RATIO EDIT TOLERANCE DEVELOPMENT USING VARIATIONS OF EXPLORATORY DATA ANALYSIS (EDA) RESISTANT FENCES METHODS

Katherine Jenny Thompson
Economic Statistical Methods and Programming Division, United States Bureau of the Census

KEY WORDS: EDA, quartile, resistant

1. Introduction

Many data items collected by the Bureau of the Census Economic Programs are subjected to ratio edits. In a ratio edit, the ratio of two correlated items is compared to upper and lower bounds, known as tolerances. Reported items that fall outside of the tolerances are considered edit failures, and one or both of the items in an edit-failing ratio are either imputed or flagged for analyst review. The efficiency of the ratio edit is consequently dependent on the selected tolerances.

In 1996, Thompson and Sigman conducted research to determine a statistical method of automatically setting tolerance limits that works well for different sets of economic data for use in the 1997 economic census. We evaluated these methods on two sets of historical data: the 1994 Annual Survey of Manufactures (ASM) and the 1992 Business Census. In both data sets, we achieved success with some variations of an Exploratory Data Analysis (EDA) method called resistant fences. The resistant fences rules flag a ratio as an outlier when it is *k* interquartile ranges outside of the first or third quartiles (*k* is a constant).

In the Business Census applications, the resistant fences methods worked best when the original distributions of ratios were symmetrized using a power transformation before applying the resistant fences methods (final tolerances were obtained using the inverse-transform on the initial limits). However, in other data sets, the symmetrizing effort has not proved worthwhile.

Lanska and Kryscio (1997) propose a variation of the resistant fences rules for asymmetric distributions: use the distance between the first quartile and the median and the distance between the third quartile and the median for the upper and lower fences instead of the interquartile range, thus elongating the fences in the direction of the skewness of the distribution. This paper investigates this method for tolerance development, in comparison with the previously described resistant methods on several sets of simulated data. In Section 2, I describe the resistant methods investigated for tolerance development. Section 3 describes the simulated data. Section 4 presents an evaluation of these methods. Section 5 presents my recommendation.

2. Resistant Methods Used for Tolerance Development

Given an ordered distribution of ratios, let q_{25} = the first quartile, q_{75} = the third quartile, m= the sample median, and H = (q_{75} - q_{25}), the interquartile range. Then,

Resistant Fences flag outliers as ratios less than $q_{25} - k\times H$ or greater than $q_{75} + k\times H$

Asymmetric Fences flag outliers as ratios less than $q_{25} - k\times(m- q_{25})$ or greater than $q_{75} + k\times(q_{75}- m)$

The value of *k* determines the fence's "rule." For resistant fences, *k*= 1.5 defines inner fences, *k* = 2 defines middle fences, and *k* = 3 defines outer fences. For asymmetric fences, *k* =3 defines inner fences, *k* = 4 defines middle fences, and *k* = 6 defines outer fences. Resistant fences methods have been used successfully at the Census Bureau to develop tolerances for several economic census sectors. The asymmetric fences methods are used in the Hidiriglou-Berthelot edit (Hidiriglou and Berthelot, 1986).

The resistant fences rules implicitly assume symmetry. When the distribution is symmetric, then the expected value of the resistant and asymmetric fences are equivalent under the same "rule." When distributions of ratios are highly skewed (as with economic data), it can be helpful to "symmetrize" the original distribution of ratios with the natural logarithm transformation or the square root transformation prior to applying the resistant fences rule [Note: The EDA method for determining an appropriate power transformation to symmetrize data described in Thompson and Sigman (1998) was not as effective. In this study, only 4% of the asymmetric distributions were symmetrized with this method; the rest used the natural log.] Resistant fences tolerances developed from the original data sets are labeled RNI (inner), RNM (middle), and RNO (outer). Resistant fences tolerances developed from symmetrized data are labeled RSI (inner), RSM (middle), and RSO (outer). Asymmetric fences tolerances are labeled AFI (inner), AFM (middle), and AFO (outer).

Note that each method uses the same **number** of interquartile ranges to set the fences. If the distribution of ratios is unskewed (symmetric), then these three methods are equivalent. If not, the difference between the RN and AN methods is the **location** of the fences, not the length from lower to upper fence [Note: this assumes that negative lower fences are permissible. If a negative lower fence is truncated at zero, then the lengths of the tolerances will differ]. The RS methods differ in both length and location.

