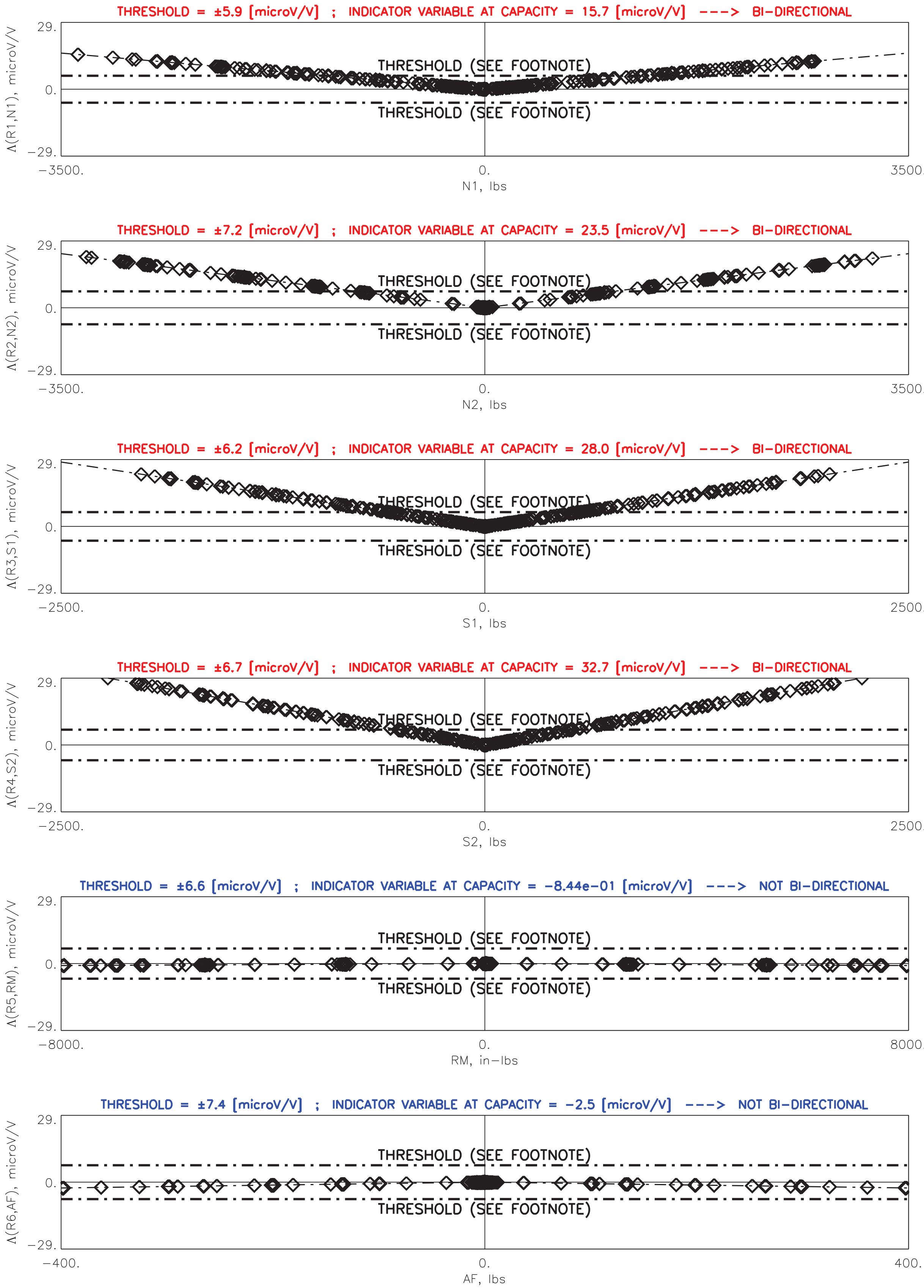

Fig. 3a Indicator variable for bi–directionality of the MK40 balance.

