A Cluster Analysis Approach To Describing Tax Data

Brian G. Raub and William W. Chen, Internal Revenue Service

The Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) produces data using information reported on tax returns. These administrative data are used by the Department of the Treasury, the Joint Committee on Taxation, and various Federal statistical agencies and are disseminated to the public via the World Wide Web and publications such as the *SOI Bulletin*. The Corporate Foreign Tax Credit (CFTC) study is in many ways typical of SOI studies. Data are collected from tax forms (in this case Form 1118) by SOI field staff and are subjected to error resolution by analysts at National Headquarters. The error-resolved data are used to create statistical tables that are published annually with descriptive text and technical notes. These statistical tables display selected aggregate fields from Form 1118 by industry, type of income, and country to which foreign taxes were paid.

The present paper will describe a population of Form 1118 filers using cluster analysis, with the goal of identifying alternative ways of organizing and analyzing tax data. A second goal is to identify new insights about this population of filers.

▶ Background

The Corporate Foreign Tax Credit is claimed by U.S. multinational firms to offset some or all of their taxes paid to foreign countries. Under U.S. tax law, U.S. corporations are taxed on income earned both in the U.S. and in foreign countries. Income earned in foreign countries may also be subject to taxation by the authorities in those foreign countries, resulting in double taxation. The foreign tax credit was adopted to alleviate this problem.

To claim the foreign tax credit, U.S. corporations file Form 1118, *Foreign Tax Credit--Corporations*. On this form, taxpayers report their incomes within broad categories such as interest, dividends, services, rents, and other. Deductions and tax liability are also reported. Further, taxpayers are required to report these items detailed by country.

For 2001, taxpayers were required to segregate their incomes, deductions, and taxes into several limitation categories, or "baskets," such as the Passive Income basket or the General Limitation Income basket. A separate foreign tax credit was calculated for each basket, with the total foreign tax credit being the sum of the separate foreign tax credits from each basket. The purpose of this provision and related limitations was to prevent taxpayers from using foreign tax credits to offset taxes on U.S.-source income, thus denying the United States tax revenues due on income earned domestically.

For Tax Year 2001, U.S. corporations claimed a combined $41.1 billion in foreign tax credits. This was the single largest type of tax credit, accounting for 86.7 percent of all credits claimed by corporations in that tax year. This credit is elective, meaning that, if the taxpayer chooses to take the credit, no deductions for those foreign taxes are available. A majority of taxpayers decide to take the credit, since it offsets the U.S. income tax dollar for dollar, unlike a deduction, which may only offset every dollar of U.S. tax by the percentage of the tax rate [1].

▶ Data Description

The 2001 CFTC study is based on a stratified, weighted sample of corporation income tax returns with a foreign tax credit that were included in the 2001 SOI sample of returns with accounting periods ending between July 2001 and June 2002. These returns were selected after administrative processing but prior to any amendments or audit examination. The corporate tax return forms included in this sample were Forms 1120, 1120S, 1120-L, 1120-PC, 1120-REIT, and 1120-RIC.

The 2001 CFTC data sets contain 2,563 returns claiming foreign tax credits. These returns are weighted up to a population estimate of 5,478 returns. For the present paper, we used a “defined population” approach by including only those returns with a sample weight of 1. This defined population of 1,075 returns accounted for an estimated 98.3 percent of the total foreign credit claimed on all returns for 2001.

▶ Cluster Analysis

Cluster analysis, or clustering, refers to a set of mathematical techniques for sorting observed data into groups so as to maximize the similarity of observations within the same group and minimize the similarity of observations across different groups. These techniques can be used to discover associations and structures within a data set that may not have been known. Cluster analysis has been widely used in the biological and social sciences to help define classification schemes or taxonomies. It has also been used to suggest new ways of describing a population in business and marketing applications.

Cluster analysis techniques can be broadly separated into two approaches, hierarchical and nonhierarchical. The hierarchical approach builds clusters of successively larger size using some measure of similarity or distance. Typical algorithms used in this approach include single linkage (nearest neighbor), complete linkage (furthest neighbor), and Ward’s Method, which minimizes the mean square distance between the center of a cluster and each member. Nonhierarchical clustering approaches also exist, including the K-means method.