3. Simulated Data

Thompson and Sigman (1998) uses historical data. There are two problems with this. First, it assumes that the edited data are entirely correct. Second, it cannot be used to determine a relationship between the effectiveness of the tolerance development methods in relation to sample size, to degree of skewness of the ratio distribution, and to correlation between ratio items. Using simulated data modeled on collected economic data allows investigation of these relationships.

The simulated data are modeled from two sources of edited historical data: six industries from the 1994 ASM; and ten industries from the 1992 Business Census. The ASM populations consisted of five selected data items: production workers (PW), other workers (OE), production workers wages (WW), other workers wages (OW), and plant hours worked by production workers (PH). The Business Census populations consisted of four selected data items: annual payroll (APR), first quarter payroll(QPR), total employment (EMP), and sales (SALES). The simulated population models have the same correlation structure for all ratio test pairs, and simulated data items match the original populations corresponding items on at least two moments and have approximately the same first and third quartiles (with the same size data sets).

From these 16 models, I randomly generated 12 multivariate populations per model (four each of sample sizes *n* = 20, 80, and 3000) of **good** data items (192 populations total). Half were "contaminated" by replacing the original values with *m* "bad" data items (*m* = [0.10**n*]) produced from each model (the 10% contamination group), using contamination models proposed by subject matter experts (see Luzi and Della Rocca (1998), Barnett and Lewis (1978), and Little (1987)). The remaining 96 populations do not contain any **bad** data items (0% contamination).

Although the errors were randomly induced in the 10% contamination populations, the probability each type of error was fixed for a given variable and was determined via data analysis of **bad** data items in the original unedited data sets and by subject matter analyst suggestions. When possible, I mimicked common reporting error patterns. Keying errors are generally independent by item within the same respondent questionnaire, as are statistical outliers. Reporting and keying errors are **not** necessarily outliers and may not be identifiable by a ratio test (e.g., swapped digits and wrong digits).

I examined three ratio tests in each population (576 separate distributions of ratios). The ASM model ratio edits (ratio distributions) are OW/OE, PH/PW, and WW/PW. The Business Census model ratio edits are APR/EMP, QPR/EMP, and SALES/EMP. A ratio is **bad** if either the numerator or the denominator is **bad.** The proportion of outlying **ratios** in a data set is not equal to the proportion of contaminated data items. Table 1 provides information on the proportion of outlying ratios by size for the 10% contamination populations. The uncontaminated populations examine the performance of each method in the presences of **no** outliers, an unrealistic situation in practice but useful in analysis, measuring the propensity for Type I errors. The 10% contamination populations are more realistic.

Table 1: Proportion of Outlying Ratios by Size in the 10% Contamination Populations

Size	Median	Mean	Minimum	Maximum
Small	0.20	0.18	0.10	0.20
Medium	0.19	0.18	0.10	0.20
Large	0.19	0.18	0.10	0.20

4. Evaluation Study

4.1 Criteria

The evaluation examines the following four statistics: Type I error rate, Type II error rate, hit rate, and outside rate. Ratio edits are hypothesis tests, and errors occur in each direction. The null hypothesis – that both data items in a ratio are good – is rejected when the ratio falls outside of the tolerances. The **Type I error rate** for a given distribution of ratios is ratio of the number of good ratios outside of the tolerances to the total number of good ratios in the distribution. A **Type II error** occurs when a bad ratio lies within the tolerances. When data items are subjected to more than one ratio test, the individual ratio test Type II error rate is a poor measure of the proportion of bad data remaining in the edited data set. Instead, the **all-ratio (overall) Type II error rate** is computed over the complete set of ratio edits applied to a population as the ratio of bad ratios that are not flagged by any ratio edit (bad ratios inside the tolerances) to the total number of bad ratios.

The **hit rate** (Granquist, 1995) is the proportion of flagged ratios that are bad (ratio of bad items outside of the tolerances to total items outside of the tolerances). Hit rates measure the operational effectiveness of an edit. The **outside rate** is the proportion of ratios outside of the tolerances. Since ratios outside of the tolerances are often flagged for analyst review, this statistic is used as a proxy for expected workload. By itself, this statistic does not yield any information about the effectiveness of the ratio edit in identifying erroneous data. In this evaluation, the outside rate is a useful evaluation tool for the 0% contamination populations.

4.2 Evaluation Methodology

I produced nine sets of edit tolerances per ratio test in each population: three using resistant fences on the original distributions (RNI, RNM, and RNO); three using resistant fences on the transformed distributions (RSI, RSM, and RSO); and three using asymmetric fences on the original distributions (AFI, AFM, and AFO). All non-zero ratios (regardless of their **good/bad** classification) are used for tolerance development.