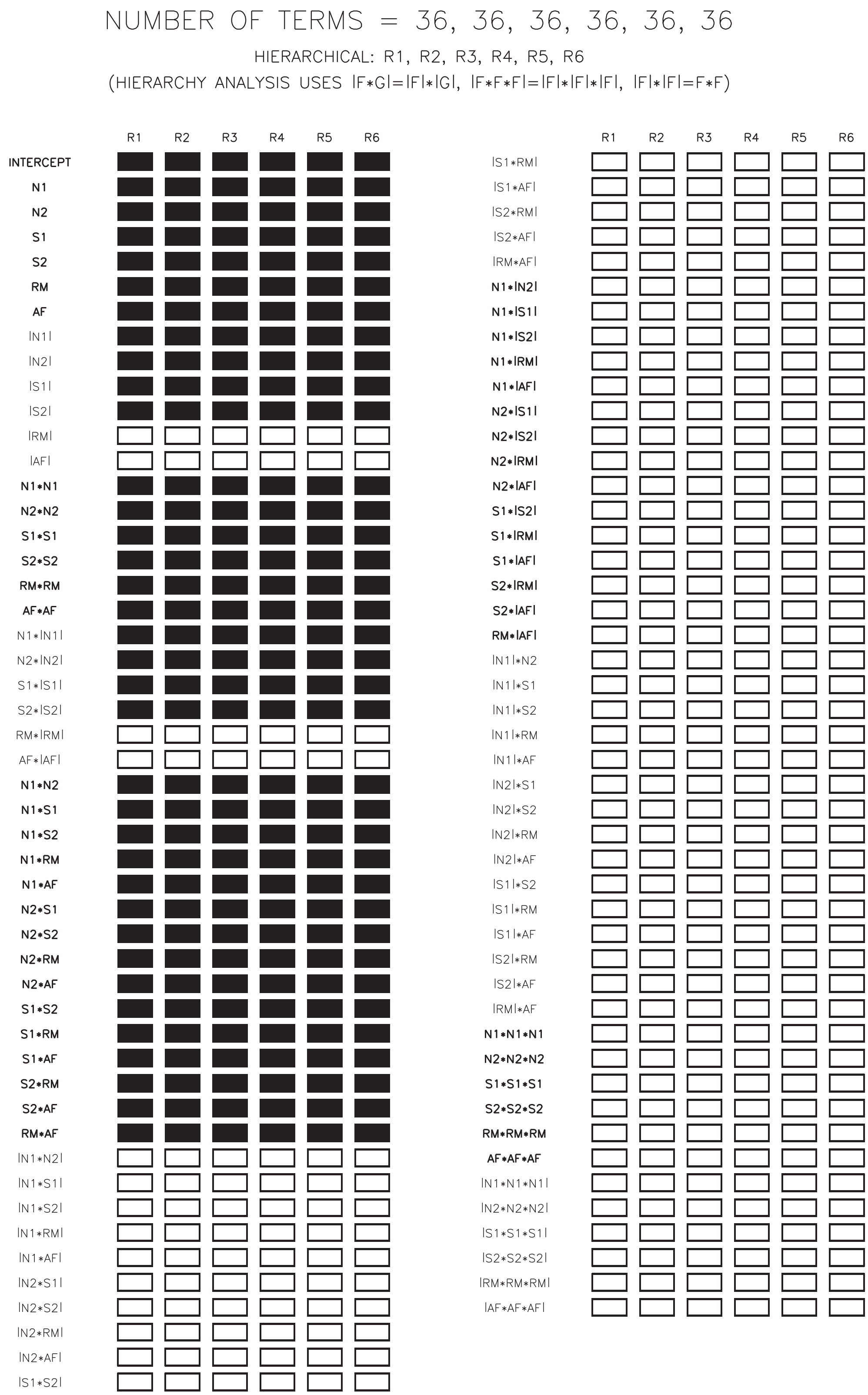

Fig. 3b First revision of the permitted math model after removal of terms related to $|AF|$ and $|RM|$.

