BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
Statistical Research Report Series
No. RR2000/01

Results of Evaluation of AGGIES for ACES

Maria Garcia and Katherine J. Thompson
Statistical Research Division
Methodology and Standards Directorate
U.S. Bureau of the Census
Washington D.C. 20233

Report Issued: 11/29/2000

Results of Evaluation of AGGIES for ACES

Maria Garcia and Katherine J. Thompson

ABSTRACT

The U. S. Census Bureau's Annual Capital Expenditures Survey (ACES) collects data about domestic capital expenditures in non-farm businesses operating within the United States. Analysts manually edit the ACES data using a specified set of editing rules. Although individual edits are straightforward, the hierarchical combination of edits are complicated with several nested levels of simultaneous balance requirements. We investigate the feasibility of replacing the current ACES editing procedures with an automated system based on National Agricultural Statistics Service's generalized edit and imputation system (AGGIES). The AGGIES system solves simultaneous linear-inequality edits using Chernikova-type algorithms for determining the minimum number of fields to change so that a record satisfies all edits. These algorithms can simultaneously deal with a large number of mathematical constraints and have been successfully applied in Statistics Canada's Generalized Edit and Imputation System and Statistics Netherlands' CherryPI system.

Key Words: data editing, Fellegi-Holt model, error localization

1. Introduction

Data collected using the Annual Capital Expenditures survey (ACES) is the primary source of information about domestic capital expenditures within the United States for federal agencies, private industry organizations, and academic researchers. ACES is a mail-out/mail back survey of companies, collecting detailed data for several related aggregate variables. The current ACES edit system automatically handles certain types of item non-response, then generates lists of edit-failing records which are manually corrected by analysts. Aside from item non-response, there is no automatic resolution of edit-failing records.

In the near future, ACES will move from its current processing system to the U.S. Census Bureau's re-engineered post-data-collection system, the Standardized Economic Processing System (StEPS). The StEPS system is composed of integrated SAS modules that perform several survey processing activities, including data editing, imputation, and estimation (Ahmed and Tasky, 1999). The StEPS editing modules have been used successfully by several other economic surveys at the U.S. Census Bureau. However, none of these surveys collect items that must satisfy several simultaneous ratio edit and balance requirements. Consequently, the StEPS edit software requires some enhancements to correctly automatically edit the ACES data.

The National Agriculture Statistics Service (NASS) has developed a generalized edit and imputation system called AGGIES (Todaro,1999) that solves simultaneous linear-inequality edits using Chernikova-type algorithms for determining the minimum number of fields to change so that a record satisfies the edits. These algorithms can simultaneously deal with a large number of mathematical constraints. Since both AGGIES and StEPS are written in SAS, a successful application of the ACES data with the AGGIES software would provide strong justification for including all or part of the AGGIES system in StEPS.

The objective of this study was to determine the feasibility of replacing all or part of the ACES current manual editing system with an automated editing software while still maintaining (or even improving) data quality. To do this, we conducted a feasibility and an evaluation study using 1996 ACES data with the AGGIES software package. The feasibility study examined whether AGGIES can be used to perform ACES editing. The evaluation study then examined the quality of the edit results. Willimack et al (2000) demonstrates that a large portion of the current ACES analysts' editing strategy does not take a Fellegi-Holt (simultaneous solution) approach. Thus, we use test data with simulated errors to perform the evaluation. ACES does not perform item imputation: instead, they use deterministic (logical) edits to replace missing values or obtain new information directly from the companies. Consequently, this study did not include any evaluation of imputation methods.

This paper provides our results, along with our recommendation. Section 2 provides background on ACES. Section 3 describes the features of AGGIES. Section 4 presents the results of our feasibility study, including detailed descriptions of edits, data groups, and edit groups. Section 5 presents our evaluation study results. Section 6 provides a discussion of our results. We make a few concluding remarks in Section 7.

2. Annual Capital Expenditures Survey

The Annual Capital Expenditures Survey (ACES) collects information about the nature and level of capital expenditures in non-farm companies, organizations, and associations in the United States. The data collected is used to prepare national measures of capital spending and to formulate fiscal and monetary policy. Different forms are mailed to the companies depending on whether they are an employer company or non-employer company. Nonemployer companies respond with a short form questionnaire (ACE-2 form). ACE-2 form responses have a high percentage of UAAs (Undeliverable as addressed), out of scope, and nonresponse (Caldwell, 1999.) Employer companies respond with a long form questionnaire (ACE-1 form). Most of the analyst review concentrates on ACE-1 form responses. This report discusses only employer companies responding with the ACE-1 form.

The ACE-1 form respondents report fixed assets and capital expenditures for the calendar year in all subsidiaries and divisions for all operations within the United States. The ACE-1 form collects data for total capital expenditures in various items of the questionnaire. Total capital expenditures is first reported in survey Item 1. Total capital expenditures is also collected in survey

Item 2 broken down by type of capital expenditures (Structures, Equipment, and Other) cross-classified by new and used capital expenditures as shown below in Figure 1.

Figure 1: Item 2 Capital Expenditures

Capital Expenditures	Total	Structures	Equipment	Other
Total Capital Expenditures	Item 201	Item 202	Item 203	Item 204
New Capital Expenditures	Item 211	Item 212	Item 213	Item 214
Used Capital Expenditures	Item 221	Item 222	Item 223	Item 224

The respondent company reports the same information for each industry in which the company operated and had capital expenditures for the survey year by completing a separate row for each industry in Item 6 of the questionnaire (see Figure 2). The data totals and details that are reported in Item 6 by industry activity must balance to the reported capital expenditures data reported in Item 2, which in turn must agree with the data reported in Item 1.

Figure 2: Item 6 Capital Expenditures by Industry

Industry Category	Total	Structures			Equipment			Other		
		Total	New	Used	Total	New	Used	Total	New	Used
Industry 1	Item 6100	Item 6101	Item 6102	Item 6103	Item 6104	Item 6105	Item 6106	Item 6107	Item 6108	Item 6109
Industry 2	Item 6110	Item 6111	Item 6112	Item 6113	Item 6114	Item 6115	Item 6116	Item 6117	Item 6118	Item 6119
●●●										
Industry XX	Item 6xx0	Item 6xx1	Item 6xx2	Item 6xx3	Item 6xx4	Item 6xx5	Item 6xx6	Item 6xx7	Item 6xx8	Item 6xx9

With the current edit system, the data is first run through a set of if-then-else edit conditions, called insert edits, in a hierarchical order. The insert edits verify that the respondents filled the forms as instructed and resolve item nonresponse using data reported elsewhere in the questionnaire. For example, if total capital expenditures is missing in Item 1 then the total capital expenditures reported in Item 2 is copied to the total capital expenditures entry in Item 1. Similarly, the capital expenditures data reported in Item 2 is copied to Item 6 in the single-industry companies. The insert edits can also correct nonresponse in item's totals by replacing missing totals with the sum of their reported details. The insert edits do not replace reported zero totals with the sum of non-missing reported details. This can occur when a keyer accidentally types a zero in a missing total, for

example.

After insert editing, the data are subjected to the general edits. The general edits check the response records and print messages indicating whether the editing condition is not satisfied, generating an edit referral listing of problem cases for review purposes. Each individual case is displayed interactively, along with a list of its edit failures. Within each case, the analysts review the edit failure listing sequentially to take the appropriate action, such as changing negative reported data, correcting rounding errors, adjusting data to satisfy balancing, or reconciling small discrepancies. Often, edit failures require an analyst telephone call to the company to determine why no data is reported or to verify reported data. According to Caldwell (1999), a large percentage of the collected data is tabulated as reported. Thus, a considerable amount of analysts' time and resources is spent in reviewing forms that were incorrectly flagged as erroneous by the edit system. Also, for those sample firms where the analysts need to change reported data, methods of resolving the same type of edit failures may vary from analyst to analyst (see Willimack, et al, 2000). This indicates that there is a need for automation of all or part of the current manual editing process.

The general edits checks that are used to generate the analyst's review referral listing can be written as a system of linear inequalities describing an acceptable record. The edits are primarily ratio edits and balance (linear equality) edits. This system of linear inequalities can be solved simultaneously. We use this system of linear inequalities in our AGGIES feasibility and evaluation studies described in Sections 4 and 5.

3. Agricultural Generalized Imputation and Edit System (AGGIES)

This section describes the generalized edit and imputation system AGGIES (Todaro,1999), developed at NASS. AGGIES uses the Fellegi-Holt model for editing and imputation (Fellegi and Holt, 1976.) This model requires that the data in each record should be made to satisfy all edits by changing the fewest possible fields. This criterion is referred to as the error localization problem.