Hit rates, Type I error rates, and outside rates were calculated separately by ratio test for each method. These statistics were then averaged within size group (*n* = 20, 80, 3000) × contamination group (0%, 10%) for **all** ratio tests (hit rates cannot be calculated in the uncontaminated populations). Similarly, all-ratio-test Type II error rates were calculated in the 10% contamination populations and averaged with size groups. These statistics are provided in the appendices which are discussed below.

Correlation between ratio items measures the ratio edit's prediction power: ratio edits can be viewed as no-intercept regression models, where the numerator is the dependent variable. Low correlation ratios have poor prediction power, and tolerances developed from these distributions are quite wide, often unusably so. Degree of skewness is equally important. If a distribution of ratios is highly positively skewed with several observations in the longer tail, then the skewness of the distribution should be accounted for in the tolerance development or too many **good** ratios in the longer tail will be erroneously flagged as outliers.

Correlation and skewness classes were calculated by pooling the sample correlations (?) and sample skewness coefficients (s_k) from all 576 distributions of ratios, regardless of size or contamination class (none or 10%). The 33rd and 66th percentiles set the cut-off for correlation or skewness class. Table 2 presents numbers of distributions of ratios within each correlation and skewness class.

Table 2: Correlation and Skewness Classifications for Simulated Data Sets (All Ratios Combined)

Size	Group	Correlation			Skewness		
		Low (? #0.85)	Medium (0.85<? #0.94)	High (0.94 < ?)	Low (s_k #2.68)	Medium (2.68< s_k#6.76)	High (6.76 < s_k)
Small (*n*=20)	0%	22	29	45	80	16	0
	10%	28	28	40	39	57	0
Medium (*n* = 80)	0%	38	30	28	64	30	2
	10%	31	31	34	5	44	47
Large (*n*=3000)	0%	40	33	23	61	14	21
	10%	44	34	22	0	1	95

When setting tolerances, we try to maximize the number of **bad** ratios outside of the tolerances (minimize Type I error rate and maximize hit rates) and maximize the number of **good** ratios inside of the tolerances (minimize the Type II error rate). It is far easier to control Type I error than Type II error through the ratio edit tolerances. Ratio edits can only catch **outlier** errors; inliers (items whose reported value is incorrect but consistent with the rest of the distribution) will not be flagged in a ratio edit.

4.3 Results

4.3.1 Characteristics of the Different Fences

None of the resistant methods guarantees a positive lower tolerance. And, in fact a non-zero lower bound is not necessarily a requirement for analysts. Negative lower bounds do have implications on all of the evaluation statistics, since low-outlier values are not detected. Table 3 presents proportion of negative lower bounds for each tolerance development method.

Table 3: Proportion of Negative Lower Bounds from Each Tolerance Development Method

	RN			RS			AF		
Size	Inner	Middle	Upper	Inner	Middle	Upper	Inner	Middle	Upper
Small	0.85	0.90	0.95	0.19	0.21	0.23	0.77	0.83	0.89
Medium	0.85	0.92	0.96	0.08	0.09	0.10	0.76	0.81	0.89
Large	0.84	0.90	0.96	0.06	0.06	0.06	0.75	0.80	0.88

In most of the data sets, the RN and AF tolerances have very close values under the same "rule." The RS tolerances have very different lengths and locations. First, the lower RS tolerance is usually positive. Second, when the distributions of ratios are positively skewed, the RS upper tolerances are located much further out in the longer tail than the corresponding RN and AF upper tolerances. Table 4 provides average ratios of upper tolerances within the same method for each rule.

Table 4: Average Ratios of Upper Tolerances Within the Same "Rule"