PHYSICAL VARIABLES & UNITS – REGRESSION COEFFICIENT ESTIMATES AND STATISTICAL METRICS (R1)

REGRESSION MODEL HIERARCHY CHARACTERISTICS = HIERARCHICAL

TERM INDEX	TERM NAME	COEFFICIENT VALUE	STANDARD ERROR	T–STATISTIC OF COEFFICIENT	P–VALUE OF COEFFICIENT	VIF (PRIMARY)	VIF (ALTERNATE)
1	INTERCEPT	−1.3915	+0.0724	−19.2079	–	–	–
2	N1	+0.3435	+0.0001	+3218.6915	< 0.0001	+3.1535	[+11.1063]
3	N2	−0.0081	+0.0001	−73.9053	< 0.0001	+2.7146	[+17.9385]
4	S1	+0.0002	+0.0001	+1.5301	+0.1262	+2.6630	[+12.7765]
5	S2	+0.0002	+0.0001	+1.6099	+0.1076	+3.1455	[+16.5158]
6	RM	+0.0005	+1.2244e−05	+40.4820	< 0.0001	+1.2300	+1.0116
7	AF	+0.0009	+0.0003	+3.5283	+0.0004	+1.0939	+1.0052
8	|N1|	−0.0004	+0.0002	−2.2613	+0.0239	+16.6365	[+16.1960]
9	|N2|	+0.0021	+0.0001	+15.7623	< 0.0001	+18.6616	[+18.6444]
10	|S1|	−0.0016	+0.0003	−6.0404	< 0.0001	+24.2112	[+24.2108]
11	|S2|	−0.0003	+0.0002	−1.4333	+0.1519	+24.0623	[+24.0623]
14	N1∗N1	+1.2580e−06	+7.0086e−08	+17.9486	< 0.0001	+17.4855	[+16.6215]
15	N2∗N2	−1.3589e−08	+5.0617e−08	−0.2685	+0.7884	+17.4322	[+17.4844]
16	S1∗S1	−9.7985e−08	+1.3672e−07	−0.7167	+0.4737	+18.8256	[+18.8276]
17	S2∗S2	+5.1810e−08	+1.1008e−07	+0.4707	+0.6379	+17.8914	[+17.8914]
18	RM∗RM	−1.3117e−07	+2.2172e−09	−59.1601	< 0.0001	+1.1603	+1.1606
19	AF∗AF	+2.6983e−06	+8.1051e−07	+3.3291	+0.0009	+1.0432	+1.0432
20	N1∗|N1|	+1.2456e−06	+5.1676e−08	+24.1036	< 0.0001	+1.3995	[+11.4671]
21	N2∗|N2|	+1.6883e−07	+4.5120e−08	+3.7419	+0.0002	+2.6652	[+18.2346]
22	S1∗|S1|	−3.3343e−07	+9.3268e−08	−3.5750	+0.0004	+1.9403	[+11.9933]
23	S2∗|S2|	−2.4281e−07	+8.6584e−08	−2.8043	+0.0051	+2.3194	[+15.1907]
26	N1∗N2	−1.4454e−07	+2.2080e−08	−6.5461	< 0.0001	+1.5788	+1.5701
27	N1∗S1	−1.1806e−06	+6.1998e−08	−19.0419	< 0.0001	+1.7335	+1.5130
28	N1∗S2	+2.2981e−08	+5.5697e−08	+0.4126	+0.6799	+1.7532	+1.5240
29	N1∗RM	−3.0356e−07	+1.9196e−08	−15.8139	< 0.0001	+1.2554	+1.0358
30	N1∗AF	−1.6866e−07	+2.8800e−07	−0.5856	+0.5582	+1.1131	+1.0233
31	N2∗S1	+1.1083e−07	+4.9463e−08	+2.2408	+0.0252	+1.5110	+1.5093
32	N2∗S2	−1.0591e−07	+4.4563e−08	−2.3766	+0.0176	+1.5228	+1.5204
33	N2∗RM	+1.3387e−08	+1.4845e−08	+0.9018	+0.3673	+1.0344	+1.0323
34	N2∗AF	+9.7052e−08	+2.4748e−07	+0.3922	+0.6950	+1.0199	+1.0227
35	S1∗S2	+3.2285e−08	+5.0522e−08	+0.6390	+0.5229	+1.8467	+1.8467
36	S1∗RM	−8.8355e−07	+2.6275e−08	−33.6266	< 0.0001	+1.4044	+1.4048
37	S1∗AF	−3.8047e−07	+4.1734e−07	−0.9117	+0.3621	+1.3182	+1.3183
38	S2∗RM	−8.6731e−08	+2.5668e−08	−3.3790	+0.0007	+1.4104	+1.4108
39	S2∗AF	+4.6004e−09	+3.8899e−07	+0.0118	+0.9906	+1.3182	+1.3183
40	RM∗AF	−3.4890e−08	+1.2391e−07	−0.2816	+0.7783	+1.0000	+1.0001

Fig. 3c Variance inflation factors of the math model of gage output $R1$ for the revised permitted math model.