With linear inequality edits, identifying the minimum number of fields to change requires iteratively solving a cardinality constrained linear program (Schiopu-Kratina, Kovar, 1989.) Rubin (Rubin, 1975) showed that the solution to the cardinality constrained linear program can be found in the vertices of the region defined by the edit set, called the feasible region. The vertices of this region can be found using an algorithm developed by Chernikova (Chernikova, 1964.) Chernikova-type algorithms have been successfully used by Statistics Canada's Generalized Edit and Imputation System, Statistics Netherlands' CherryPI system, and NASS's AGGIES.

AGGIES is an interactive automated system composed of various modules written in SAS/IML and SAS/AF for editing numeric, continuous, and non-negative economic data. The user specifies the edits interactively as linear inequalities that describe an acceptable record. The acceptable records are contained in the feasible region described by the specified set of linear inequality edits. There is an option to specify a partition of the data and edit set into groups. The edit groups define the conditions that the records in the data group must satisfy. Using the data partition allows the user to run different data records only through the sets of edits involved with

those records.

Once the edit set and edit set partition are ready, the Check Edits module checks the edits for logical consistency, redundancy, determinacy and hidden equalities. A consistent set of edits defines a non-empty feasible region. If the module finds an inconsistent set of edits, then the software outputs a set of edits that if removed will result on a consistent set of edits. The edit set needs to be analyzed and modified until it is consistent. This module also checks the edit set for redundancy. An edit is redundant if removing it from the edit set will not change the feasible region. Since redundant edits do not add any information about the feasible region, they should be removed from the edit set to avoid slowing down the system. After all the redundant edits are removed, the module checks the remaining set of edits for hidden equalities. The Check Edits module also determines lower and upper bounds for the set of response variables.

When the minimal set of non-redundant edits is available, the data groups are submitted to the Edit Summary module. This module determines the number of records that pass the edits and the number of records that fail the edits for each data group and each edit in an edit group.

The edit-failing records identified in the Edit Summary module need to be corrected. For each edit-failing record, the Error Localization module identifies a minimum set of fields to change so that the record satisfies the edits. Reliability weights can be assigned to the variables, with a higher weight indicating higher reliability. If the user assigns reliability weights to the fields during error localization, then this module identifies a weighted minimum set of fields to change so that each edit-failing record passes the edits. The error localization uses Chernikova's algorithm to generate vertices (Chernikova, 1964.) The algorithm is iterative, generating a matrix containing the edges of a region as its columns by solving a series of linear programs at each iteration. These columns correspond to the vertices of some convex cone that represents the possible set of fields that have to be corrected so that an edit-failing record passes the edits. This algorithm is computationally intensive and requires a large amount of computer storage. Since records that contain a high proportion of errors require a large amount of time for error localization, the software has an upper (wall) time limit on the amount of processing time allowed for each record.

When the error localization is complete, we have the minimum set of variables that require imputation for each record that fails at least one edit. The AGGIES imputation module provides six different imputation options, and for each edit-failing record imputation is performed while ensuring that the imputed fields will not continue to fail the edits.

4. Feasibility Study: Adapting AGGIES to edit ACES data

4.1 Test data

For the feasibility study, we developed a test data set consisting of 2,230 records of companies operating in at most four highly correlated industries from the 1996 ACES data collection (1996 ACES data set consists of 42,608 records.) The test deck consisted of records of companies operating in at most four industries in the health sector (Offices & Clinics of Medical Doctors;

Nursing and Personal Care Facilities; Hospitals; and Health and Allied Services, not elsewhere classified.) There are four sample industries in the data set from five strata defined according to payroll size. All records had been previously subjected to the insert edits. Each record contains a census identification number, stratum, representative industry code, a variable counting the number of industries for which the company reported data in Item 6, and all reported data items. Data item variable names correspond to the item codes defined in the original data sets. For example, data for survey item code 201 (total capital expenditures) is represented by AGGIES variable _00201.

The ACES data is stored in a database, and if no data are either reported or inserted for an item, the database does not contain a record for that item. However, AGGIES requires a flat file. When creating a file with one record per company, the data corresponding to these items is identified as missing. This data is ineligible for imputation, so we replaced these values with zeros to prevent AGGIES from automatically flagging the item for imputation.

4.2 Edits

We used the ACES edit failure specifications to write linear inequality edits describing an acceptable record. The edits followed closely the edit failures checks as they appear in the ACES edit box. We tried to maintain consistency between AGGIES edit identifiers and ACES edits identifiers. For example, ACES edit 17 identifies a failing record if:

> "201 not blank or 0 and, 201/payroll $\geq$ U or 201/payroll $\leq$ L, where L and U are respectively the lower and upper limits for the given industry and strata."

We specified AGGIES edits EDLYZ and EDUYZ, where Y corresponds to the industry code and Z corresponds to the strata, as:

EDLYZ: L(Y,Z) * payroll - _00201 $\leq$ 0
EDUYZ: _00201 - U(Y,Z) * payroll $\leq$ 0.

In our test data set with four industry codes and five strata, there are forty linear inequality edits corresponding to ACES edit 17.

Item 6 of the questionnaire collects information on capital expenditures by industry. The respondents must complete a separate row of the questionnaire for each industry in which their company operated and had capital expenditures during the survey year. The sum of expenditures by industry in the total column is detailed by Structures, Equipment and Other (not classified as either structures or equipment.) Further the sum of the total for each of the details is subdetailed by New and Used capital expenditures (see Figure 2.). The fields in Item 6 must satisfy additive relationships between the totals, details and subdetails. Also, the sum of the totals over all industries should be balanced to the totals reported in Item 2. These edits are to be applied depending on how many rows are filled in Item 6 of the questionnaire (how many industries in which the company reported capital expenditures.)

For example, the consistency check for total new capital expenditures (Item 211) and reported new capital expenditures by industry (ACES edit 29) identifies an edit failure if,

$$211 <> \sum 6XX2 + \sum 6XX5 + \sum 6XX8,$$

where XX represents rows 10, 11, ..., 99 of the questionnaire. Corresponding to this edit, we defined AGGIES edit ED291 if the respondent reported capital expenditures in only one industry,

$$\_00211 = \_06102 + \_06105 + \_06108.$$

If the respondent reported capital expenditures in two industries then the corresponding balancing equation is ED292:

$$\_00211 = \_06102 + \_006112 + \_06105 + \_06115 + \_06108 + \_06118.$$

Note that the number of variables that must satisfy this balance requirement depends on the number of industries reported in Item 6.

Similarly, all Item 6 edits are applied to each record depending on the number of industries reported for the company. In this survey, Item 6 reporting pattern varies from record to record since the industries and number of industries operating within each company differs from company to company. To tell the software which edits should be applied to each record, we defined a new variable, NUMIND, which keeps count of the number of industries in which the company had reported capital expenditures. The creation of this new variable was the key to successfully implementing the current ACES edits in AGGIES: it is used in the data group partition to prevent AGGIES from identifying ineligible fields for correction in error localization (see Section 4.3). Appendix 1 provides a complete list of ACES linear inequality edits for AGGIES.

4.3 Edit Groups and Data Groups

Data Groups partition the data set into several disjoint sets of records. Different edit sets can be defined so that different data groups are subjected only to pertinent edits. This reduces the overall computational effort and processing time, particularly in the computationally intensive error localization module, and it prevents ineligible data item fields from being flagged for deletion/imputation. The ratio checks between capital expenditures and payroll (see ED17YZL and ED17YZU above) require that the data are partitioned by sampling strata and sample industry code for each responding unit. The capital expenditures to payroll ratio edit test is applied only if the responding unit reported nonzero capital expenditures on Item 201. Consequently, we must further partition the data according to whether the respondent reported nonzero capital expenditures.

We previously described the assignment of a new variable, NUMIND, that contains the number of industries in which the company reported capital expenditures on Item 6 of the questionnaire. The linear equality edits corresponding to the consistency checks between Item 2 and Item 6 are several simultaneous balance requirements where the balance constraints vary according

to the number of industries in which the company had reported capital expenditures. We used the NUMIND variable to further partition the data according to the number of industries in which a company had reported capital expenditures. The creation of this new variable was the key to successfully implementing the current ACES edits in AGGIES because it prevents AGGIES from identifying ineligible fields for correction in error localization. In summary, the survey records are partitioned in data groups according to the reported capital expenditures, sampling strata, sample industry for the responding unit, and the number of industries in which the company had reported capital expenditures.

For example, we define the first data group as containing all those records for which the reported capital expenditures is nonzero, belong to strata 10, the industry code for the representative sample industry is 801, and reported capital expenditures in only one industry. Using the AGGIES variable names data group one contains all records where:

_00201 > 0 and STRATA = 10 and SAMPLE_I = 801 and NUMIND = 1.