Size		Pooled Ratios			Correlation Classification									Skewness Classification								
					Low			Medium			High			Low			Medium			High		
	Method	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS	RN/ RS	RN/ AF	AF/ RS
S	Inner	0.5	0.9	0.6	0.5	0.9	0.6	0.5	0.9	0.6	0.6	1.0	0.7	0.7	0.9	0.7	0.4	0.9	0.5	N/A	N/A	N/A
	Middle	0.5	0.9	0.5	0.5	0.9	0.5	0.5	0.9	0.5	0.6	1.0	0.6	0.6	0.9	0.7	0.4	0.9	0.4	N/A	N/A	N/A
	Outer	0.4	0.9	0.4	0.4	0.9	0.4	0.4	0.9	0.4	0.5	1.0	0.5	0.6	0.9	0.6	0.2	0.9	0.2	N/A	N/A	N/A
M	Inner	0.5	0.9	0.5	0.4	0.9	0.5	0.5	0.9	0.5	0.7	1.0	0.7	0.7	1.0	0.8	0.5	0.9	0.5	0.5	0.9	0.5
	Middle	0.4	0.9	0.5	0.3	0.9	0.4	0.4	0.9	0.4	0.6	1.0	0.6	0.7	1.0	0.7	0.4	0.9	0.4	0.4	0.9	0.4
	Outer	0.3	0.9	0.3	0.2	0.9	0.3	0.3	0.9	0.3	0.5	1.0	0.5	0.6	1.0	0.6	0.2	0.9	0.3	0.3	0.9	0.3
L	Inner	0.5	0.9	0.5	0.4	0.9	0.4	0.5	0.9	0.6	0.8	1.0	0.8	0.8	1.0	0.8	0.3	0.8	0.4	0.5	0.9	0.5
	Middle	0.4	0.9	0.4	0.3	0.9	0.3	0.4	0.9	0.4	0.7	1.0	0.7	0.7	1.0	0.7	0.2	0.8	0.3	0.4	0.9	0.4
	Outer	0.3	0.9	0.3	0.2	0.9	0.2	0.3	0.9	0.3	0.6	1.0	0.6	0.6	1.0	0.6	0.1	0.8	0.2	0.2	0.9	0.3

Table 5 provides average ratios of tolerance length within the same "rule." On average, the RS tolerances are twice as wide (upper tolerance - lower tolerance) as the RN and AF tolerances under the same "rule," making this the much more conservative tolerance development method. Since the RN and AF lower fences are truncated at zero (negative lower bounds are not reasonable for these ratio tests), the AF tolerances are **slightly** wider than the RN tolerances. Because the AF upper tolerances are usually slightly larger than the RN tolerances, the RN Type I error rates and hit rates are higher.

Table 5: Average Ratios of Tolerance Length

Size		RN/RS	RN/AF	AF/RS
Small	Inner	0.6	0.9	0.6
	Middle	0.5	0.9	0.5
	Upper	0.4	0.9	0.4
Medium	Inner	0.5	0.9	0.5
	Middle	0.4	0.9	0.4
	Upper	0.3	0.9	0.3
Large	Inner	0.5	0.9	0.5
	Middle	0.4	0.9	0.4
	Upper	0.3	0.9	0.3

In summary, the RN and AF tolerances are more likely to be negative, so low-outliers are rarely flagged. This effects the Type I error rates, which are lower than the corresponding RS Type I error rates because no **good** small ratios are erroneously flagged. It also effects the Type II error rates, since **bad** small ratios remain unflagged. Hit rates will generally be higher with the RS tolerances than with the RN or AF tolerances because only extreme outliers at either end are flagged. And, in the uncontaminated populations, the RS outside rates are more likely to be the nominal 0% because the tolerances are so conservative.

4.3.2 Effect of Skewness of Distribution of Ratios

The effect of the skewness of the distribution of ratios on each tolerance development method depends on the size of the data set. With the small distributions of ratios, better results are obtained using the AFM limits, thus accounting for skewness in the distribution without transforming the data. When the data sets are small, symmetrizing the distributions prior to applying the resistant fences rules has no benefit. In fact, it can result in unusually wide tolerances. If the data set contains at least one outlier, the RS upper tolerance is usually much larger than any observation in the data set (large outliers are very influential in small data sets), regardless of the "rule" used.

As the sample size increases, the advantage of RS over AF rules depends on the degree of skewness. In both the medium and large data sets, the RSM methods appeared to best balance low Type I and Type II error rates and high hit rates when the degree of skewness is **high** (they are also more likely to yield nominal outside-rates when no outliers are present). Neither the RN or AF methods flag low-outliers in highly skewed populations, increasing the Type II error rates. Within "rule," most of the same high-outliers are flagged by the three sets of tolerances (the exception is the outer fences rule for RS tolerances, which is overly conservative). This is also true for the moderately skewed distributions of ratios in the **large** data sets, so again the RSM method appears to be the best choice. A poor second is the AFM tolerances: the hit rates are not nearly as high and the Type II error rates are not nearly as low.

It is harder to find the "best" method for the moderately skewed medium-sized distributions of ratios. The AF tolerances yield slightly lower Type I error rates, and the RS tolerances yield slightly higher hit rates in the medium sized distributions. However, in these populations the values of the RSI and AFM tolerances are very close. In this case, the RS symmetrizing effort is unjustified as it is when the skewness of the distribution ratios is low; the AFM tolerances are the safest choice.

