A COMPARISON OF ADJUSTMENT TECHNIQUES TO SUPPLEMENT A NETWORK FLOW DISCLOSURE AVOIDANCE SYSTEM

Colleen M. Sullivan, U.S. Bureau of the Census
Commerce/Census/ESMD, Washington, D.C. 20233-5800

1. INTRODUCTION

The United States Bureau of the Census has the responsibility to collect data regarding economic sectors and to publish this data without violating confidentiality laws. Collected data contain sensitive data values that if directly published could identify an individual establishment's data. To protect this data, we suppress the cells containing these values in the publication. We call these suppressed cells primary suppressions. However, since data appear in additive tables, additional cells also must be suppressed in the publication. We call these additional suppressed cells complementary suppressions. Our objective in applying complementary suppressions is to minimize the sum of the data values chosen as complementary suppressions. We employ this objective because we wish to provide the greatest amount of usable data to our data users. (Sullivan (1992a) presents a more complete discussion of disclosure topics.)

Furthermore, the Census Bureau uses complementary suppressions to ensure that a data user cannot estimate the value of a sensitive data cell within a predefined interval. That is, when choosing complementary suppressions for some primary suppression with true value X, we ensure that it cannot be estimated within a smaller interval than [X-L,X+U] where L is the amount of lower protection required by X, and U is the amount of upper protection required by X. We call this the amount of protection required by the primary suppression. Kelly, et al. (1991) discusses protection levels in greater detail.

The Economic Divisions of the Bureau employ a cell suppression technique that uses network flow methodology to apply complementary suppressions to tabular data. We implement this computationally fast technique using a program we developed that prepares the tabular data for input into the commercially available Minimum Cost Flow program of Glover, Klingman, and Mote. The use of network flow methodology as a cell suppression technique is discussed in greater detail in Sullivan and Zayatz (1991), Sullivan and Rowe (1992), and Sullivan (1992b). In the remainder of this paper, "MCF" will refer to using the front-end program we developed along with the Minimum Cost Flow program.

MCF individually checks each primary suppression contained in a table to ensure it cannot be derived by using non-suppressed cells from the table. If a primary suppression is determined by MCF to be derivable, a set of cells is chosen to serve as complementary suppressions. Cells are chosen to serve as complementary suppressions based on their corresponding cost and capacity. The cost of a cell corresponds to the cost incurred by removing the cell from the publication. The cost of a cell can be the data value contained in the cell or some other value based on subject matter expertise. The capacity represents the amount of protection the cell contributes to the protection required by the primary suppression. Sullivan (1992b) and Jewett (1992) describe costs and capacities in greater detail.

Since MCF merely approximates an optimal integer programming formulation, it may choose more cells than necessary to serve as complementary suppressions. Therefore, we have developed two adjustment techniques to supplement the MCF program. I will refer to the first technique as the cost adjustment method and the second as the hybrid method. The objective of both techniques is to remove the superfluous complementary suppressions applied by MCF, resulting in the publication of a larger quantity of data.

The optimal technique for choosing complementary suppressions is the integer programming routine outlined in Section 2. However, since this routine is computationally impractical for census tables, the MCF approximation is used. Although MCF is computationally fast, it often oversuppresses due to the structure of the objective function (See Section 3). This paper discusses two adjustment techniques to remove superfluous complementary suppressions applied by MCF. Sections 4.1 and 4.2 present the framework of the cost adjustment method and the hybrid method, respectively. Section 5 reports the results of applying MCF, the cost adjustment method, and the hybrid method to several sets of actual data.[1] Section 6 defines and discusses what we call "backtracking" and how this affects each method. Section 7 presents some concluding remarks.

2. THE IP FORMULATION

The cell suppression problem has a theoretical integer programming (IP) statement that produces an optimal solution (Sullivan and Rowe, 1992). In this routine, there is an indicator variable, I_{ij}, that is restricted to be zero or one. Thus, we consider decisions in which just

two outcomes are possible: we either assign a complementary suppression to a particular table cell or we do not. The IP formulation minimizes the sum of the values chosen as complementary suppressions while maintaining the confidentiality of the primary suppression within a specified tolerance level. The objective function for the IP formulation is of the form:

$$\min_{I_{ij}} c \quad \text{where } c = \sum_{i=1}^{m+1} \sum_{j=1}^{n+1} e_{ij} I_{ij}$$

where m is the number of internal rows, n is the number of internal columns, e_{ij} is the cost of the entry in row i column j of the table, and I_{ij} is 1 if e_{ij} is suppressed, and 0 otherwise. This objective is subject to the additivity and protection allowance constraints. These are not shown here, but are described fully in Sullivan and Rowe (1992).