In our test data set, we had one hundred sixty data groups with nonzero capital expenditures. After defining the data groups, we assign the edits to each data group. Appendix 2 contains our complete set of data groups and associated edit groups.

4.4 Set of Non-redundant Edits

After setting up the ACES edit set and edit set partition (edit groups), we used the Check Edits module to test the edit set for inconsistency, redundancy, determinacy and hidden equalities. Our edit set contained a hidden equality that was correctly identified by the AGGIES software. The system did not identify any inconsistency or determinacy in any edit group. However for each group it determined several redundant edits. We used this information to determine which edits should be removed from the edit set to avoid slowing down the system. AGGIES identified the following ACES edits as redundant:

ED08	1*_00201-1*_00202-1*_00203-1*_00204 = 0
ED09	1*_00211-1*_00212-1*_00213-1*_00214 = 0
ED10	1*_00221-1*_00222-1*_00223-1*_00224 = 0
ED11	1*_00201-1*_00211-1*_00221 = 0
ED12	1*_00202-1*_00212-1*_00222 = 0
ED13	1*_00203-1*_00213-1*_00223 = 0
ED14	1*_00204-1*_00214-1*_00224 = 0
ED24100	1*_06100-1*_06101-1*_06104-1*_06107 = 0
ED25100	1*_06101-1*_06102-1*_06103 = 0
ED26100	1*_06104-1*_06105-1*_06106 = 0
ED27100	1*_06107-1*_06108-1*_06109 = 0
ED281	1*_00201-1*_06100 = 0
ED291	1*_00211-1*_06102-1*_06105-1*_06108 = 0
ED301	1*_00221-1*_06103-1*_06106-1*_06109 = 0

ED311	1*_00212-1*_06102 = 0
ED321	1*_00222-1*_06103 = 0
ED331	1*_00213-1*_06105 = 0
ED341	1*_00223-1*_06106 = 0
ED351	1*_00214-1*_06108 = 0
ED361	1*_00224-1*_06109 = 0
ED371	1*_00202-1*_06101 = 0
ED381	1*_00203-1*_06104 = 0
ED391	1*_00204-1*_06107 = 0

This set of edits can be partitioned into three subsets:

E_1 ED08, ED09, ED10, ED11, ED12, ED13, and ED14

E_2 ED24100, ED25100, ED26100, ED27100, ED291, and ED301

E_3 ED281, ED311, ED321, ED331, ED341, ED351, ED361, ED371, ED381, and ED391,

where each edit subset can be identified as redundant in the presence of the other two subsets. For example, E_2 edit ED24100, and edits ED281, ED371, ED381, ED391 from E_3 imply E_1 edit ED08. The Check Edits module output serves as a tool to tell the analyst that a subset of the originally specified edits could be specified in the case of redundancy. In this case, using only the non-redundant edits for error localization will not result in the failing data records satisfying all edits simultaneously. We removed all of the edits in E_1 to obtain a minimal set of non–redundant edits.

4.5 Error Localization

Error localization preserves as much of the originally reported data as possible by identifying for each edit-failing record the fewest possible fields to change so that the record satisfies the simultaneous edit requirements. This is not necessarily the approach currently used by the ACES analysts (Willimack et al, 2000.)

The error localization solution is not always unique. In fact, the system randomly chooses a solution when more than one solution is identified. This affects the repeatability of the results and a possible comparison with the results obtained from manually editing the data. One way to assess this variability is repeatedly running the software and comparing the average data from running the software several times with the manually edited data.

It is possible to significantly reduce the variability or even eliminate it if we assign reliability weights to the data fields (Todaro, 1999). As in many surveys, certain reported ACES data items are consistently more reliably reported than others. For example, we may have more confidence in the reported total capital expenditures from Item 201 than in the reported capital expenditures aggregated over all industries in which the company operated (Item 6 totals). The AGGIES error localization module allows the user to specify reliability weights for each data item.

We examined frequencies of edit changes to ACES historical reported data to specify an initial set of reliability weights, where the higher weight indicates higher reliability. Item reliability weights ranged from 10 (Item 111) to one (for sub-details in Item 6.) In general, the reliability weight for a total item in a balance edit is one unit larger than the weight of the associated details. This initial set of weights was reviewed and modified by the ACES subject matter experts. The complete set of reliability weights appears in Appendix 3.

As mentioned above, using reliability weights with error localization can possibly reduce or eliminate the variability of the results. It was suggested by W. Winkler (private communication) that we could get a unique error localization solution by assigning reliability weights, $1 + 1/p_i$ to field i, where p_i denotes the ith prime. In this survey, we want to keep the weights closed to the originally assigned set of reliability weights, therefore we assigned reliability weights, $w_i + 1/p_i$ to field i, where w_i denotes the originally prescribed reliability weight. In AGGIES, the reliability weights need to be re-entered every time the error localization module is run. We modified the code to reuse the reliability weights from a previous run.

Willimack et al (2000) states that the analysts rarely change a reported zero value without direct confirmation from the company. To preserve this requested edit feature, we added a goldplating option to the existing AGGIES software. Goldplating assigns a higher reliability weight to variables with a reported value of zero, thus making it less likely that the reported zero value will be flagged for deletion by error localization. For this to be effective, reported zeros must be reliably distinguished from zeros that are in the file for any other reason (e.g., keying, blanks).

4.6 Results

Our initial test runs used the complete test set of 2,230 records. The results from these initial test runs were not very encouraging since the error localization took a considerable amount of computer time. Most of the time was consumed by records that passed the allowable time limit for error localization. These records must be reviewed by an analyst outside the automatic editing system.

Upon examining the unresolved cases, we realized that the Fellegi-Holt editing approach is not optimal for resolving all types of ACES balance edit failures. For example, if the percentage difference between the reported totals and aggregated details is small (say five-percent), it may be preferable to rake the details to the total, thus preserving the reported distributions. Similarly, an aggregated industry level reported capital expenditures (Item 6) value that is considerably smaller than the total capital expenditures reported earlier in Item 2 could indicate uncollected industry information. A Fellegi-Holt (AGGIES) approach would automatically correct the data fields provided in the data groups. However, to correctly resolve/investigate this imbalance, an analyst should contact the company or do further research.

Based on these initial results, it was very clear that we needed to combine the Fellegi-Holt approach with certain outside edits available in the StEPS system to properly edit the ACES data.

Together with the subject matter experts we:

1. Identified the cases that **must** be resolved by an analyst prior to submitting to AGGIES.
2. Identified quick data fixes that should be implemented prior to submitting to AGGIES.
3. Produced specifications for deciding which cases will be handled by analysts and which cases will be submitted to AGGIES.

Such pre-editing will greatly reduce processing time while improving data quality. Simple fixes (not requiring Fellegi-Holt type "detective work") can be resolved without using computationally intensive error localization. Records containing fatal errors that must eventually be reviewed by an analyst will be automatically flagged for review (See Appendix 4 for the complete proposed flow of StEPS/AGGIES editing.) These decisions allowed us to eliminate 444 records from our initial test set. Of these, 438 records had either reported zero capital expenditures or with no reported capital expenditures by industry, and six records had uncorrectable keying errors. Of the remaining 1,786 records, there were 592 records failing at least one edit. Our subsequent runs used these 592 edit-failing records.

Our first comparison examined the effect of changing the error localization time limit. Using a stand–alone laptop to avoid network traffic we ran five separate tests, each time increasing the allowable time limit for error localization by ten seconds. If the time limit allowed for error localization is exceeded for any record, the software stops processing that record and proceeds with the next one. Table 1 displays an overview of the performance of the software with these time limits. When the allowable time limit was ten seconds per record, AGGIES error localized 77.5% of the records. When we increased the allowable time limit to 60 seconds per record, AGGIES processed 93.4% of the records.

Table 1: Error Localization Time Limit Results

Time Limit for Error Localization (seconds)	Number of Records Edited [1]	Number of Records Passing Time Limit for Error Localization	Error Localization Time [2] (Hours)
10	459 (77.5%)	133 (22.5%)	0:57
20	507 (85.6%)	85 (14.4%)	1:24
30	531 (89.7%)	61 (10.3%)	1:44
40	540 (91.2%)	52 (8.8%)	1:59
60	553 (93.4%)	39 (6.6%)	2:08

[1]Out of 592 edit-failing records.

[2]All wall-time is equal to CPU time.