For the present data set, we chose hierarchical clustering since this set of techniques is available in SAS’s PROC CLUSTER. We clustered a sample of our data set using each of the 11 methods available in SAS and ultimately selected Ward’s Method for two main reasons. First is the efficiency of this method, useful given the relatively large number of observations (1,075) and clustering variables (9). Second is the tendency of this method to create clusters of relatively equal size. We noted a strong tendency for other clustering algorithms to create clusters with very few observations. Although the existence of these outliers may be an interesting outcome in a subject-matter sense, allowing very small clusters could create a disclosure problem [2].

In Ward’s Method, the distance between two clusters is defined as

D_{KL} = distance between clusters C_K and C_L

$$D_{KL} = \left\|\overline{x}_K - \overline{x}_L\right\|^2 /(1/N_K + 1/N_L)$$

where

$C_K =$ ${}_K$th cluster, subset of {1,2,…,n}

$x_i =$ ${}_i$th observation

N_K = number of observations in C_k

X_K = mean vector for cluster C_K

$\|x\|$ = Euclidian length of the vector x, that is, the sum of the squares of the elements of x.

If the distance between observations x and y ,d(x,y)= $\|x-y\|^2/2$, then the combinatorial formula is

$$D_{JM} = (N_J + N_K)D_K + (N_J + N_L)D_{JL} - N_J D_{KJ})/$$

$$(N_J + J_M)$$

The distance between two clusters is the ANOVA sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation [3].

To define our clustering variables, we started by considering the main variables in the CFTC study data sets: selected data from Form 1120; gross income and deduction items from Form 1118, Schedule A; foreign tax items from Schedule B, Part I; and foreign tax credit computation items from Schedule B, Parts II and III. The first variable of interest that we identified was the total foreign tax credit, which is calculated on Form 1118, Schedule B, Part III and carried over to Form 1120. One concern that we identified immediately is that the total foreign tax credit amount varies significantly by corporation and is strongly correlated to the overall size of the corporation. Therefore, clustering on this variable in its original form would tend to create clusters based primarily on the size of the corporation. This clustering would add little to our current knowledge of the filer population and would likely fail to capture relationships between other clustering variables. To overcome this limitation, we standardized this variable by taking the ratio of the total foreign tax credit to the corporation's income tax liability.

Since the types of income, deductions, and taxes reported by taxpayers are important elements of the CFTC study, we chose to use a set of variables that capture these elements. As deductions and taxes for each income type are closely correlated with the gross income for that type, we decided that including deduction and tax variables in our clustering would add little value. Thus, we focused only on gross income for each type--dividends, interest, rents, services, and other. We also standardized each of the gross income variables into a ratio by dividing the total for each type of gross income by the total gross income for the corporation. These ratios became five of our clustering variables.

The final data element of the CFTC data set that we used in our cluster analysis was foreign-source country of the gross income reported by each corporation. Defining clustering elements based on country proved to be somewhat challenging, however, since there are over 300 countries in our system, and it was necessary to limit the number of clustering variables for the sake of efficiency. Ultimately, we decided to create variables for the top three countries as defined by amount of total gross income. These three countries, Canada, Japan, and the United Kingdom, combined for 32.6 percent of the total gross income reported by the firms in our defined population. The corresponding clustering variables were defined as the ratio of gross income allocated to each country to the total amount of gross income for each company. Figure 1 summarizes the clustering variables by description and the names we assigned.

Determining the number of clusters to be used in this cluster analysis was largely a heuristic process.

Figure 1.--Clustering Variables

Variable Name	Variable Description
FTC	Foreign tax credit divided by income tax liability
Dividends	Dividend income divided by total gross income
Interest	Interest income divided by total gross income
Rents	Rents income divided by total gross income
Services	Services income divided by total gross income
Other	Other income divided by total gross income
UK	UK-source income divided by total gross income
Japan	Japan-source income divided by total gross income
Canada	Canada-source income divided by total gross income

From a subject-matter standpoint, we began with the assumption that it made sense to look for at least three clusters but that more than eight clusters would become cumbersome and provide less valuable insight into our defined population. After considering the output from these options, we concluded that viewing our data in four clusters provided the most insight into our data and could be described most effectively. We named these clusters "High Dividend Firms," "Low CFTC/Other Income Firms," "Interest/ Service Firms," and "High CFTC/Manufacturing Firms."