To illustrate, suppose Table 1.1 depicts 4 products (5 rows) produced in 4 counties (5 columns) and we consider e_{ij}, the cost of suppressing the entry in row i and column j, to be the cell value.

Table 1.1. A Two-Dimensional Table

	c1	c2	c3	c4	c5
r1	1255	300	240	230	485
r2	300	50	100	100	50
r3	240	100	20	100	20
r4	230	100	100	15	15
r5	485	50	20	15	400

Suppose that the table entry in row 5, column 5 (e_{55}) is a primary suppression; i.e., e_{55} is considered too sensitive to be released. Further, suppose we want to protect e_{55} by an upper and lower protection of at least 65 units. That is, we want to prevent users from estimating the value of e_{55} any finer than the predefined range $335 < P < 465$. If the integer programming formulation is applied to this problem, cells r2/c2, r2/c5, r4/c4, r4/c5, r5/c2 and r5/c4 in Table 1.1 are chosen as complementary suppressions. The cost of this suppression scheme, according to the objective function employed in the IP formulation, is c = 50(1) + 50(1) + 50(1) + 15(1) + 15(1) + 15(1) = 195

This result is optimal. However, Kelly (1990) has shown that the cell suppression problem is NP-hard; that is, there is no known polynomial time algorithm to solve the problem with optimal results every time, and all known methods take exponential time. This implies that for large tables, as many census tables are, the IP formulation is an impractical choice.

3. THE MCF FORMULATION

Due to the unreasonable amount of time required by the IP formulation, the Census Bureau utilizes a heuristic known as the MCF program that employs network flow methodology. MCF is described in more detail in Sullivan and Zayatz (1991), Rowe (1991), and Sullivan (1992b).

Each suppression problem can be viewed as a specialized linear programming problem whose objective is to minimize the sum of the products of the cost of the cell chosen to serve as a complementary suppression and the amount of uncertainty it contributes to achieving the required protection of the primary suppression. The objective function follows:

$$\min_{\Delta_{ij},\delta_{ij}} z \quad \text{where } z = \sum_{i=1}^{m+1} \sum_{j=1}^{n+1} e_{ij}(\Delta_{ij}+\delta_{ij})$$

where m is the number of internal rows, n is the number of internal columns, e_{ij} is the entry in row i column j, and Δ_{ij} and δ_{ij}, in general, is the amount of uncertainty e_{ij} contributes to achieving the required upper or lower protection. Again, this objective is subject to the additivity and protection allowance constraints. These are not shown here, but are described fully in Sullivan and Rowe (1992).

Again suppose e_{55} in Table 1.1 is determined to be a primary suppression, and by some primary suppression rule it requires 65 units of protection. Also, suppose the value in all other table cells represents both the cost and the capacity of the cell. MCF would choose cells r2/c2, r2/c5, r3/c3, r3/c5, r4/c4, r4/c5, r5/c2, r5/c3, and r5/c4 in Table 1.1 to serve as the set of complementary suppressions. Jewett (1992) explains in more detail why MCF chooses certain complementary patterns and why MCF can oversuppress. The cost of this suppression scheme, according to the objective function employed in MCF, is: z=15(15)+15(15)+15(15)+20(20)+20(20)+20(20)+50(30)+50(30)+50(30)=6375, where the number in braces is the amount of protection the corresponding cell contributes to the amount of protection required by the primary suppression.

The reason MCF does not choose the same suppression pattern as IP is that the MCF objective function would consider the cost of that scheme to be: z=50(50)+50(50)+50(50)+15(15)+15(15)+15(15)=8175.

This shows us that MCF fails to recognize the fact that a cell is suppressed regardless of whether it contributes 1 unit to the amount of protection required by the primary suppression, or a large number of units. Although MCF calculates the cost of its suppression scheme as 6375, we only consider the sum of the values suppressed in the table to be our cost. Therefore, we would consider the cost of the MCF suppression scheme to be 255.

Comparing the complementary suppressions chosen by MCF with those chosen by the integer programming formulation, we see that MCF is guilty of oversuppression. For the above example, the total data value lost to complementary suppressions is 255 for MCF, and only 195 for the IP formulation.