Almost 90% of the edit-failing records were error localized when we imposed a time limit of 30 seconds per record, and there were no substantial improvements in amount of records error-localized by increasing the time limit. Tables 2 and 3 present the results from the third run (30 second time limit). Table 2 displays the total number of variables identified to be changed by both AGGIES and the current manual editing procedures. Notice that the number of items changed using the current editing procedures is considerably higher than the number of items identified for change by AGGIES. This result was expected: Willimack et al (2000) reports that analysts generally resolve ACES edit failures by sequentially replacing reported totals with aggregated lower details (e.g., replace Item 2 totals with aggregated Item 6 details). Very little attempt is currently made to resolve all edit failures simultaneously, as done in AGGIES.

Table 2: Total Number of Items Changed by the Current System and by AGGIES

Number of Industries	Total Number of Items	Total Number of Companies with failing records	Number of Items Changed by Manual editing	Number of Items Changed by AGGIES
1	34	543	3606 (19.5%)	1074 (5.8%)
2	44	32	279 (19.8%)	35 (2.5%)
3	54	17	165 (18.0%)	23 (2.5%)
4	64	0	0	0

Table 3 displays the same comparison for a subset of the survey items consisting of Capital Expenditures as reported in Items 1 and 2, and total capital expenditures as the sum of the reported capital expenditures for each industry in which the company had reported capital expenditures. This table also shows that AGGIES consistently made less changes than the current system, preserving more of the originally reported data.

Table 3: Total Number of Times Items Changed by Current Production System and by AGGIES

Number of Industries	Number of Records	Total Capital Expenditures Reported in Item 111		Total Capital Expenditures Reported in Item 201		Sum of Total Capital Expenditures Reported in Item 6 by Industries	
		Current	AGGIES	Current	AGGIES	Current	AGGIES
1	543	148	84	100	85	144	59
2	32	10	6	3	5	9	1
3	17	3	2	2	2	4	1
4	0	0	0	0	0	0	0

The results presented in Tables 1 through 3 demonstrate that it is indeed feasible to edit the ACES data using the AGGIES software. The AGGIES results are not, however, comparable to those

obtained using the current edit procedures: the edit approach is entirely different. To evaluate the quality of the AGGIES edit approach, we needed to conduct a separate evaluation study.

5. Evaluation Study: Quality of AGGIES Editing of ACES Data

5.1 Test data

To test the performance of statistical editing procedures, Granquist (1997) proposes comparing three different files: a file of "true" data values; a file of "raw" data (contaminated by errors); and a file of "cleaned" data (edited data). The edit procedure is evaluated by comparing the true and clean data files.

We use this approach in our evaluation study. Our true data set consisted of 567 1996 ACES cases which satisfied all edits. To produce the raw data, we randomly introduced nonsampling errors into every record in the data set using models proposed by subject matter experts and by Luzi and Della Rocca (1998). Many of the induced errors may not be detectable by AGGIES (e.g., rounding errors where the entire questionnaire is reported in units, combinations of more than one keying error in a balance edit). Table 4 presents frequency distributions of the number of contaminated items per record by number of reported industries. This test deck contains an unrealistically high proportion of records with a large number of errors by design, providing some insight into the types of edit-failing records that can be corrected in AGGIES and testing the limitations of the software in terms of number of edit-failing items.

Table 4: Number of Contaminated Items Per Record

Number of Industries	Number of Errors				Number of Records
	1 - 5	6 - 10	11 - 15	15 +	
1	270	137	42	3	452
2	48	17	7	8	80
3	14	6	5	4	29
4	4	0	1	1	6

We evaluate the edit results using the Manzari and Della Rocca's Indices described below in Section 5.2.

5.2 Evaluation Criteria

Manzari and Della Rocca (1999) proposed the following accuracy indices to evaluate the quality of the editing process.

Table 5: Accuracy Indices (Manzari and Della Rocca, 1999)

Index	Calculation
I1: fraction of true data correctly handled	*d/(c+d)*
I2: fraction of modified data correctly handled	*a/(a+b)*
I3: fraction of total data correctly handled	*(a+d)/(a+b+c+d)*

Where,

modified data = raw data that does not equal true data ("contaminated" items)
a = number of modified data identified to be changed by the editing system
b = number of modified data identified not to be changed by the editing system
c = number of true data identified to be changed by the editing system
d = number of true data identified not to be changed by the editing system

The first two indices are somewhat analogous to the Type I and Type II error measurements used in statistics. An edit is a hypothesis test, where the null hypothesis is that all items involved in the edits are correct. A Type I error flags a "true" value as an error and Type II error fails to flag a "modified" value as an error. With these definitions, the Type I error rate is (1-I1) and the Type II error rate is (1-I2). Thus, I2 measures the power of the edit test (the probability of correctly flagging the incorrect values). Some caution must be used in interpreting these indices with the AGGIES/Fellegi-Holt editing approach. Recall that the Fellegi-Holt edit approach attempts to preserve the maximum amount of reported data. A successful application of this edit approach would yield "high" (close to 1) I1 and I3 indices. However, low I2 indices do not necessarily indicate a failure of this edit approach, since AGGIES searches for a **minimal** deletion set, not the set of all incorrect items (so, for example, if an edit involving two erroneous values can be satisfied by correcting one item, AGGIES will select only one item for correction).

5.2 Results

Because the edit sets are quite different, we separately examine the single-industry company and multi-industry company results. Single-industry companies are instructed to skip Item 6 of the questionnaire; the ACES production system automatically copies the reported capital expenditures from Item 2 into the appropriate Item 6 entries. For this reason, we do not include any Item 6 edits or Item 6/Item 2 consistency edit checks in the single-industry company edit group. Instead, we edit Items 1 and 2 and assume that the corrected data will be copied to Item 6 as a post-edit operation. Multi-industry company edit groups include all Item 6 edits and Item 6/Item 2 consistency checks.

Table 6 displays the three sets of accuracy indices for a subset of items in single-industry companies. We ran three separate tests, each using a different set of item reliability weights. The first run used the set of subject-matter expert defined weights described in Section 4.5. With this set

of weights, AGGIES was more likely to flag a total item for deletion than two details even if both details were incorrectly reported, so in the second run, we assigned a reliability weight of one to all details and a weight of two to all totals. The third run used AGGIES default weights (all weights equal one.) We used a 30 seconds per record time limit for error localization in all runs. Of the 452 single-industry companies in our test deck, there were 69 time limit exceeded records in Run 1, 72 in Run 2, and 73 in Run 3. The data for any record that exceeded the time limit was not used in the calculations.

Table 6: Evaluation Indices for Single- Industry Companies (in Percents)

Run	Index	Item 111	Item 201	Item 202	Item 203	Item 211	Item 212	Item 213	Item 221	Item 222	Item 223
Run 1	**I1**	99	95	97	94	97	84	79	99	99	99
	I2	72	89	13	9	55	45	39	48	24	28
	I3	92	93	72	68	86	73	65	94	91	92
Run 2	**I1**	99	95	94	91	97	85	78	99	99	99
	I2	69	89	13	8	47	39	31	51	26	29
	I3	91	93	70	65	83	73	63	93	91	92
Run 3	**I1**	99	95	94	92	97	85	78	99	99	99
	I2	70	89	13	8	46	39	31	52	28	30
	I3	92	93	70	65	84	73	63	93	91	92

Clearly, AGGIES is performing quite well in terms of correctly preserving true reported data (low Type I error). The I1 indices are all very high (generally more than 84%), indicating that AGGIES is almost never flagging true data (good values) for deletion. The one exception to this is Item 213, which has an I1 index ranging from 78% to 79%. For this item, the edit system is introducing some new errors in the data.

As expected, our results are not as strong in terms of power. In all runs, the values for the I2 index for Items 111 and 201 (total capital expenditures) are greater than 69% and greater than 85% respectively, indicating a high probability of AGGIES correctly flagging erroneous data values for these items. For most other items – the detail items – the value of the I2 index is less than 50%, indicating that the modified values are not consistently identified as erroneous by the editing system. The differences in I2 indices for totals and details are largely a function of the edits: Items 111 and 201 are involved in many more consistency edits than the other detail items. Varying the reliability weights does not appear to improve the I2 indices.

Overall, however, AGGIES is performing quite well for the single-industry data, as evidenced by the I3 index values. Regardless of item reliability weight (run), the I3 indices are greater than 90% for Items 111, 201, 221, 222, and 223, indicating high editing accuracy for these

items. The I3 indices for the other items – the detail items with low I2 indices – are a bit lower. The lowest value of index I3 is 63% (Runs 2 and 3), corresponding to Item 213, the item with the lowest I1 value.

Tables 7 and 8 display the indices for the multi-industry companies. Table 7 displays the results for a subset of variables in Items 1 and 2 of the questionnaire while Table 8 presents results for the totals reported by industry in Item 6 of the questionnaire.