4.3.3 Effect of Correlation Between Ratio Items on Tolerance Development

With the small data sets, I was unable to find any relationship between the correlation of ratio items and success of a given tolerance development method. In the medium and large data sets, as the correlation between ratio items increases, the Type I error rate and hit rates (within each tolerance development method) increase as well. This is reasonable, since the tolerances are narrower when the ratio items are highly correlated (the interquartile range is small in this case). More ratios, and consequently more **bad** ratios, are flagged as the correlation increases. Consequently, when the correlation between ratio items is high, substantial improvements in Type II error rates and hit rates are achieved with smaller values of *k* (use *k* for inner or middle fences rules).

4.4 Discussion

Effective tolerance development begins with data analysis. From available historical data, calculate samples sizes and skewness coefficients for each distribution of ratios, as well as sample correlation between pairs of ratio items in each distribution. Additional subject matter expertise is invaluable: these experts know which items are traditionally misreported and can describe whether a particular distribution of ratios is typical or appears to be an anomaly. Also, verify unimodality before proceeding: none of these methods work with bimodal data. Finally, verify that the distributions of ratios have a nonzero length interquartile range. Assuming that all of these requirements are satisfied and the calculated statistics appear fairly reasonable, follow these guidelines to get an initial set of tolerances:

5. If the sample size is small, use AFM rules to set tolerance limits.
6. If the sample size is reasonable (probably greater than 50) then use the degree of skewness of the distribution and the sample size to determine whether AF or RS methods are preferable (large and highly skewed data sets should probably be symmetrized). Before selecting RS methods, examine the effect of the potential symmetrizing power transformations on the data: if power transformations do not usually reduce the skewness, then AF methods are better. Once the method has been selected (RS or AF), use the degree of correlation between ratio items to select an initial

"rule" (value of *k*). With highly correlated data, use a small value of *k* (use inner or outer fences). Otherwise, use a large value of *k* (outer fences) to minimize Type I error.

If historical data are not available, the safest approach to use RSM methods on large distributions of ratios (more than 1000 observations) and use AFM methods otherwise. When using RSM methods, always include a comparison between the skewness coefficient of the untransformed data to the power transformed distribution(s).

5. Conclusion

This paper presents a variety of resistant methods for setting ratio edit tolerances. All of these methods are based on sample quartiles and have a breakdown point of 25% (i.e., up to one fourth of the ratios can be replaced with minimal expected change in results). Based on this study's results, I have developed guidelines that use characteristics of the distributions of ratios to determine which method should be used to develop initial tolerances. These guidelines are developed from a study that uses simulated data. A logical next step is to verify and refine these guidelines with actual (historical) data.

From an operational perspective, it is appealing to use one method/computer program to develop tolerances for all of the distributions of ratios in a data set. From a statistical perspective, this is not optimal. Each distribution of ratios is unique, and it is unreasonable to assume that the same model would work equally well on several sets of data. However, the approach that I recommend is quite flexible and can be easily implemented automatically with very few input parameters.

ACKNOWLEDGMENTS

I thank Nash Monsour, Amy Newman-Smith, and Richard Sigman for their useful comments, and Richard Sterner, Dennis Shoemaker, and William Winkler for their helpful suggestions on the error simulation process. This paper reports research and analysis undertaken by Census Bureau Staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties and to encourage discussion.

REFERENCES

Barnett,V. and Lewis, T. (1978). *Outliers in Statistical Data.* New York: John Wiley and Sons, Inc.

Granquist, L. (1995). Improving the Traditional Editing Process. *Business Survey Methods.* New York: John Wiley and Sons.

Hidiroglou, M.A. and Bertholot, J.M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. *Survey Methodology*, **12**, 73-83.

Lanska, Douglas J. and Kryscio, Richard J. (1997). Modified Box Plots for Asymmetric Distributions. Poster Session at the Joint Meetings of the American Statistical Association.

Little, R. and Smith, P. (1987). Editing and Imputation for Quantitative Survey Data. *Journal of the American Statistical Association,* **82**, 58-68.

Luzi, O. and Della Rocca, G. (1998). A Generalised Error Simulation System in an Internet/Intranet Environment. *Proceedings for the Conference of European Statisticians Statistical Standards and Studies, Section on Statistical Data Editing (Methods and Techniques),* United Nations Statistical Commission on Economic Commission for Europe.