4. ADJUSTMENT TECHNIQUES

We have developed two techniques to supplement the MCF program. The objective of both techniques is to remove the superfluous complementary suppressions applied by MCF. Each technique begins with using MCF to choose a set of complementary suppressions for a given primary suppression. As mentioned earlier, each primary suppression has an associated amount of required protection that in census work is set by either an n-k or p percent primary suppression rule (see Sullivan 1992a). Essentially, the protection given to a primary suppression is used to prevent data users from estimating the value of individual respondents in a cell too closely. All other cells in the table are given a cost and capacity as discussed in Section 1. Based on the costs and capacities of the cells, a set of complementary suppressions is selected by MCF. Once the set of complementary suppressions for a given primary suppression is chosen, we can apply either the cost adjustment method or the hybrid method.

4.1 The Cost Adjustment Method

When applying the cost adjustment method to a set of complementary suppressions chosen by MCF, the costs of the selected cells are adjusted to give cells that contribute little protection to the primary suppression a high cost and cells that provide a large amount of protection a low cost.

Jewett (1992) proposes that the adjusted cost of suppressing the cells already chosen by MCF to serve as complementary suppressions be based on the following: If the cell capacity of the suppressed cell is strictly less than the required protection of the primary suppression, then Adjusted Cost = Required Protection - Cell Capacity + 10. Otherwise, Adjusted Cost = 1. Therefore, the cost of the cells suppressed by MCF as complementary suppressions are adjusted as follows:

Values in R5/C2; R2/C2; R2/C5:
Adjusted Cost = 65 - 50 + 10 = 25.
Values in R5/C3; R3/C3; R3/C5:
Adjusted Cost = 65 - 20 + 10 = 55.
Values in R5/C4; R4/C4; R4/C5:
Adjusted Cost = 65 - 15 + 10 = 60.

MCF is then rerun using these adjusted costs, and excluding all previously non-suppressed cells. By performing this rerun of MCF, we are sometimes able to eliminate superfluous suppressions.

For example, after adjusting the costs and rerunning MCF, cells r2/c2, r2/c5, r3/c3, r3/c5, r5/c2, and r5/c3 in Table 1.1 are suppressed as complementary suppressions. Notice that the three cells containing the value of 15 were released ("unsuppressed") by the cost adjustment method. Therefore, the total data value lost to complementary suppressions is 210, compared to 255 for MCF used alone, and 195 for the IP method.

4.2 The Hybrid Method

When applying the hybrid method to a set of complementary suppressions chosen by MCF, a suppression tree is used to obtain all suppression cycles applied by MCF, the cost of each suppression cycle and the amount of protection each cycle contributes to the primary suppression. Sullivan and Rowe (1992) explain the hybrid method, suppression trees, and suppression cycles in greater detail.

Using information from the suppression tree to construct a set of constraints to be used in a zero-one integer programming (IP) formulation, the hybrid method attempts to refine the MCF solution by removing a subset of the complementary suppressions. The IP formulation attempts to minimize the cost associated with the complementary suppressions subject to there being sufficient protection for the primary suppression. This is accomplished by using an indicator variable, I_j, for each cycle to decide whether to suppress ($I_j = 1$) or unsuppress ($I_j = 0$) the cycle.

Applying the hybrid method to Table 1.1 when cell r5/c5 is a primary suppression would result in the suppression of cells r2/c2, r2/c5, r4/c4, r4/c5, r5/c2, and r5/c4.

Notice that the three cells containing the value of 20 were not suppressed by the hybrid method. Therefore, the total data value lost to complementary suppressions is 195 for the hybrid method compared to 210 for the cost adjustment method, 255 for MCF, and 195 for the IP method.

5. RESULTS

For a given primary suppression, MCF chooses a set of complementary suppressions. Then the cost adjustment method or the hybrid method attempts to

release some of the superfluous complementary suppressions, if any, applied by MCF for a given primary suppression. Since both adjustment methods operate on a set of complementary suppressions applied by MCF, they will never suppress more data for a single primary suppression than applied by MCF. However, there is usually more than one primary suppression per table. Since both adjustment methods minimize the value suppressed individually for each primary suppression, the overall data suppressed by an adjustment method for a table can be greater than the overall data suppressed by MCF for the same table.

Chart 1 presents the results from individually running several geographical relationships through MCF, the cost adjustment method, and the hybrid method. The number of geographical relationships tested for each data set is shown. For each method, the number of times the method came up with the "best" result is indicated. Note "best" result is determined by the method that suppressed the least amount of data value as complementary suppressions. The number of "best" results may add to more than the indicated number of geographical relationships for the data set because results may coincide for two or three methods.