Table 7: Evaluation Indices for Items 1 and 2 in Multi-Industry Companies

Weights	Index	Item 111	Item 201	Item 202	Item 203	Item 211	Item 212	Item 213	Item 221	Item 222	Item 223
Run 1	**I1**	100	98	100	100	100	98	100	100	100	100
	I2	90	100	29	39	0	7	0	0	25	0
	I3	99	99	83	85	80	79	80	94	96	96
Run 2	**I1**	100	98	100	100	100	100	100	100	100	100
	I2	90	100	29	37	0	7	0	0	25	0
	I3	99	99	83	83	81	80	81	94	96	96
Run 3	**I1**	100	98	100	100	100	100	100	100	100	100
	I2	90	100	29	37	0	7	0	0	25	0
	I3	99	99	83	83	81	80	81	94	96	96

Table 8: Evaluation Indices for Item 6 Totals in Multi-Industry Companies

Weights	Index	Item 6100	Item 6110	Item 6120	Item 6130
Run 1	**I1**	100	100	100	100
	I2	50	38	0	0
	I3	94	93	86	75
Run 2	**I1**	100	98	94	100
	I2	13	0	0	0
	I3	90	87	81	75
Run 3	**I1**	100	98	94	100
	I2	50	38	0	0
	I3	94	91	81	75

In all runs and for all items the values of the I1 index indicate that the error localization algorithm is performing well and most of the true data is not flagged for change by the edit. The values of the I2 index range between 0% and 100%. A zero value for an item's I2 index indicates that AGGIES failed to flag any of the modified data values for deletion. Again, this can be a function of the types of errors as well as a function of the edits. The accuracy of the editing process, measured by index I3, is more than 75% for all items.

6. Discussion

The primary purpose of this research was to determine the feasibility as well as the advisability of using the Fellegi-Holt editing approach on the ACES data. The results presented in Sections 4 and 5 demonstrate both the advantages and disadvantages of such an approach.

The main advantage of the Fellegi-Holt approach used by AGGIES is the **simultaneous** consideration of multiple edits. This is a particular strength with the ACES data, since the hierarchical combination of balance edits provides a good deal of information about the reliability of certain data item values. Furthermore, because AGGIES can simultaneously deal with a large number of mathematical constraints, we can greatly strengthen the existing edit. For example:

- We can add ratio tests comparing reported capital expenditures to payroll, namely tests of $\sum$6XX0/payroll (aggregated industry capital expenditures/payroll) and 111/payroll, all using the same upper and lower bounds as the 201/payroll test. This will help determine which of the three reported capital expenditures value(s) is most reliable (since 111, 201, $\sum$6XX0 are expected to be equal);
- We can add ratio tests of current year reported capital expenditures to prior year capital expenditures. In large strata companies, these ratios may not vary much from year to year. With any company, a ratio value larger than 1,000 can indicate rounding errors (values reported in the wrong units) on the current questionnaire;
- We can vary the ratio tests depending on data group. For example, in many industries capital expenditures is not strongly correlated with payroll (e.g., temporary employment agencies, which may have small employment but large expenditures on equipment). In these cases, other ratio tests – such as capital expenditures to sales – may be more appropriate.

This approach is not, however, without its disadvantages. First, it requires pre-editing (e.g., inserting aggregated details for zero or non-reported totals). Second, it does not preserve the distribution of reported details, so is probably not the best approach when the discrepancies between reported total and aggregated details are small. Third, it does not error localize all records with failed edits, especially when the percentage of failed data items is high. This third caveat can also be viewed as a strength, since these probably are the records that require particular analyst investigation. Finally, we were unable to elicit control over the edit results by using item weights, although again we suspect that this is a consequence of nested hierarchy of ACES balance edits (not a function of the editing algorithm).

On balance, we feel that the benefits of using this editing approach in conjunction with the pre-editing described in Section 4.6 outweigh the disadvantages. In general, AGGIES performed quite well with the ACES data.

This is not, however, a wholehearted endorsement of the AGGIES software package. Implementing ACES edits in AGGIES required a great deal of data manipulation. Moreover, the AGGIES software would require several other modifications for a full-scale production test on

ACES data. Examples include developing deterministic/logical imputation edit modules for exact balance solutions (residual imputation), modifying screens to properly define the data groups, and allowing more than twenty variables per edit (needed to edit data for companies reporting capital expenditures in more than six industries.)

Some issues still need to be addressed. For example, we modified the error localization routine by including a gold-plating feature, changing the weight of an item if the reported value for the item is zero. However, we are unable to distinguish reported zeros from missing responses in our input data sets, so the gold-plating feature may prevent legitimate missing values from being flagged for imputation.

7. Conclusion

In this paper we presented the results of a study to determine whether the AGGIES software can be used to perform ACES data editing. The results of our feasibility study were encouraging, so we proceeded to evaluate the accuracy of the results. The results for the evaluation study show that AGGIES performs well in terms of preserving a high proportion of "true" data, although the error localization failed to identify a high percentage of the erroneous data. For now, our recommendation is to further test the existing data edit/groups with no reliability weights on different data sets.

Why do we need further testing? The missing piece in our evaluation is the quality of the edit results: we don't know whether the minimal solution sets are reasonable solutions. Our next step is to work with the subject matter experts on evaluating the quality of the results. The subject matter experts will provide a test deck of ACES unedited data, which we will run through the existing ACES AGGIES edit. The analysts will then review the results on a record-by-record basis. The analysts can decide if the solution obtained using the Fellegi-Holt approach – changing the minimum number of variables so that a record satisfies the edits – is acceptable. We can also modify the reliability weights, edit groups, data groups, time limit for error localization, and pre-editing procedures until the majority of AGGIES edit solutions are acceptable. Other future analysis items include detailed analysis of characteristics of records that exceeded the error localization time limit and investigation into alternative imputation solutions for linear equality edits.

8. References

Ahmed, S. and Tasky, D. (1999). "The Standard Economic Processing System: A Generalized Integrated System for Survey Processing," *Proceedings of the Section on Survey Research Methods,* Alexandria, VA: American Statistical Association, to appear.

Barron, J. and Prince, S. (1999). "1998 ACES Edit Specification," unpublished internal memorandum. Washington, DC: U.S. Bureau of the Census.

Caldwell, C. (1999). "Response Characteristics for the 1997 ACES," unpublished internal memorandum. Washington, DC: U.S. Bureau of the Census.

Chernikova, N. V. (1964), "Algorithm for Finding a General formula for the Nonnegative Solutions of a System of Linear Equations," *U.S.S.R. Computational Mathematics and Mathematical Physics,* No. 5, 228-233.

Cotton, C. (1999). "Functional Description of the Generalized Edit and Imputation System," Statistics Canada Technical Report.

Fellegi, I. P. and Holt, D. (1976) "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association, No. 71, 17-35.*

Granquist, L. (1997). "An Overview of Methods of Evaluating Data Editing Procedures." *Proceedings for the Conference of European Statisticians Statistical Standards and Studies, Section on Statistical Data Editing (Methods and Techniques),* United Nations Statistical Commission and Economic Commission for Europe, pp. 112-123.

Luzi, O. and Della Rocca, G. (1998). "A Generalized 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 and Economic Commission for Europe.

Manzari, A. and Della Rocca, G. (1999), "A Generalized System Based on a Simulation Approach to Test the Quality of Editing and Imputation Procedures," Working paper No. 13, Conference of European Statisticians, Work Session on Statistical Data Editing, United Nations Statistical Commission and Economic Commission for Europe.

Rubin, D. S. (1975), "Vertex generation in Cardinality Constrained Linear Programs, " *Operations Research,* No. 23, 555-565.

Schiopu-Kratina, I., and Kovar, J. G. (1989), "Use of Chernikova's Algorithm in the Generalized Edit and Imputation System," Statistics Canada, Methodology Branch Working Paper No. BSMD-89-001E.

Todaro, T. (1999). "Evaluation of the AGGIES Automated Edit and Imputation System," National Agricultural Statistics Service, USDA, Washington D.C., RD Research Report No. RD-99-01.

Willimack, D., Anderson, A., and Thompson, K. (2000). "Using Focus Groups to Identify Analysts' Editing Strategies in an Economic Survey." *Proceedings of the International Conference on Establishment Surveys II,* Alexandria, VA: American Statistical Association, to appear.

Acknowledgment

The authors would like to thank Richard Sigman and William Winkler for their helpful comments on earlier versions of this manuscript. The authors also wish to thank the ACES experts at the Company Statistics Division: Charles Funk, James Barron, Amy Newman-Smith, Sara Prebble and John Seabold for the helpful discussions leading to the complete flow of StEPS/AGGIES editing proposed in Appendix 4.