▶ Clustering Results

Figure 2 displays the number of observations in each cluster.

Figure 2. --Cluster Summary

Cluster	Number of Observations
High Dividend Firms	295
Low CFTC/Other Income Firms	201
Interest/Service Firms	367
High CFTC/Manufacturing Firms	208

The relative similarity in the number of observations in each cluster is consistent with our choice of Ward's Method for our clustering algorithm, while the absence of very small clusters serves our requirement of protecting taxpayer confidentiality.

In comparing the makeup of the four clusters below, we will use the average of each variable for the firms in the respective cluster, expressed as a percentage rather than a pure ratio for ease of use.

The "High Dividend Firms" cluster is summarized in Figure 3. Dividends is the dominant income variable with an average of 72.0 percent, while the average Interest, Rents, and Services are all below 5.0 percent. The average FTC for "High Dividend Firms" is 16.7 percent, below the overall average of 32.4 percent for companies in our defined population. The UK variable has the highest average value among the four clusters at 15.4 percent, while the average Japan variable is the lowest among the clusters at 0.9 percent.

Figure 3.--"High Dividend Firms" Summary

Variable	Average Percentage Value
FTC	16.7
Dividends	72.0
Interest	3.1
Rents	4.7
Services	1.6
Other	6.7
UK	15.4
Japan	0.9
Canada	18.8

As seen in Figure 4, the average company in "Low CFTC/Other Income Firms" has a significantly different set of characteristics. For this group, the dominant income variable is Other, with an average of 82.8 percent. In contrast, the average Services and FTC values in this cluster are the lowest among the four clusters at 0.6 percent and 8.3 percent, respectively. The average country variables for this cluster are middling--with neither a high nor a low for any country variable among the clusters.

Figure 4.--"Low CFTC/Other Income Firms" Summary

Variable	Average Percentage Value
FTC	8.3
Dividends	4.1
Interest	4.9
Rents	5.7
Services	0.6
Other	82.8
UK	13.5
Japan	4.9
Canada	16.8

Summary statistics for "Interest/Service Firms" appear in Figure 5. For companies in this cluster, Interest, Rents, and Services incomes combine for nearly all of the gross incomes, with an average Interest of 33.4 percent, an average Rents of 31.1 percent, and an average Services of 23.2 percent. The average FTC for companies in this cluster is below the average of all the companies in our defined population at 15.8 percent. Among the country variables, the average Canada and Japan values are the highest of any cluster, 23.1 percent and 8.1 percent, respectively, while the average UK value is the lowest at 9.2 percent.

Figure 5.--"Interest/Service Firms" Summary

Variable	Average Percentage Value
FTC	15.8
Dividends	5.7
Interest	33.4
Rents	31.1
Services	23.2
Other	4.4
UK	9.2
Japan	8.07
Canada	23.1

Figure 6 displays the variable averages for companies in "High CFTC/Manufacturing Firms." Other is the dominant income variable with an average of 36.0 percent, followed by Dividends and Rents with 28.8 percent and 15.0 percent, respectively. The average FTC of companies in this cluster is dramatically larger than for any other cluster at 80.2 percent. Among the country variables, the average Canada value is the lowest of the four clusters at 7.1 percent, as is the combined average of the three country variables, 24.6 percent.

Figure 6.--“High CFTC/Manufacturing Firms” Summary

Variable	Average Percentage Value
FTC	80.2
Dividends	28.8
Interest	5.3
Rents	15.0
Services	1.7
Other	36.1
UK	12.4
Japan	5.2
Canada	7.1

▶ Industry Analysis

One additional element of note in the CFTC data is the industry classification of the companies filing Form 1118. Using industry classification in our cluster analysis, however, proved infeasible. Although each corporation in our defined population has a six-digit industry code assigned to it using the North American Industry Classification System (NAICS), this number is of an ordinal, rather than cardinal, nature. Therefore, although the NAICS code could be used as a clustering value, interpreting and describing the meaning of the industry code in the clustering output would be problematic. However, because industry classification is an element of interest, we analyzed the industry breakdown for each cluster ex post facto.