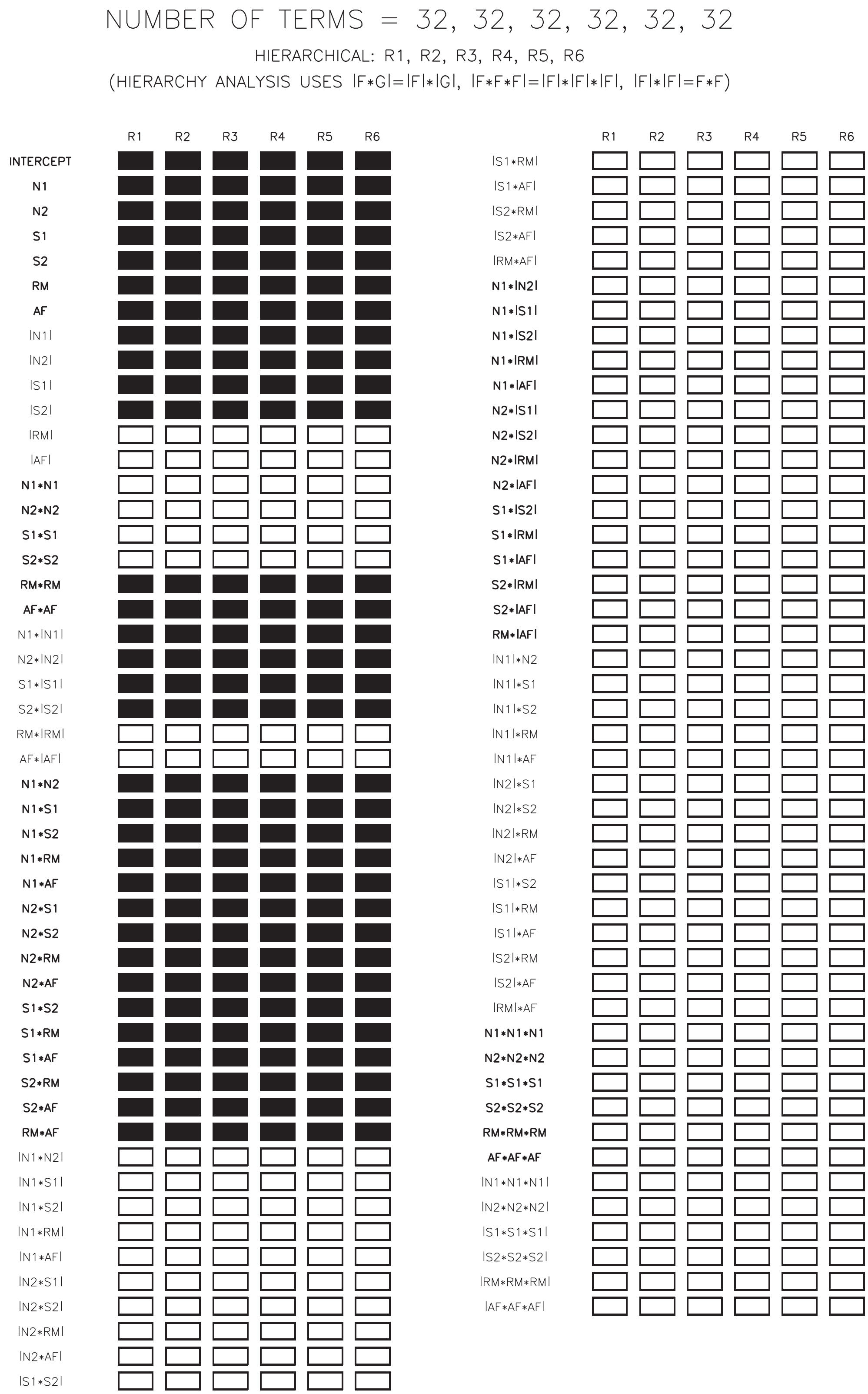

Fig. 3d Second revision of the permitted math model after removal of square terms of $N1$, $N2$, $S1$, and $S2$. (second revision is used as the upper bound for both the original and simplified search algorithm)