Thompson, Katherine J. and Sigman, Richard S. (1998). Statistical Methods for Developing Ratio Edit Tolerances for Economic Data, to appear in *Journal of Official Statistics*.

Appendix One

Evaluation Statistics for Small Data Sets

		TOTAL		Correlation Classification						Skewness Classification			
Group		0%	10%	0%			10%			0%		10%	
				L	M	H	L	M	H	L	M	L	M
TYPE I	RNI	0.16	0.14	0.28	0.09	0.14	0.17	0.13	0.13	0.16	0.16	0.13	0.15
	RNM	0.09	0.08	0.09	0.10	0.08	0.10	0.08	0.08	0.09	0.08	0.08	0.09
	RNO	0.06	0.07	0.07	0.06	0.07	0.07	0.06	0.08	0.07	0.06	0.07	0.07
	RSI	0.09	0.08	0.06	0.09	0.10	0.10	0.07	0.08	0.10	0.08	0.07	0.08
	RSM	0.08	0.08	0.00	0.08	0.09	0.09	0.08	0.08	0.09	0.07	0.08	0.08
	RSO	0.07	0.07	0.00	0.08	0.07	0.06	0.08	0.06	0.08	0.07	0.09	0.06
	AFI	0.10	0.08	0.08	0.11	0.11	0.09	0.07	0.09	0.11	0.07	0.08	0.09
	AFM	0.10	0.08	0.08	0.09	0.10	0.07	0.07	0.10	0.11	0.06	0.07	0.09
	AFO	0.08	0.07	0.07	0.08	0.08	0.06	0.06	0.08	0.10	0.06	0.06	0.07
HITRATE	RNI	N/A	0.48	N/A	N/A	N/A	0.39	0.47	0.58	N/A	N/A	0.20	0.67
	RNM	N/A	0.50	N/A	N/A	N/A	0.42	0.46	0.60	N/A	N/A	0.20	0.70
	RNO	N/A	0.52	N/A	N/A	N/A	0.41	0.49	0.65	N/A	N/A	0.17	0.76
	RSI	N/A	0.49	N/A	N/A	N/A	0.43	0.41	0.61	N/A	N/A	0.18	0.69
	RSM	N/A	0.47	N/A	N/A	N/A	0.37	0.43	0.60	N/A	N/A	0.16	0.67
	RSO	N/A	0.39	N/A	N/A	N/A	0.29	0.28	0.59	N/A	N/A	0.08	0.60
	AFI	N/A	0.51	N/A	N/A	N/A	0.43	0.47	0.61	N/A	N/A	0.21	0.70
	AFM	N/A	0.50	N/A	N/A	N/A	0.42	0.45	0.61	N/A	N/A	0.19	0.71
	AFO	N/A	0.51	N/A	N/A	N/A	0.41	0.47	0.63	N/A	N/A	0.12	0.77
OUTSIDE	RNI	0.05	0.09	0.04	0.05	0.06	0.08	0.07	0.12	0.05	0.11	0.07	0.11
	RNM	0.03	0.08	0.02	0.03	0.03	0.07	0.06	0.10	0.02	0.09	0.05	0.10
	RNO	0.02	0.06	0.01	0.02	0.02	0.06	0.05	0.08	0.01	0.06	0.04	0.08
	RSI	0.03	0.09	0.01	0.03	0.04	0.08	0.06	0.12	0.03	0.05	0.04	0.12
	RSM	0.02	0.06	0.00	0.01	0.03	0.06	0.04	0.10	0.01	0.04	0.02	0.09
	RSO	0.01	0.04	0.00	0.00	0.01	0.03	0.02	0.07	0.01	0.02	0.01	0.06
	AFI	0.05	0.09	0.03	0.05	0.05	0.08	0.06	0.12	0.04	0.07	0.06	0.11
	AFM	0.04	0.08	0.02	0.03	0.04	0.06	0.06	0.11	0.03	0.06	0.04	0.10
	AFO	0.02	0.06	0.01	0.02	0.03	0.05	0.04	0.09	0.02	0.06	0.03	0.08
TYPE II	RNI	N/A	0.54										
	RNM	N/A	0.57										
	RNO	N/A	0.59										
	RSI	N/A	0.48										
	RSM	N/A	0.55										
	RSO	N/A	0.64										
	AFI	N/A	0.53										
	AFM	N/A	0.56										
	AFO	N/A	0.59										