Appendix 1: ACES Linear Inequality Edits for AGGIES

Variables' names in AGGIES correspond to ACES's items, the variable corresponding to Item 201 is _00201. The AGGIES's edits follow the numbering in the ACES edit specifications, for example, ED05 corresponds to ACES's Edit 5; ED24XX0 corresponds to Edit 24 for row XX, XX = 10, 11, ..., 99; ED28X corresponds to Edit 28 for numind = X.

ED05	-1*_00101-1*_00111-1*_00121+1*_00131+1*_00141 = 0
ED06	-1*_00141+1*_00151 <= 0
ED07	1*_00111-1*_00201 = 0
ED08	1*_00201-1*_00202-1*_00203-1*_00204 = 0
ED09	1*_00211-1*_00212-1*_00213-1*_00214 = 0
ED10	1*_00221-1*_00222-1*_00223-1*_00224 = 0
ED11	1*_00201-1*_00211-1*_00221 = 0
ED12	1*_00202-1*_00212-1*_00222 = 0
ED13	1*_00203-1*_00213-1*_00223 = 0
ED14	1*_00204-1*_00214-1*_00224 = 0
ED18	1*_00204-1*_00302-1*_00312-1*_00322 = 0
ED19	-1*_00211+1*_00411 <= 0
ED20	-1*_00211+1*_00511 <= 0
ED24100	1*_06100-1*_06101-1*_06104-1*_06107 = 0
ED24110	1*_06110-1*_06111-1*_06114-1*_06117 = 0
ED24120	1*_06120-1*_06121-1*_06124-1*_06127 = 0
ED24130	1*_06130-1*_06131-1*_06134-1*_06137 = 0
ED25100	1*_06101-1*_06102-1*_06103 = 0
ED25110	1*_06111-1*_06112-1*_06113 = 0
ED25120	1*_06121-1*_06122-1*_06123 = 0
ED25130	1*_06131-1*_06132-1*_06133 = 0
ED26100	1*_06104-1*_06105-1*_06106 = 0
ED26110	1*_06114-1*_06115-1*_06116 = 0
ED26120	1*_06124-1*_06125-1*_06126 = 0
ED26130	1*_06134-1*_06135-1*_06136 = 0
ED27100	1*_06107-1*_06108-1*_06109 = 0
ED27110	1*_06117-1*_06118-1*_06119 = 0
ED27120	1*_06127-1*_06128-1*_06129 = 0
ED27130	1*_06137-1*_06138-1*_06139 = 0
ED281	1*_00201-1*_06100 = 0
ED282	1*_00201-1*_06100-1*_06110 = 0
ED283	1*_00201-1*_06100-1*_06110-1*_06120 = 0
ED284	1*_00201-1*_06100-1*_06110-1*_06120-1*_06130 = 0
ED291	1*_00211-1*_06102-1*_06105-1*_06108 = 0
ED292	1*_00211-1*_06102-1*_06105-1*_06108-1*_06112-1*_06115-1*_06118 = 0

ED293 1*_00211-1*_06102-1*_06105-1*_06108-1*_06112-1*_06115-1*_06118
-1*_06122-1*_06125-1*_06128 = 0
ED294 1*_00211-1*_06102-1*_06105-1*_06108-1*_06112-1*_06115-1*_06118
-1*_06122-1*_06125-1*_06128-1*_06132-1*_06135-1*_06138 = 0
ED301 1*_00221-1*_06103-1*_06106-1*_06109 = 0
ED302 1*_00221-1*_06103-1*_06106-1*_06109-1*_06113-1*_06116-1*_06119 = 0
ED303 1*_00221-1*_06103-1*_06106-1*_06109-1*_06113-1*_06116-1*_06119
-1*_06123-1*_06126-1*_06129 = 0
ED304 1*_00221-1*_06103-1*_06106-1*_06109-1*_06113-1*_06116-1*_06119
-1*_06123-1*_06126-1*_06129-1*_06133-1*_06136-1*_06139
ED311 1*_00212-1*_06102 = 0
ED312 1*_00212-1*_06102-1*_06112 = 0
ED313 1*_00212-1*_06102-1*_06112-1*_06122 = 0
ED314 1*_00212-1*_06102-1*_06112-1*_06122-1*_06132 = 0
ED321 1*_00222-1*_06103 = 0
ED322 1*_00222-1*_06103-1*_06113 = 0
ED323 1*_00222-1*_06103-1*_06113-1*_06123 = 0
ED324 1*_00222-1*_06103-1*_06113-1*_06123-1*_06133 = 0
ED331 1*_00213-1*_06105 = 0
ED332 1*_00213-1*_06105-1*_06115 = 0
ED333 1*_00213-1*_06105-1*_06115-1*_06125 = 0
ED334 1*_00213-1*_06105-1*_06115-1*_06125-1*_06135 = 0
ED341 1*_00223-1*_06106 = 0
ED342 1*_00223-1*_06106-1*_06116 = 0
ED343 1*_00223-1*_06106-1*_06116-1*_06126 = 0
ED344 1*_00223-1*_06106-1*_06116-1*_06126-1*_06136 = 0
ED351 1*_00214-1*_06108 = 0
ED352 1*_00214-1*_06108-1*_06118 = 0
ED353 1*_00214-1*_06108-1*_06118-1*_06128 = 0
ED354 1*_00214-1*_06108-1*_06118-1*_06128-1*_06138 = 0
ED361 1*_00224-1*_06109 = 0
ED362 1*_00224-1*_06109-1*_06119 = 0
ED363 1*_00224-1*_06109-1*_06119-1*_06129 = 0
ED364 1*_00224-1*_06109-1*_06119-1*_06129-1*_06139 = 0
ED371 1*_00202-1*_06101 = 0
ED372 1*_00202-1*_06101-1*_06111 = 0
ED373 1*_00202-1*_06101-1*_06111-1*_06121 = 0
ED374 1*_00202-1*_06101-1*_06111-1*_06121-1*_06131 = 0
ED381 1*_00203-1*_06104 = 0
ED382 1*_00203-1*_06104-1*_06114 = 0
ED383 1*_00203-1*_06104-1*_06114-1*_06124 = 0
ED384 1*_00203-1*_06104-1*_06114-1*_06124-1*_06134 = 0
ED391 1*_00204-1*_06107 = 0

ED392 1*_00204-1*_06107-1*_06117 = 0
ED393 1*_00204-1*_06107-1*_06117-1*_06127 = 0
ED394 1*_00204-1*_06107-1*_06117-1*_06127-1*_06137 = 0
EDU80110 -0.4931*PAYROLL+1*_00201 <= 0
EDU80121 -0.2616*PAYROLL+1*_00201 <= 0
EDU80122 -0.1412*PAYROLL+1*_00201 <= 0
EDU80123 -0.6341*PAYROLL+1*_00201 <= 0
EDU80124 -4.6250*PAYROLL+1*_00201 <= 0
EDU80510 -1.3356*PAYROLL+1*_00201 <= 0
EDU80521 -2.1462*PAYROLL+1*_00201 <= 0
EDU80522 -3.9991*PAYROLL+1*_00201 <= 0
EDU80523 -2.9358*PAYROLL+1*_00201 <= 0
EDU80524 -0.0027*PAYROLL+1*_00201 = 0
EDU80610 -0.9369*PAYROLL+1*_00201 <= 0
EDU80621 -0.3143*PAYROLL+1*_00201 <= 0
EDU80622 -0.5975*PAYROLL+1*_00201 <= 0
EDU80623 -0.3035*PAYROLL+1*_00201 <= 0
EDU80624 -0.2500*PAYROLL+1*_00201 <= 0
EDU80910 -0.9414*PAYROLL+1*_00201 <= 0
EDU80921 -0.1828*PAYROLL+1*_00201 <= 0
EDU80922 -0.8461*PAYROLL+1*_00201 <= 0
EDU80923 -0.4444*PAYROLL+1*_00201 <= 0
EDU80924 -1.7143*PAYROLL+1*_00201 <= 0
EDL80110 0.0005*PAYROLL-1*_00201 <= 0
EDL80121 0.0005*PAYROLL-1*_00201 <= 0
EDL80122 0.0006*PAYROLL-1*_00201 <= 0
EDL80123 0.0019*PAYROLL-1*_00201 <= 0
EDL80124 0.0043*PAYROLL-1*_00201 <= 0
EDL80510 0.0002*PAYROLL-1*_00201 <= 0
EDL80521 0.0003*PAYROLL-1*_00201 <= 0
EDL80522 0.0015*PAYROLL-1*_00201 <= 0
EDL80523 0.0019*PAYROLL-1*_00201 <= 0
EDL80610 0.0021*PAYROLL-1*_00201 <= 0
EDL80621 0.0447*PAYROLL-1*_00201 <= 0
EDL80622 0.0635*PAYROLL-1*_00201 <= 0
EDL80623 0.0089*PAYROLL-1*_00201 <= 0
EDL80624 0.0098*PAYROLL-1*_00201 <= 0
EDL80910 0.0003*PAYROLL-1*_00201 <= 0
EDL80921 0.0008*PAYROLL-1*_00201 <= 0
EDL80922 0.0028*PAYROLL-1*_00201 <= 0
EDL80923 0.0014*PAYROLL-1*_00201 <= 0
EDL80924 0.0263*PAYROLL-1*_00201 <= 0