Chart 1. Results from Independent Geographical Relationships

Data Set	No. of Geo. Rels	No. of geographical relationships for which the indicated method came up with the "best" result.		
		MCF	Hybrid	Cost Adj.
Retail ST 23	29	15	21	27
Retail ST 27	114	105	109	113
Retail ST 56	27	25	26	27
Serv .ST 23	29	23	21	24
Serv. ST 56	27	25	26	27

The results show that in 218 of 226 geographical relationships, the cost adjustment method either suppressed the least data, or tied with MCF or the hybrid method for suppressing the least data as complementary suppressions.

Since table cells sometimes appear in more than one table, some tables must be reprocessed several times. For instance, one table may have the state broken into its MSA parts, and the next table may take an MSA and break it into its county parts. When this occurs, we must process these two tables separately and carry information back and forth between the two. This is called backtracking.

Chart 2. Results from Relationships Requiring Backtracking

Data Set	MCF		Cost Adjustment		Hybrid	
	Value	No.	Value	No.	Value	No.
Serv. St24	27,691,666	1546	27,708,554	1532	27,846,305	1536
Serv. St56	1,041,115	302	935,379	292	935,295	293
Retail St23	5,002,021	365	4,006,538	316	4,948,502	355
Serv. St23	1,258,327	377	1,293,715	390	1,287851	391
Retail St56	1,883,890	176	1,652,408	167	1,901,210	172
Retail St27	20,567,208	1011	22,265,223	986	21,830,658	1036

Chart 2 presents the results from running the indicated data sets that required backtracking through MCF, the cost adjustment method, and the hybrid method. In this chart, value indicates the total published data value suppressed as complementary suppressions, and number indicates the number of published cells suppressed as complementary suppressions.

Comparing all three methods when backtracking is taken into account, MCF suppressed the least amount of data for three (Services--ST24, Services--ST23, Retail--ST27) of the six data sets, the cost adjustment method suppressed the least amount of data for two (Retail--ST23, Retail--ST56) of the six data sets, and the hybrid suppressed the least for one (Services--ST56) data set.

Comparing the cost adjustment method to MCF used alone, the cost adjustment method improved three of the six data sets. However, the cost adjustment method did suppress fewer cells for two data sets that had less data value suppressed using MCF.

6. BACKTRACKING

Since the cost adjustment method and the hybrid method both try to release complementary suppressions applied by MCF, they should never suppress more data for a single primary suppression than suppressed by MCF used alone. However, since cells sometimes appear in more than one table, some tables must be reprocessed several times. This is what we refer to as backtracking. For example, consider Tables 6.1 and 6.2.

Both tables contain data for three SIC's adding to a total, where more detail is provided for SIC 3. The

first table shows an MSA and it's three counties, while the second table shows county 3 further divided into its place parts. Notice that data for county 3 appear in both tables. Since it is theoretically impossible to create a network to capture the hierarchical structure of the SIC's and geographical regions (and, if it was possible, we would more than likely run into computer storage constraints) we must process these two tables separately and carry information (which cells are suppressed) back and forth between the two. Due to this phenomenon we call backtracking, both the cost adjustment method and the hybrid method may suppress more data than MCF used alone.

For example, suppose the value in SIC 1, county 1 is a primary disclosure that requires 100 units of protection. MCF would process Table 6.1 and choose cells sic2/cnty1, sic1/cnty2, sic1/cnty3, sic2/cnty2, and sic2/cnty3 as complementary suppressions. With this suppression pattern, the primary is protected by the required 100 units. We would then process Table 6.2, carrying over the complementary suppressions in county 3 for SIC 1 and SIC 2. These two suppression would need to be protected by 70 units each since they provided 70 units of protection in Table 6.1. Using MCF on Table 6.2, cells sic1/place1, sic1/place2, sic2/place1, and sic2 /place2 would be suppressed as complementary suppressions. The two complementary suppressions in county 3 are now protected by 70 units as required. At this point, we are done; we do not have to go back and process Table 6.1 again.

Now let us see what the cost adjustment and the hybrid method would do for these two tables. Immediately after the complementary suppressions are chosen for the primary suppression in the Table 6.1, and before we process the Table 6.2, we try to release some suppressions applied by MCF. The application of either the cost adjustment method or the hybrid method would result in the suppression of cells sic1/cnty3, sic2/cnty1, and sic2/cnty3. We may look at this table and claim we suppressed less data than MCF suppressed. But realistically we cannot make this claim until we process Table 6.2 since there is overlap between the two.