Appendix Two

Evaluation Statistics for Medium Data Sets

	Group	TOTAL		Correlation Classification						Skewness Classification					
		0%	10%	0%			10%			0%			10%		
				L	M	H	L	M	H	L	M	H	L	M	H
TYPE I	RNI	0.08	0.06	0.06	0.07	0.11	0.05	0.05	0.08	0.06	0.10	0.15	0.08	0.06	0.05
	RNM	0.06	0.05	0.05	0.06	0.09	0.04	0.04	0.07	0.05	0.08	0.13	0.05	0.05	0.05
	RNO	0.05	0.04	0.04	0.05	0.08	0.04	0.03	0.05	0.04	0.06	0.10	0.04	0.04	0.04
	RSI	0.07	0.07	0.03	0.07	0.12	0.04	0.05	0.09	0.06	0.07	0.14	0.05	0.07	0.07
	RSM	0.07	0.06	0.03	0.07	0.10	0.03	0.05	0.08	0.08	0.06	0.11	0.03	0.06	0.06
	RSO	0.07	0.06	0.05	0.07	0.08	0.04	0.04	0.07	0.10	0.05	0.08	0.02	0.06	0.06
	AFI	0.07	0.06	0.05	0.06	0.13	0.05	0.06	0.09	0.06	0.10	0.14	0.09	0.06	0.06
	AFM	0.06	0.05	0.04	0.06	0.10	0.03	0.03	0.07	0.05	0.08	0.10	0.05	0.04	0.05
	AFO	0.06	0.05	0.04	0.05	0.10	0.04	0.03	0.07	0.05	0.06	0.09	0.05	0.05	0.05
HITRATE	RNI	N/A	0.44	N/A	N/A	N/A	0.40	0.43	0.48	N/A	N/A	N/A	0.12	0.43	0.48
	RNM	N/A	0.46	N/A	N/A	N/A	0.44	0.46	0.48	N/A	N/A	N/A	0.25	0.44	0.51
	RNO	N/A	0.47	N/A	N/A	N/A	0.43	0.47	0.49	N/A	N/A	N/A	0.06	0.46	0.51
	RSI	N/A	0.53	N/A	N/A	N/A	0.52	0.55	0.52	N/A	N/A	N/A	0.28	0.53	0.55
	RSM	N/A	0.50	N/A	N/A	N/A	0.49	0.49	0.52	N/A	N/A	N/A	0.20	0.51	0.53
	RSO	N/A	0.42	N/A	N/A	N/A	0.38	0.39	0.48	N/A	N/A	N/A	0.00	0.40	0.48
	AFI	N/A	0.45	N/A	N/A	N/A	0.40	0.44	0.50	N/A	N/A	N/A	0.11	0.43	0.50
	AFM	N/A	0.48	N/A	N/A	N/A	0.45	0.45	0.52	N/A	N/A	N/A	0.26	0.46	0.52
	AFO	N/A	0.47	N/A	N/A	N/A	0.43	0.46	0.52	N/A	N/A	N/A	0.02	0.47	0.52
OUTSIDE	RNI	0.07	0.11	0.06	0.06	0.09	0.09	0.09	0.15	0.05	0.11	0.15	0.11	0.11	0.11
	RNM	0.05	0.09	0.04	0.04	0.07	0.08	0.07	0.12	0.03	0.09	0.13	0.07	0.09	0.09
	RNO	0.04	0.07	0.03	0.03	0.05	0.06	0.05	0.09	0.02	0.07	0.10	0.04	0.07	0.07
	RSI	0.04	0.12	0.02	0.03	0.07	0.09	0.09	0.18	0.02	0.07	0.14	0.08	0.12	0.13
	RSM	0.02	0.10	0.01	0.01	0.05	0.06	0.07	0.15	0.01	0.05	0.11	0.03	0.09	0.11
	RSO	0.01	0.06	0.00	0.01	0.04	0.03	0.04	0.11	0.01	0.03	0.08	0.02	0.06	0.07
	AFI	0.06	0.11	0.05	0.05	0.10	0.08	0.08	0.16	0.04	0.11	0.14	0.11	0.11	0.11
	AFM	0.05	0.09	0.04	0.04	0.08	0.07	0.06	0.13	0.03	0.09	0.10	0.06	0.08	0.10
	AFO	0.03	0.08	0.03	0.02	0.06	0.06	0.05	0.11	0.01	0.07	0.09	0.04	0.07	0.08
TYPE II	RNI	N/A	0.58												
	RNM	N/A	0.62												
	RNO	N/A	0.67												
	RSI	N/A	0.41												
	RSM	N/A	0.50												
	RSO	N/A	0.63												
	AFI	N/A	0.56												
	AFM	N/A	0.60												
	AFO	N/A	0.65												