Appendix 2: Edit/Data Groups for Companies reporting Nonzero Capital Expenditures

Group Number	Number of Industries	Sample Industry	Stratum	Edits
1	1	801	10	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80110, EDL80110
2	1	801	2A	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80121, EDL80121
3	1	801	2B	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80122, EDL80122
4	1	801	2C	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80123, EDL80123
5	1	801	2D	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80124, EDL80124
6	1	805	10	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80510, EDL80510
7	1	805	2A	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80521, EDL80521
8	1	805	2B	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80522, EDL80522
9	1	805	2C	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80523, EDL80523
10	1	805	2D	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80524, EDL80524
11	1	806	10	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80610, EDL80610
12	1	806	2A	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80621, EDL80621
13	1	806	2B	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80622, EDL80622
14	1	806	2C	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80623, EDL80623
15	1	806	2D	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281- ED391, EDU80624, EDL80624
16	1	809	10	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80910, EDL80910
17	1	809	2A	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80921, EDL80921

18	1	809	2B	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80922, EDL80922
19	1	809	2C	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80923, EDL80923
20	1	809	2D	ED05-ED07, ED18-ED20, ED24100 - ED27100, ED281 - ED391, EDU80924, EDL80924
21	2	801	10	ED05-ED07, ED18-ED20, ED24100-ED271000, ED24110-ED27110, ED282-ED392, EDU80110, EDL80110
22	2	801	2A	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80121, EDL80121
23	2	805	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80510, EDL80510
24	2	805	2A	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80521, EDL80521
25	2	806	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80610, EDL80610
26	2	806	2A	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80621, EDL80621
27	2	806	2B	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80622, EDL80622
28	2	806	2C	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80623, EDL80623
29	2	809	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED282-ED392, EDU80910, EDL80910
30	3	805	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED24120-ED27120, ED283-ED393, EDU80510, EDL805910
31	3	806	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED24120-ED27120, ED283-ED393, EDU80610, EDL80610
32	3	806	2A	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED24120-ED27120, ED283-ED393, EDU80621, EDL80621
33	3	809	2A	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED24120-ED27120, ED283-ED393, EDU80921, EDL80921
34	4	806	10	ED05-ED07, ED18-ED20, ED24100-ED27100, ED24110-ED27110, ED24120-ED27120, ED24130-ED27130, ED283-ED393, EDU80610, EDL80610

Appendix 3: Item Reliability Weights

Reliability Weights for Items 1 – 5 and Item 6 Totals

Item 1		Item 2		Items 3, 4, 5		Item 6 totals	
Item	Weight	Item	Weight	Item	Weight	Item	Weight
101	9	201	8	302	6	6100	7
111	10	202	7	312	5	6110	7
121	9	203	7	322	4	6120	7
131	9	204	6	411	6	6130	7
141	9	211	7	511	6		
151	8	212	6				
		213	6				
		214	5				
		221	6				
		222	5				
		223	5				
		224	4				

Reliability Weights for Item 6 Totals, New and Used in Structures, Equipment and Other

	Total	Weight	New	Weight	Used	Weight
Structures	6101	6	6102	4	6103	3
	6111	6	6112	4	6113	3
	6121	6	6122	4	6123	3
	6131	6	6132	4	6133	3
Equipment	6104	6	6105	4	6106	3
	6114	6	6115	4	6116	3
	6124	6	6125	4	6126	3
	6134	6	6135	4	6136	3
Other	6107	2	6108	1	6109	1
	6117	2	6118	1	6119	1
	6127	2	6128	1	6129	1
	6137	2	6138	1	6139	1

Appendix 4: Proposed Flow of StEPS/AGGIES editing

Eliminating Scenarios: Cases that Must Be Resolved By An Analyst Prior to Any Automatic Editing

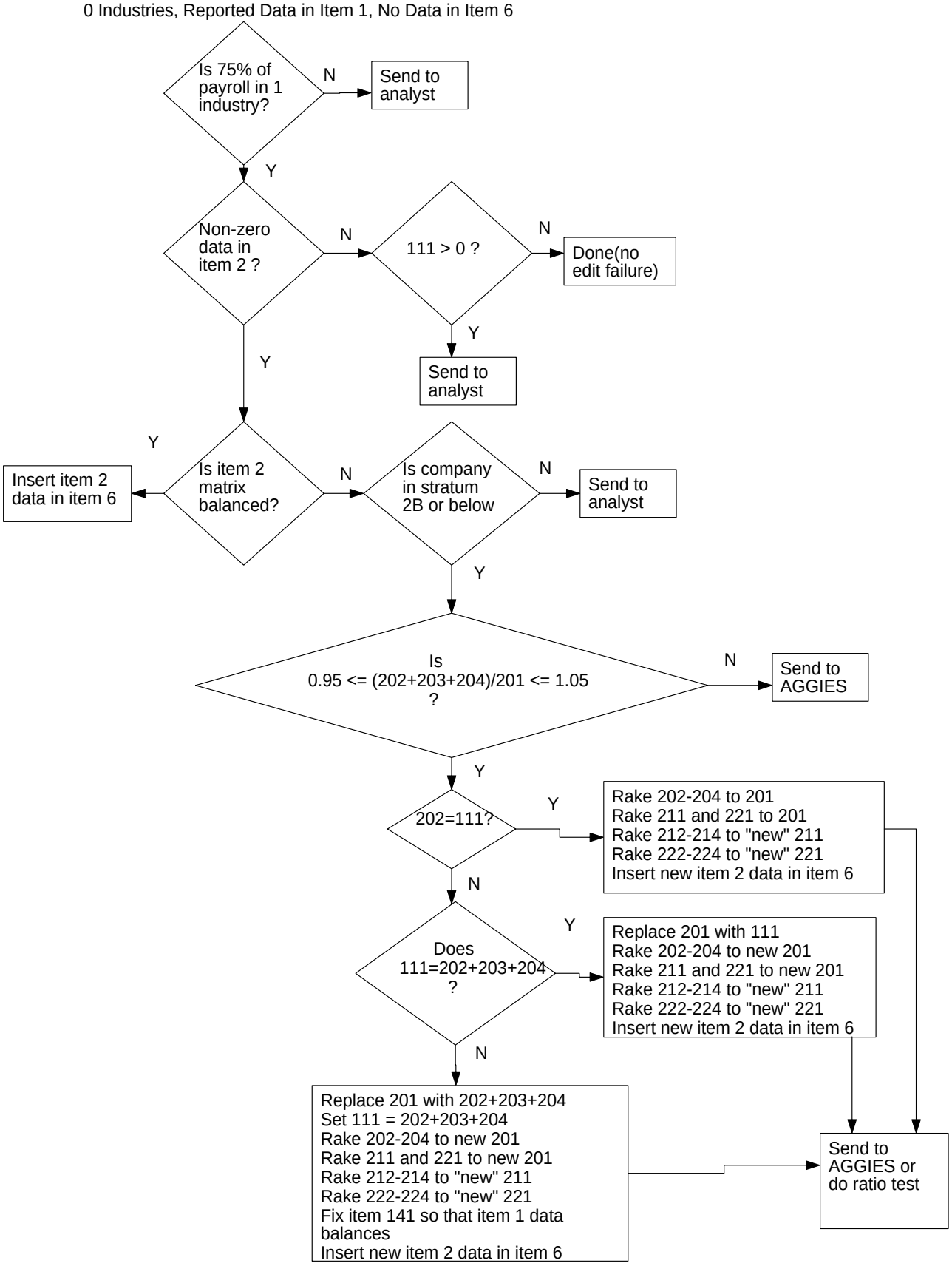
8. Invalid or duplicate industry
9. Blank industry when there is more than one expected industry
3. Large[3] difference between item 6 totals and item 201
4. Strata 1 or 2A cases that do not report capital expenditures (item 111 or 201 is blank/reported zero)
5. Strata 1 or 2A cases that do not report fixed assets (item 101 is blank/reported zero)
6. Any cases where 204, 214, or 224 > 0
7. Any cases where 6XX7, 6XX8, or 6XX9 > 0
8. Negative values for all items except 131 or 151
9. Cases where 411 > 211
10. Cases where 511 > 211
11. Cases where ratio of current year sales to prior year sales are out of tolerance, if both years' sales are reported (Ratio edit of $\left(a \le \frac{\text{Sales00}}{\text{Sales01}} \le b \right)$, where tolerances *a* and *b* will be determined later).