In Table 6.2, we must now protect the two complementary suppressions contained in county 3 by 100 units each. We first apply MCF to the suppression in sic1/cnty3 and immediately apply either the cost adjustment method or the hybrid method. This results in the suppression of cells sic1/place1, sic3/cnty3, sic3/place1, sic31/cnty3, and sic31/place1. We must then apply MCF to the complementary suppression in sic2/cnty3 that we carried over from the Table 6.1. Applying MCF and immediately applying either the cost adjustment method or the hybrid method would result

Table 6.1: MSA Part of Two Dimensional Table

	MSA 1	Cnty 1	Cnty 2	Cnty 3
Total	52,500	1,200	1,060	50,240
SIC 1	270	P	30	140
SIC 2	230	100	30	100
SIC 3	52,000	1,000	1,000	50,000
SIC 31	26,000	500	500	25,000
SIC 32	26,000	500	500	25,000

Table 6.2: County Part of Two Dimensional Table

	County 3	Place 1	Place 2
TOTAL	50,240	40,130	10,110
SIC 1	140	100	40
SIC 2	100	30	70
SIC 3	50,000	40,000	10,000
SIC 31	25,000	20,000	5,000
SIC 32	25,000	20,000	5,000

in the suppression of cells sic2/place1, sic1/place2, and sic2/place2.

Unfortunately, while processing the second table we suppressed cells sic3/cnty3 and sic31/cnty3. These two cells also appear in Table 6.1. We must carry these two suppressions back to the first table and ensure they are protected. We first apply MCF to the suppression in cell sic3/cnty3 in Table 6.1, and apply either the cost adjustment method or the hybrid method to see if any suppressions can be released. Then we apply MCF to find complementary suppressions for the suppression in sic31/cnty3, and again we apply either the cost adjustment method or the hybrid method to see if we can improve upon the MCF solution. When the cost adjustment or the hybrid method is completed on Tables 6.1 and 6.2 cells sic1/cnty3, sic2/cnty1, sic2/cnty3, sic3/cnty1, sic3/cnty3, sic31/cnty1, sic31/cnty3, sic1/place1, sic1/place2, sic2/place1, sic2/place2, sic3/place1, and sic31/place1 are suppressed as complementary suppressions. Now both tables have more data value suppressed than if we had not attempted to reduce the number of complementary suppressions chosen to protect the original primary suppression in county 1 for SIC 1. Therefore, sometimes by trying to improve our solution for one primary suppression, our overall results are worsened.

7. CONCLUSION

Although both methods discussed above approach the problem by working with an initial solution provided by MCF, they are quite different. The cost adjustment method examines the set of complementary suppressions applied by MCF for a particular primary suppression and finds new closed paths (cycles) from the set to protect the primary suppression. Any previously applied complementary suppressions not contained in the new closed paths are released from the suppression scheme. However, this could still cause oversuppression. The hybrid method never reconfigures the closed paths used to protect the primary suppression. It determines which, if any, closed paths can be removed from the suppression scheme. It is important to note that the hybrid method can suppress less data than the cost adjustment method and vice versa.

As stated earlier, the cost adjustment method and the hybrid method will never suppress more data than suppressed by MCF for a single primary suppression since they both work with a subset of data provided by MCF. However, there is usually more than one primary suppression per table and many tables overlap (equivalent cells appear in more than one table). This can cause the cost adjustment method and the hybrid method to suppress more (or less) data than MCF for a given data set.

Based on the results shown in Section 5, I recommend using the cost adjustment method for processing the economic censuses. When backtracking is taken into account, the cost adjustment method suppressed less total data value than MCF for three of six data sets, and suppressed fewer data cells for an additional two data sets. However, since there does not seem to be a notable difference between the results for MCF and the cost adjustment method, I recommend that if CPU time is ever a concern that MCF be used alone instead of the cost adjustment method.

ENDNOTE:

1. Six sets of data were obtained from the 1987 censuses. They are as follows: (1) Services data for Maryland, Wyoming, and Maine, and (2) Retail data for Maine, Minnesota, and Wyoming.

REFERENCES

Jewett, Robert (1992), "Disclosure Analysis for the 1992 Economic Census," US Bureau of the Census, Economic Statistical Methods Division, working paper.