Our industry analysis reveals significant differences between clusters. Although Manufacturing, the largest industry among the firms in our defined population, represents a significant portion of the observations in each cluster, its contribution to the clusters ranged from 26.2 percent of “Interest/Service Firms” to 63.9 percent of “High CFTC/Manufacturing Firms.” Mining, Utilities, and Construction companies are distributed relatively evenly between the clusters, with a low of 4.0 percent and a high of 7.2 percent. The remaining four industries make up more widely varied portions of the cluster totals. The Finance, Insurance, Real Estate, and Rental and Leasing industry makes up a low of 4.3 percent of “High CFTC/Manufacturing Firms” but a high of 33.6 percent of “High Dividend Firms.” Information companies comprise 3.7 percent of “High Dividend Firms” but 8.2 percent of “High CFTC/Manufacturing Firms.” Services companies make up only 6.0 percent of “Low CFTC/Other Income Firms” but 23.2 percent of “Interest/Service Firms.” Distribution and Transportation companies make up 8.2 percent of “High CFTC/Manufacturing Firms” but 17.4 percent of “Low CFTC/Other Income Firms.”

The industry distribution of “High Dividend Firms,” shown in Figure 7, reveals that Finance, Insurance, Real Estate, Rental, and Leasing is the dominant industry, comprising 33.6 percent of this cluster. This is the highest percentage of firms in this industry among the four clusters. The 13.2 percent of companies in the Services industry was the second highest among the clusters, while the 3.7 percent of companies in the Information industry was the lowest.

Figure 7.--“High Dividend Firms” Selected Industry Breakdown

Industry	Percent of Total
Mining, Utilities, and Construction	6.4
Manufacturing	30.2
Distribution and Transportation	11.9
Information	3.7
Finance, Insurance, Real Estate, Rental and Leasing	33.6
Services	13.2

The industry distribution of “Low CFTC/Other Income Firms,” shown in Figure 8, reveals that companies in the Distribution and Transportation industry represent a larger share than in any other cluster, with 17.4 of the total. In contrast, companies in the Services industry represent a smaller share of the total, 6.0 percent, than in any other cluster.

Figure 8.--“Low CFTC/Other Income Firms” Selected Industry Breakdown

Industry	Percent of Total
Mining, Utilities, and Construction	4.0
Manufacturing	39.8
Distribution and Transportation	17.4
Information	7.5
Finance, Insurance, Real Estate, Rental and Leasing	23.4
Services	6.0

Figure 9 displays the industry distribution of “Interest/Service Firms.” This cluster has the highest concentration of companies in the Services industry, 23.2 percent, and the lowest concentration of companies in the Manufacturing industry, 26.2 percent. “Interest/Service Firms” has 367 members, the most among the four clusters.

Figure 9.--“Interest/Service Firms” Selected Industry Breakdown

Industry	Percent of Total
Mining, Utilities, and Construction	6.0
Manufacturing	26.2
Distribution and Transportation	12.8
Information	6.8
Finance, Insurance, Real Estate, Rental and Leasing	24.0
Services	23.2

As seen in Figure 10, manufacturing firms dominate the “High CFTC/Manufacturing Firms” cluster, with 63.9 percent of the total, while the other industry groups each comprise 8.2 percent or less of the total.