Quick Fixes

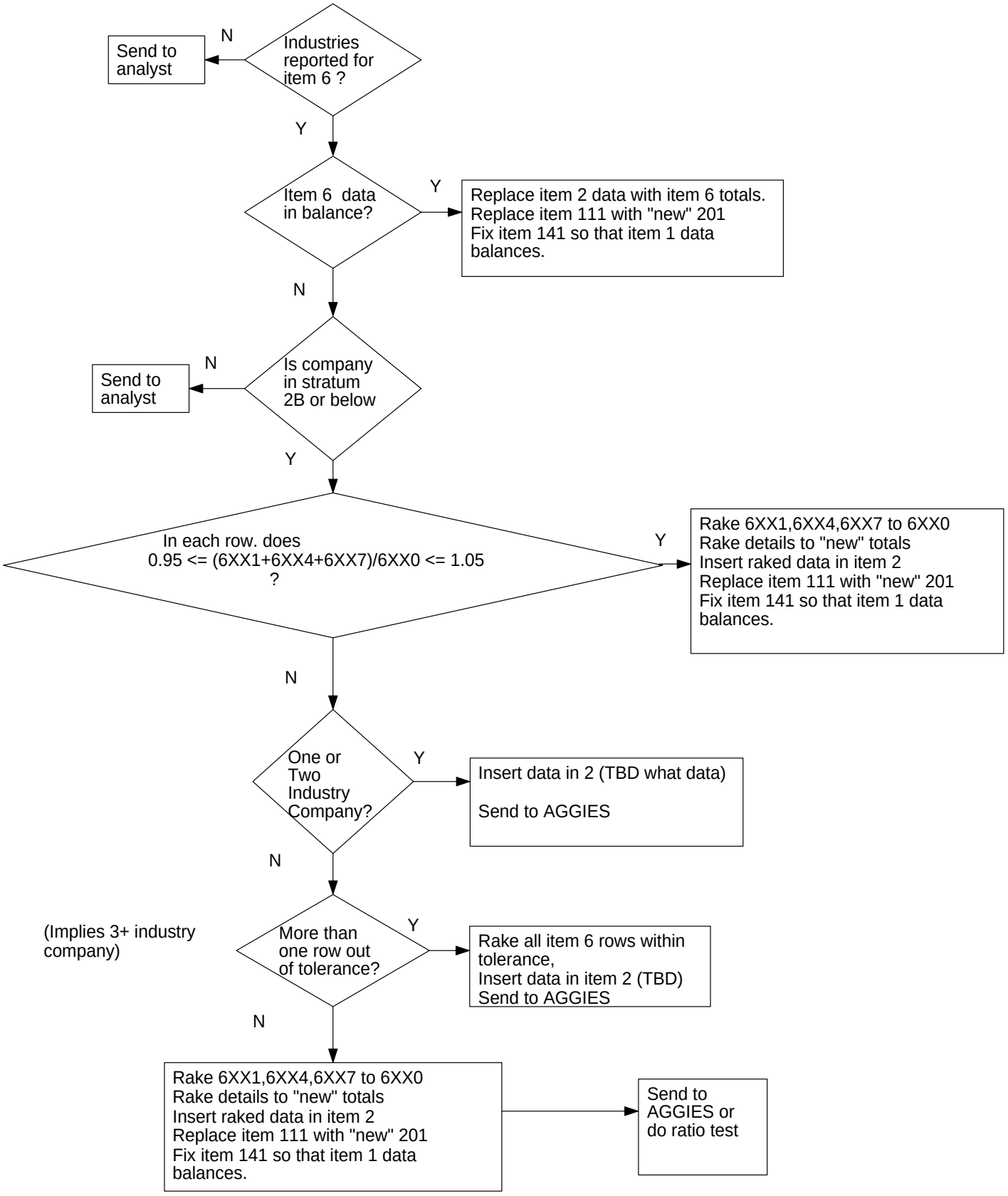
1. Change all negative signs on items 131 and 151 to positive
2. If any data is reported under industry '010,' then delete and subtract for 111 and 211.

[3]Large to be defined later.

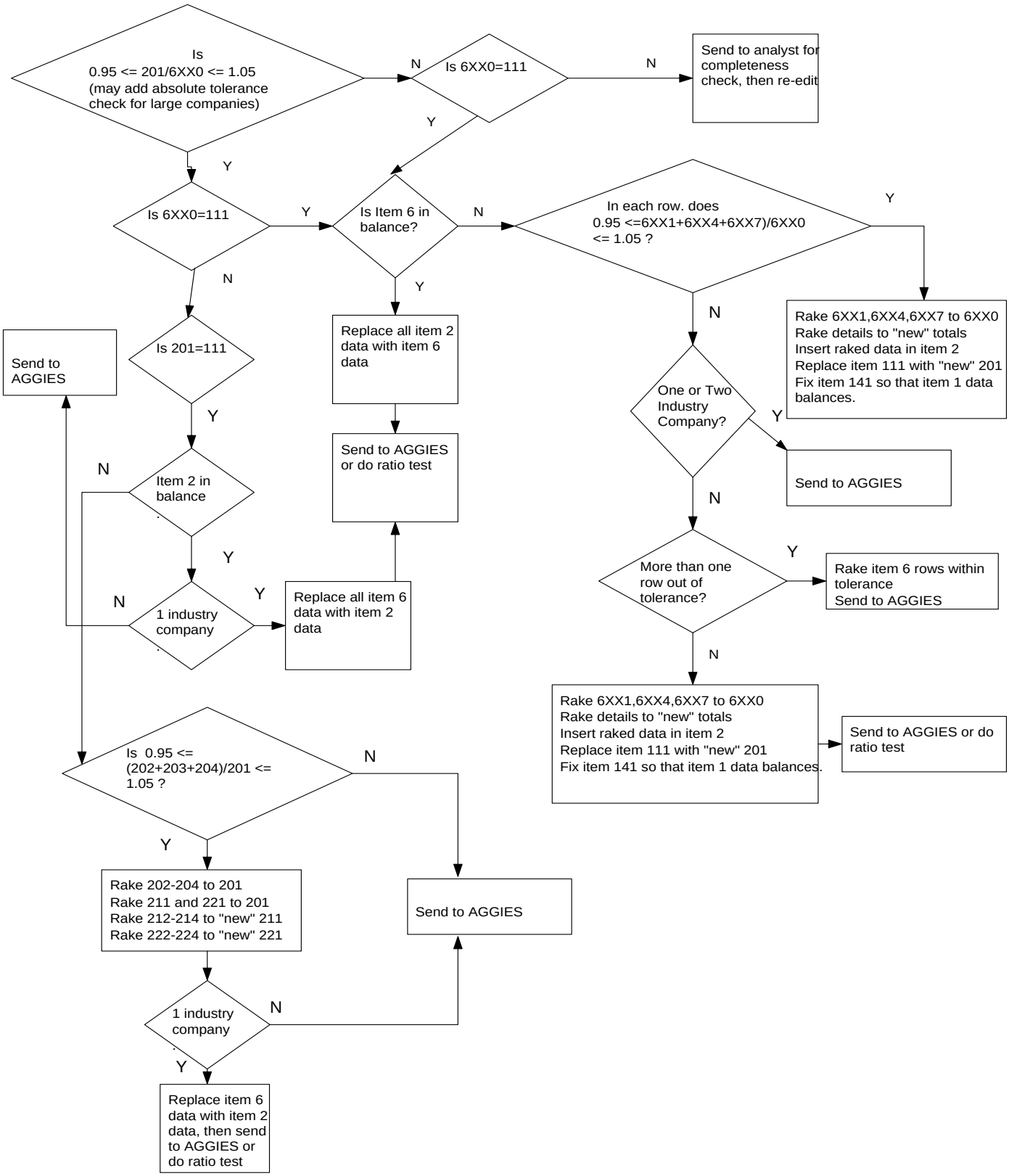
0 Industries, Reported Data in Item 1, No Data in Item 6



Data Reported in Item 6, No Reported Data in Items 1 or 2



Item 2 and Item 6 have Data, But 201 NE 6XX0



Item 201 and Item 6XX0 not equal, Item 6 balanced, item 2 not balanced