Kelly, J.P. (1990) "Confidentiality Protection in Two and Three-Dimensional Tables," Ph.D. Dissertation, University of Maryland, College Park, Maryland 20742.

Kelly, J.P., Golden, B.L., and Assad, A.A. (1991), "Cell Suppression: Protection for Sensitive Tabular Data," Working Paper Series MS/S 91-014, College of Business and Management, University of Maryland, College Park, Maryland 20742.

Rowe, Errol (1991), "Some Considerations in the Use of Linear Networks to Suppress Tabular Data," American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods, pp. 357-362.

Sullivan, Colleen M. (1992a), "An Overview of Disclosure Principles," SRD Research Report Series, NO. RR-92/09, Bureau of the Census, Statistical Research Division, Washington, D.C.

Sullivan, Colleen M. (1992b), "The Fundamental Principles of a Network Flow Disclosure Avoidance System," SRD Research Report Series, No. RR-92/10, Bureau of the Census, Statistical Research Division, Washington, D.C.

Sullivan, Colleen M. and Errol Rowe (1992), "A Data Structure and Integer Programming Technique to Facilitate Cell Suppression Strategies," American Statistical Association, 1992 Proceedings of the Section on Survey Research Methods.

Sullivan, Colleen M. and Laura Zayatz (1991), "A Network Flow Disclosure Avoidance System Applied to the Census of Agriculture," American Statistical Association, 1991 Proceedings of the Section on Survey Research Methods.

University of Texas at Austin, "MCF PROGRAM DOCUMENTATION: Minimum Cost Flow Optimization Software," Center for Business Decision Analysis, College and Graduate School of Business Administration, Austin, Texas.

* This paper reports the general results of research undertaken by the Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

A long version of this paper, ESMD Report Series ESMD-9301 "A Comparison of Cell Suppression Methods" , is available upon request from the author.

STRATEGY FOR PUBLISHING THE PARTIAL HERFINDAHL-HIRSCHMANN INDEX

Richard W. Graham, Bureau of the Census
FOB #3, Room 2682, Washington, DC

KEY WORDS: Disclosure avoidance, statistical disclosure, Herfindahl-Hirschmann Index

I. INTRODUCTION

The Census Bureau receives many requests from various sponsors for special tabulations of the quinquennial economic censuses. While every attempt is made to accommodate these requests, the Census Bureau strictly adheres to its standard publication procedures, including compliance with the confidentiality rules as mandated by Title 13, United States Code, section 9. As interpreted, the Census Bureau includes statistical nondisclosure as a point of confidentiality.

The Census Bureau has received several requests from the Food Policy Marketing Center (FPMC) at the University of Connecticut for data pertaining to grocery stores, Standard Industrial Classification (SIC) 5411. Specifically, these requests have focused on measures of market share and concentration by metropolitan area, including the Herfindahl-Hirschmann Index (HHI). Until recently, all such requests for HHI data have been denied because the risk of statistical disclosure (deriving close estimates) of confidential data was deemed unacceptable. Census Bureau policy prescribed that the release of HHI data along with other concentration measures posed too great a risk because users could derive close estimates of individual respondent data - clearly a violation of both Census Bureau policy and Title 13.

In Census Bureau publications, concentration measures generally are based on sales, revenues, or value of shipments.[1] These concentration ratios, as they are called in Census Bureau publications, provide the component shares of total economic activity accounted for by the largest firms within a given published cell. For example, the concentration ratio for the eight largest firms in a given published cell, denoted C_8, represents the proportion of total economic activity accounted for by those eight largest firms.

[1] Sales will be used throughout the paper as a general term for sales, revenues, or value of shipments.

The Herfindahl-Hirschmann Index is a rather simple index to calculate but a more complex statistic to interpret. The index is calculated by summing the squares of the individual market shares (concentration ratios) of the participants (firms) within a given published cell (Hirschmann, 1945; Herfindahl, 1950; and Department of Justice and Federal Trade Commission, 1992). Thus, H_8 is the sum of the squares of the market shares of the eight largest firms.

Interpreting the HHI is somewhat less straightforward. Once understood, however, it does provide much more meaningful information about the market and its participants than the pure concentration ratios. The HHI provides both insight into the distribution of market shares of the participants used in the calculations and also the composition of the market outside of the participants used in the calculations (Justice, 1992).