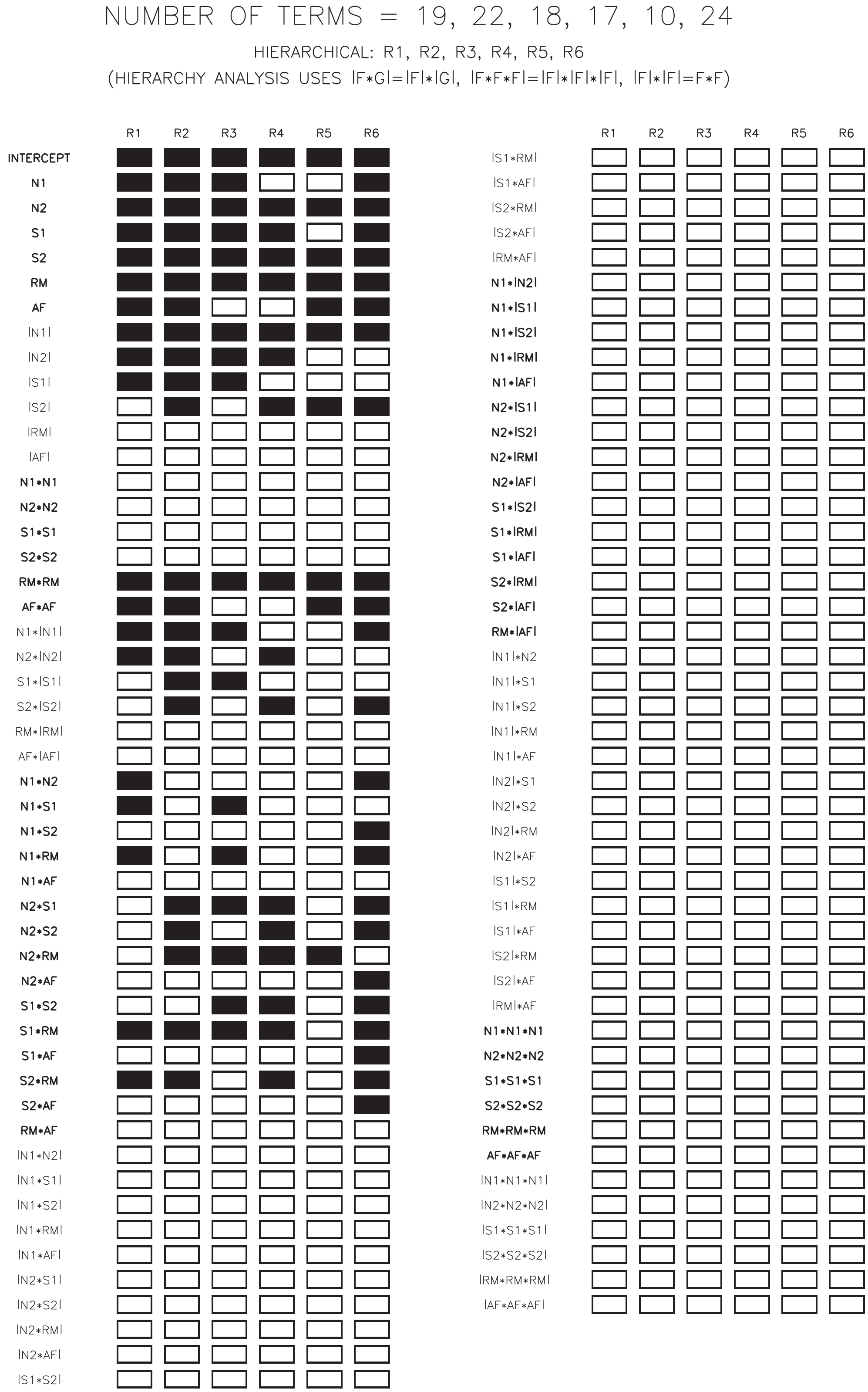

Fig. 4a Optimized math model obtained after application of <u>original</u> search algorithm. (regression model search was completed after ≈ 28 minutes of CPU time)