Figure 10.--“High CFTC/Manufacturing Firms” Selected Industry Breakdown

Industry	Percent of Total
Mining, Utilities, and Construction	7.2
Manufacturing	63.9
Distribution and Transportation	8.2
Information	8.2
Finance, Insurance, Real Estate, Rental and Leasing	4.3
Services	8.2

▶ Implications

To gauge the effectiveness of cluster analysis in gaining insight to our data, we should consider its value to analysts both within SOI and outside. To SOI analysts who work with the CFTC data, some of the output of this cluster analysis may seem relatively obvious and merely confirms prior knowledge about our defined population. An example of this kind of result is that firms in the “High CFTC/Manufacturing” cluster, dominated by manufacturing companies, claim the highest average foreign tax credit as a percentage of their income tax liabilities. On the other hand, at least one output of our cluster analysis was somewhat surprising: the relationship between reporting primarily Other gross income and offsetting a relatively smaller portion of tax liability with foreign tax, revealed in the “Low CFTC/Other Income Firms” cluster. Although it may have been possible to find this relationship by exhaustively querying our data files, cluster analysis has here served a useful function by pointing us in the right direction for further inquiry.

To those outside SOI who use CFTC data, our cluster analysis may also have value. Because, in most cases, users outside the Department of the Treasury do not have access to our data files, their ability to use our data is limited by what we provide in the published tables or in requested special tabulations. For example, while our published data tables do include summary statistics by industry and by country, they do not capture both relationships together as does our cluster analysis with the ex post facto industry distribution. Here again, the output from our cluster analysis may serve a useful function in revealing areas for further research.

▶ Limitations

The 2001 Corporate Foreign Tax Credit statistics quoted in this article do not represent the final amounts credited that year. Complete foreign tax credit statistics for 2001 would reflect the results of any audits. Also, some corporations did not file Form 1118 because they did not have a U.S. income tax liability and were, thus, unable to credit any foreign taxes paid, accrued, or deemed paid for 2001. Finally, other corporations could have deducted their foreign taxes from their gross incomes instead of claiming a foreign tax credit.

As noted above, our analysis used only those firms from our sample with a weight of 1, i.e., those not weighted up to represent a greater part of the population estimates. This group of companies combined to claim 98.3 percent of all CFTC tax credits. Thus, while our analysis includes the large companies that claim an overwhelming majority of the total dollar amount of credits, it excludes many small companies that claim comparatively small CFTC's.

The output of our cluster analysis depended to a significant extent on choices made about our clustering techniques and our selection of clustering variables. As noted above, selecting which clustering algorithm to use and the number of clusters in the output is largely a heuristic process. Our set of clustering variables does not take into account several broad elements of the CFTC data sets, including "limitation baskets," data from Schedules F, G, H, I, and J, and country detail other than for Canada, Japan, and the UK.

▶ Conclusion

Cluster analysis can be a useful set of techniques for exploring and describing data sets, including those produced by SOI based on tax return data. By identifying relationships among the variables that are not immediately obvious to internal or external researchers, clustering can enhance knowledge of the data set and serve as the starting point for further research. The costs of cluster analysis should be manageable in many applications, since widespread software tools such as SAS® include clustering capability.

One challenge in using cluster analysis for data sets like those produced by SOI is that these tools may add the most value for data sets with a very large number of observations and/or variables where relationships may be more difficult to identify by other techniques. However, these data sets may also be the most difficult to model for efficient clustering. In these cases, an alternative algorithm such as SAS's PROC FASTCLUS may be more appropriate, though at a loss of power and flexibility relative to PROC CLUS.

Another potential challenge in using cluster analysis on data sets like those produced by SOI presents itself for those which use sampling and weighting. Many data sets are significantly less "top-heavy" in dollar terms than the CFTC data set. In these cases, using only returns with a weight of 1 might entail the exclusion of many observations of interest from the clustering analysis. In the alternative, using returns with a weight of greater than 1 would require additional statistical considerations. The tradeoffs between these approaches could be analyzed using a Pareto analysis of the observations in the data set.

Thus, while cluster analysis can be a useful tool for data exploration and description in applications such as SOI's Corporate Foreign Tax Credit project, further study is needed to assess its potential costs and benefits for larger data sets.

▶ Endnotes

[1] For more background on the Corporate Foreign Tax Credit, see Luttrell, Scott, "Corporate Foreign Tax Credit, 2000," *Statistics of Income Bulletin*, Fall 2004, Volume 24, Number 2.

[2] The Internal Revenue Code prohibits the IRS from releasing information that could be used to identify specific taxpayers.

[3] Description of Ward's Method adapted from *SAS/STAT User's Guide, Version 6*.