While publication of the Herfindahl-Hirschmann Index in and of itself poses no disclosure risks or violations of Census Bureau disclosure rules, standardizing the HHI (Greenberg, 1993) may allow close estimates of individual respondent data to be derived. The Census Bureau views the derivation of close estimates to be a disclosure risk and a breach of confidentiality pursuant to Title 13.

Although the Census Bureau has in the past published the Herfindahl-Hirschmann Index as part of its standard publications (Census of Manufactures, 1982 and 1987), risk of disclosure was greatly reduced or nonexistent. The HHI data in these publications were presented for the 50 largest firms or the cell total for the United States, whichever was smaller. Census Bureau policy is such that where a sufficient number of firms exist the threat of disclosure is significantly diminished or nonexistent when the largest firms do not account for the major proportion of the published cell.

The requests by the FPMC for special tabulations included sales totals, concentration ratios, and Herfindahl-Hirschmann Indices for the four, eight, and twenty largest firms and the entire market (published cell total) by metropolitan area. For the special tabulations for 1977 and 1982, the Census Bureau denied publication of the HHI because the risk of

disclosure was deemed too great when provided with other measures of concentration (Monsour, 1980). The Census Bureau did, however, provide the other requested data subject to normal disclosure rules.

In 1989, the FPMC once again made its special tabulation request for sales totals, concentration ratios, and Herfindahl-Hirschmann Indices by metropolitan area for the cell total and the four, eight, and twenty largest firms. Because of these repeated requests for the HHI data, the Census Bureau revisited the HHI and confidentiality issue. The objective of this paper is to present the Census Bureau's methodology for applying the theoretical foundation developed in Greenberg and develop a criteria by which the HHI data may be published.

As in Greenberg, sales is the measure by which the HHI is calculated over a given geographic area (published cell). To maintain continuity, the following notation is used:

S_T	total sales for a cell having N participants
S_i	sales for the i^{th} largest firm in the cell for $1 \leq i \leq N$
$s_i = S_i/S_T$	market (cell) share of firm i
$V_K = \sum_{i=1}^{K} S_i$	total sales of the K largest firms for $1 \leq K \leq N$
$C_K = \sum_{i=1}^{K} s_i$	concentration ratio for the K largest firms for $1 \leq K \leq N$
$H_K = \sum_{i=1}^{K} s_i^2$	Herfindahl-Hirschmann Index for the K largest firms for $1 \leq K \leq N$
$G_K = H_K/(C_K)^2$	defined as the relative Herfindahl-Hirschmann Index for the K largest firms for $1 \leq K \leq N$.

In his paper, Greenberg develops a theory of risk based on an analysis of G_K by asking: given G_K, how closely can one estimate the value (sales) of a cell respondent? Greenberg answers this question by developing formulas and expressing his findings as both maximum and continuous minimum functions of q_1,[2] where $q_1 = S_1 / V_4 = s_1 / C_4$, in terms of a substitute parameter, b, where $b = G_K$.

From Greenberg, the maximum function for q_1, M, expressed in terms of G_K is:

$$M = (1/K)[1 + (K-1)[(KG_K-1)/(K-1)]^{1/2}] \qquad 1/K \leq G_K \leq 1$$

and the continuous minimum function for q_1, m, expressed in terms of G_K is:

$$m = \begin{cases} (1/K)[1 + [(KG_K-1)/(K-1)]^{1/2}] & 1/K \leq G_K \leq 1/(K-1) \\ \vdots & \vdots \\ [1/(K-J)][1 + [((K-1)G_K - 1)/(K-J-1)]^{1/2}] & 1/(K-J) \leq G_K \leq 1/(K-J-1) \text{ [for } 0 \leq J \leq K-2] \\ \vdots & \vdots \\ 1/2[1 + (2G_K-1)^{1/2}] & 1/2 \leq G_K \leq 1. \end{cases}$$

As noted in Greenberg, when G_K approaches either of the tails in the distribution (i.e., 1/K in the lower tail or 1 in the upper tail) the risk of statistical disclosure is unacceptable. In the case of G_4, the relative HHI for the four largest firms, the range is 1/4 to 1.

What Greenberg does not address, however, is at what points in the distribution does the G_K value present the risk of disclosure? As G_K moves closer to the midpoint from either tail, when does G_K become acceptable?

In Section II, the minimum and maximum functions, as developed by Greenberg, are applied to arbitrary disclosure rules to illustrate the methodology used to release the Herfindahl-Hirschmann Index data.