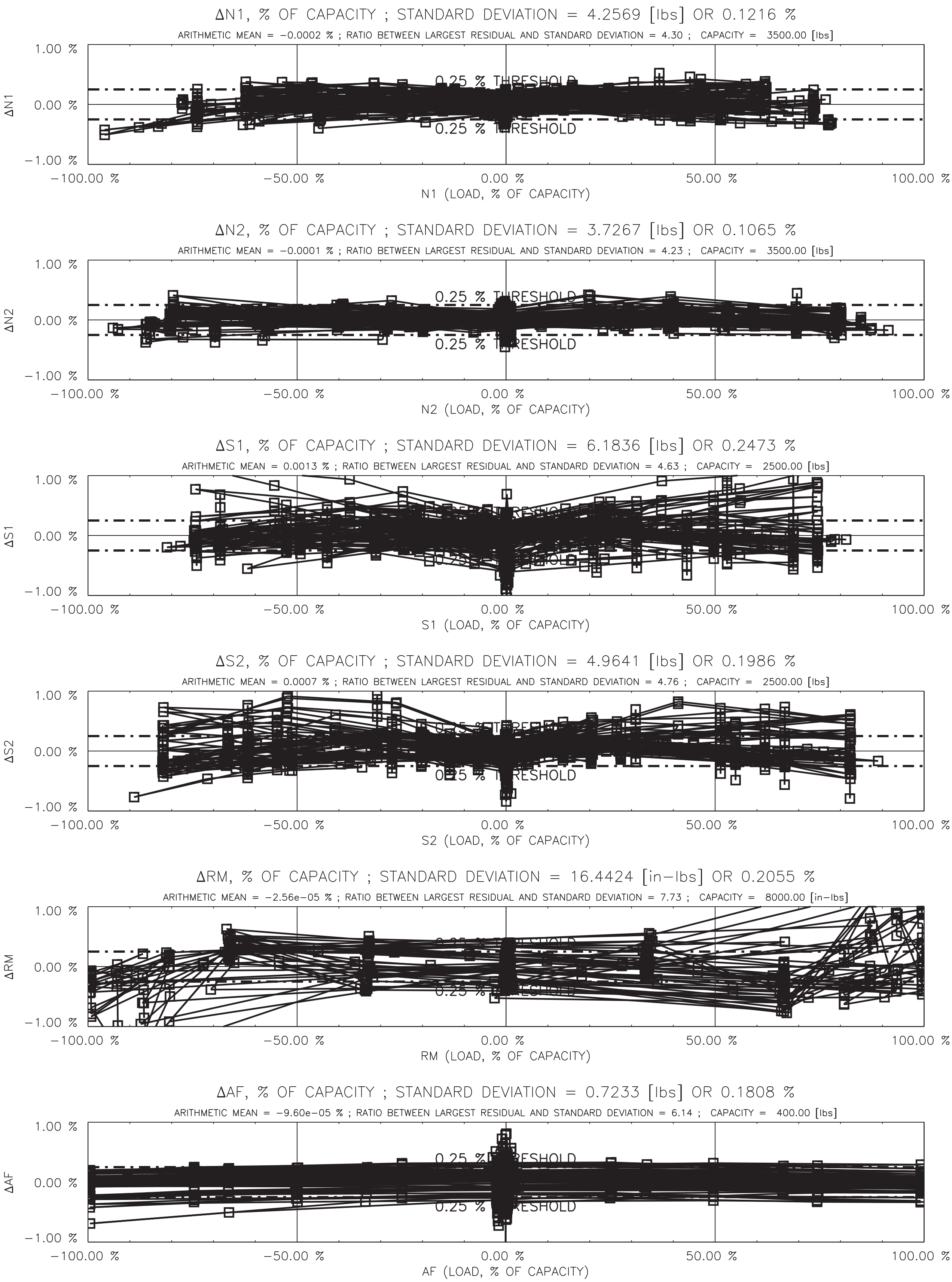

Fig. 4b Load residuals for optimized math model of original search algorithm.