Evaluation Statistics for Large Data Sets

		TOTAL		Correlation Classification						Skewness Classification				
	Group	0%	10%	0%			10%			0%			10%	
				L	M	H	L	M	H	L	M	H	M	H
TYPE I	RNI	0.07	0.06	0.05	0.06	0.11	0.04	0.05	0.10	0.04	0.08	0.11	0.03	0.06
	RNM	0.05	0.04	0.03	0.05	0.10	0.03	0.04	0.09	0.03	0.07	0.09	0.01	0.04
	RNO	0.03	0.03	0.02	0.03	0.08	0.01	0.02	0.07	0.01	0.04	0.07	0.00	0.03
	RSI	0.04	0.04	0.01	0.05	0.09	0.01	0.04	0.10	0.02	0.03	0.09	0.00	0.04
	RSM	0.03	0.05	0.00	0.04	0.07	0.01	0.04	0.09	0.02	0.03	0.07	0.00	0.05
	RSO	0.04	0.04	0.01	0.05	0.05	0.02	0.04	0.07	0.04	0.03	0.04	0.00	0.04
	AFI	0.07	0.06	0.04	0.08	0.09	0.03	0.08	0.08	0.05	0.07	0.10	0.02	0.06
	AFM	0.05	0.05	0.02	0.07	0.07	0.02	0.07	0.07	0.04	0.05	0.08	0.01	0.05
	AFO	0.04	0.04	0.01	0.06	0.06	0.01	0.05	0.08	0.03	0.03	0.06	0.00	0.04
HIT RATE	RNI	N/A	0.44	N/A	N/A	N/A	0.41	0.43	0.49	N/A	N/A	N/A	0.49	0.44
	RNM	N/A	0.45	N/A	N/A	N/A	0.43	0.46	0.48	N/A	N/A	N/A	0.63	0.45
	RNO	N/A	0.46	N/A	N/A	N/A	0.45	0.48	0.48	N/A	N/A	N/A	0.74	0.46
	RSI	N/A	0.52	N/A	N/A	N/A	0.54	0.53	0.47	N/A	N/A	N/A	0.95	0.52
	RSM	N/A	0.50	N/A	N/A	N/A	0.50	0.51	0.50	N/A	N/A	N/A	0.96	0.50
	RSO	N/A	0.45	N/A	N/A	N/A	0.42	0.45	0.52	N/A	N/A	N/A	0.93	0.45
	AFI	N/A	0.45	N/A	N/A	N/A	0.43	0.44	0.51	N/A	N/A	N/A	0.54	0.45
	AFM	N/A	0.47	N/A	N/A	N/A	0.44	0.46	0.52	N/A	N/A	N/A	0.69	0.47
	AFO	N/A	0.47	N/A	N/A	N/A	0.45	0.47	0.53	N/A	N/A	N/A	0.76	0.47
OUTSIDE	RNI	0.07	0.11	0.05	0.07	0.10	0.08	0.10	0.18	0.04	0.10	0.14	0.05	0.11
	RNM	0.05	0.09	0.03	0.05	0.09	0.07	0.08	0.14	0.02	0.08	0.12	0.04	0.09
	RNO	0.03	0.07	0.02	0.03	0.06	0.06	0.06	0.12	0.01	0.05	0.09	0.03	0.07
	RSI	0.03	0.12	0.01	0.03	0.08	0.08	0.11	0.21	0.01	0.03	0.11	0.11	0.12
	RSM	0.02	0.10	0.00	0.02	0.07	0.06	0.09	0.18	0.01	0.01	0.08	0.09	0.10
	RSO	0.01	0.07	0.00	0.01	0.04	0.03	0.06	0.14	0.00	0.01	0.05	0.04	0.07
	AFI	0.06	0.11	0.04	0.08	0.09	0.07	0.12	0.18	0.04	0.08	0.13	0.05	0.12
	AFM	0.05	0.10	0.02	0.06	0.07	0.06	0.10	0.17	0.03	0.06	0.10	0.04	0.10
	AFO	0.03	0.08	0.01	0.05	0.05	0.05	0.08	0.14	0.02	0.03	0.08	0.03	0.08
TYPE II	RNI	N/A	0.55											
	RNM	N/A	0.59											
	RNO	N/A	0.63											
	RSI	N/A	0.39											
	RSM	N/A	0.46											
	RSO	N/A	0.58											
	AFI	N/A	0.54											
	AFM	N/A	0.56											
	AFO	N/A	0.60											