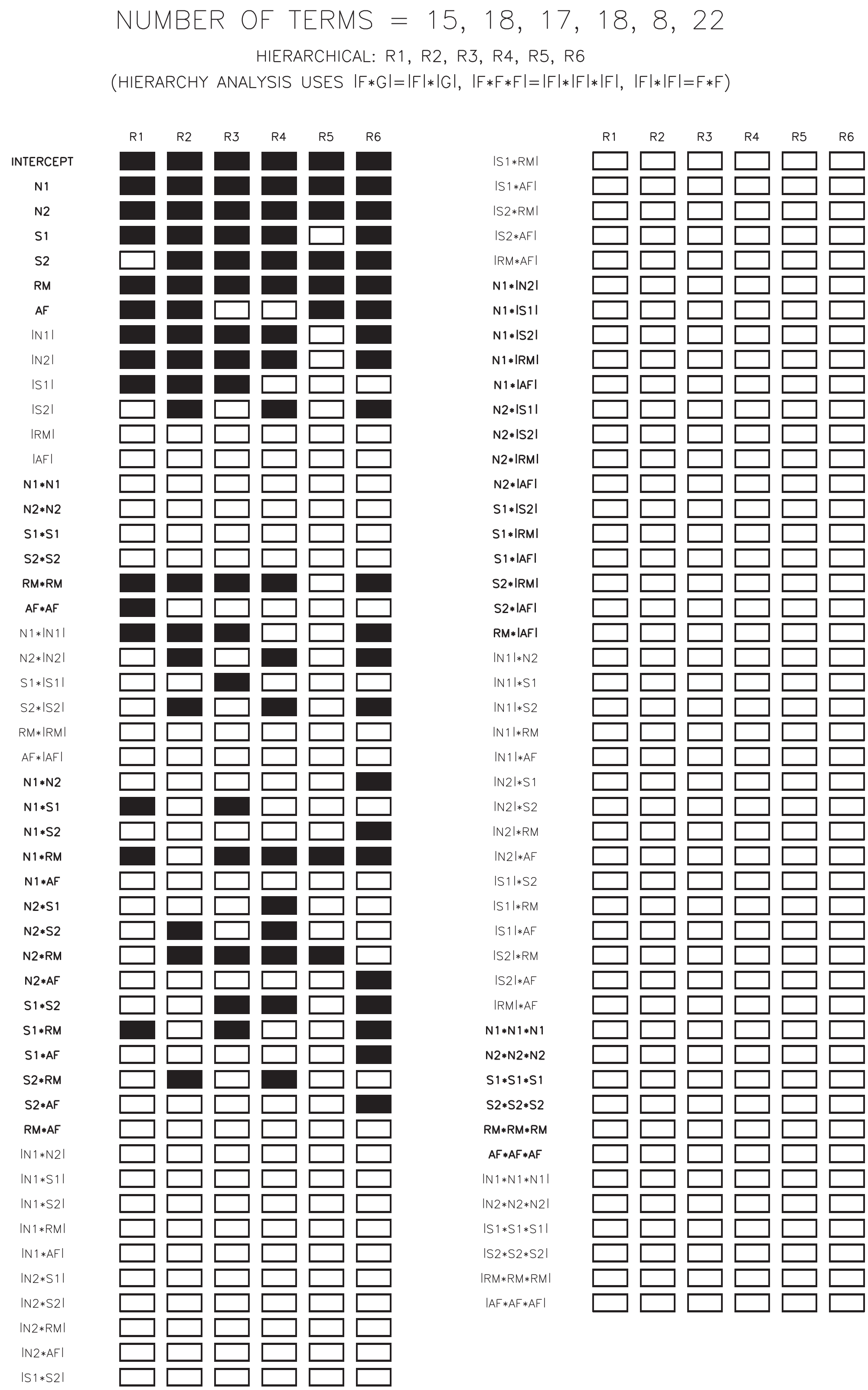

Fig. 5a Optimized math model obtained after application of <u>simplified</u> search algorithm.
(regression model search was completed after ≈ 3 minutes of CPU time)

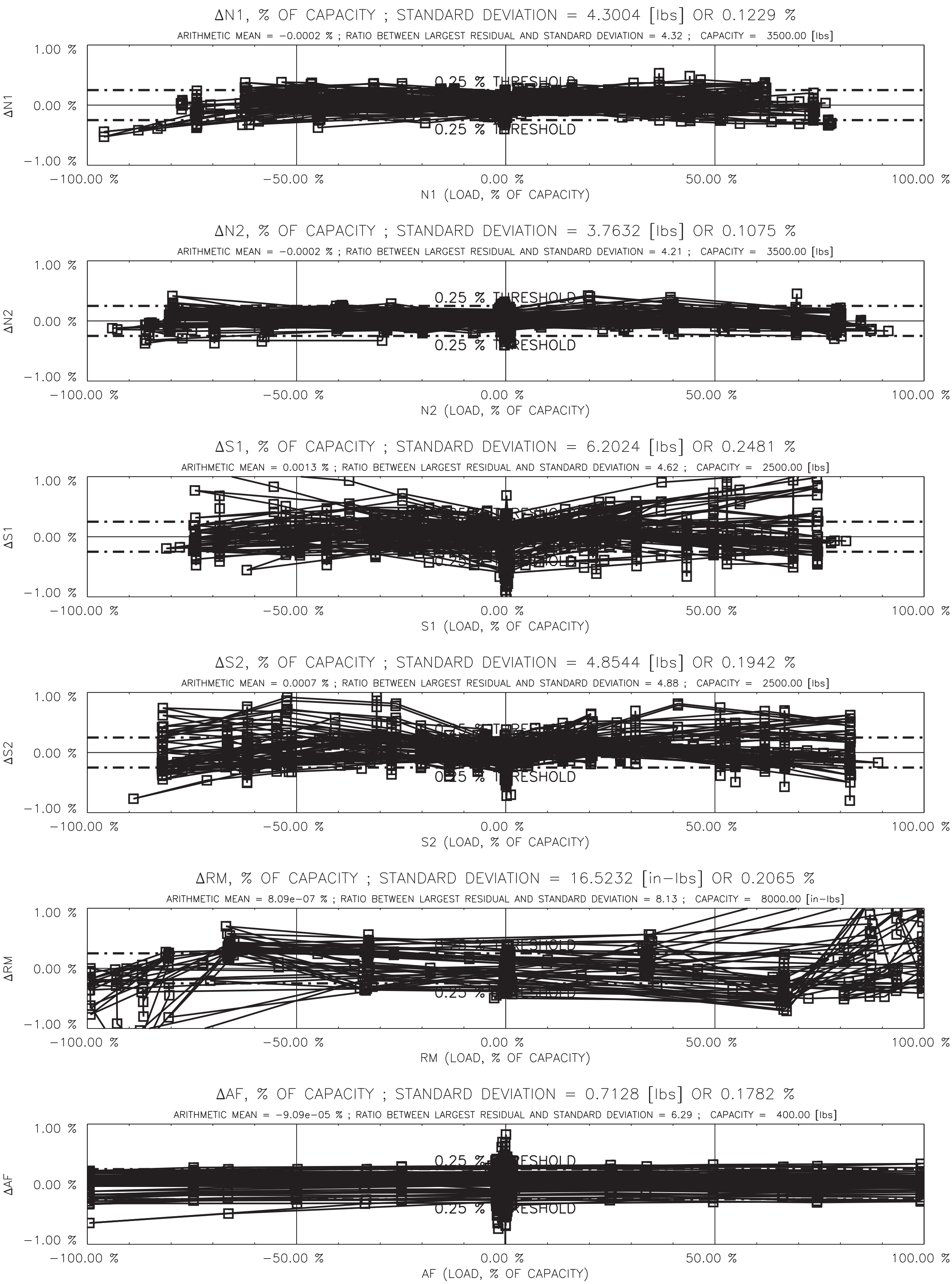

Fig. 5b Load residuals for optimized math model of simplified search algorithm.