II. GENERAL DISCLOSURE POLICY FOR THE RELATIVE HERFINDAHL-HIRSCHMANN INDEX

Census Bureau policy prevents the publication of its official disclosure rules. Thus for purposes of this paper we make the arbitrary assumption that close estimates are defined as estimates within +/-10 percent of the true value of the respondent data for a given cell.

Although Census Bureau plans were to publish only sales totals, concentration ratios, and HHI data for the total cell, and the four, eight, and twenty largest firms; the disclosure analysis procedures required

[2]Greenberg uses q_1 since typically sales of the largest firm is the most visible target for which primary interest lies. We note, however, that the analysis applies for each q_i in the cell.

analysis beyond the relative HHI (G_K) data for the published figures.

As Greenberg points out, by subtracting out its own known sales value, any firm within a given cell could derive G_{K-1} and utilize the analysis in Greenberg to potentially derive a close estimate of the respondent data. In the case of G_4, any firm knowing its own sales and knowing itself to be one of the four largest firms, could subtract out its sales and derive G_3 and hence estimates for sales, S_1, of the largest firm. Of course, completing the analysis of G_4 also could produce close estimates for sales of the largest firm; however, the fewer the number of firms the greater the risk of developing close estimates.

A. Acceptable ranges through indirect release of G_4.

Using our stated assumption that estimates within +/- 10 percent are unacceptable, we begin our analysis by viewing Figure 1. (Figure 4. in Greenberg). The horizontal (X) scale measures the relative HHI (G_4), and the vertical (Y) scale measures the difference (d) between the maximum and minimum estimates of q_1, M - m. Our assumption then follows that any point on the graph below .1, or 10 percent, clearly falls in the tails of the distribution and presents an obvious risk of statistical disclosure for that cell. Hence, publishing the HHI for the particular cell would be a clear violation of the disclosure rules.

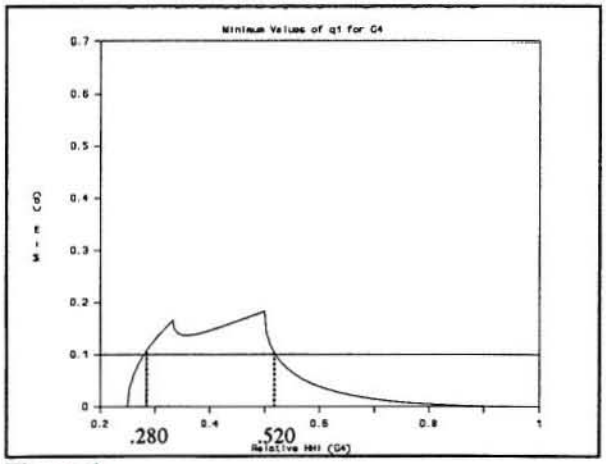


Figure 1

From Figure 1., it is clear that only a small range exists over which the potential for publishing the HHI may be acceptable. For the relative HHI G_4, the range where the difference equals or exceeds 10 percent is .280 to .520. When G_4 falls within this range the potential exists for publishing H_4. But because any one of the participants in the cell could derive G_3 from the published data through substraction of its own sales, the analysis must also be conducted on G_3 before H_4 may be published.

Below are four examples which show the values for C_4, H_4, G_4, the minimum and maximum computed values of q_1, and d.

Example #1:

$C_4 = .893$ $H_4 = .4861$ $G_4 = .610$

Max $q_1 = .770$ Min $q_1 = .735$

$d = M - m = .035$

Example #2:

$C_4 = .722$ $H_4 = .1731$ $G_4 = .332$

Max $q_1 = .498$ Min $q_1 = .333$

$d = .165$

Example #3:

$C_4 = .668$ $H_4 = .1945$ $G_4 = .436$

Max $q_1 = .623$ Min $q_1 = .464$

$d = .159$

Example #4:

$C_4 = .546$ $H_4 = .0759$ $G_4 = .255$

Max $q_1 = .311$ Min $q_1 = .270$

$d = .041$

Examples 1 and 4 clearly illustrate that to publish H_4, undesirable results are expected as G_4 falls within either tail of the distribution, example 1 in the upper tail and example 4 in the lower. The derived value for d in each of these examples is well below the minimum 10 percent level. One could derive estimates for q_1 in example 1 within +/- 3.5 percent and hence very close estimates of sales for the largest firm, S_1.

Conversely, examples 2 and 3 illustrate that publishing H_4 would not present the risk of allowing users to compute estimates of q_1 within the prescribed 10 percent level. In fact, example 2 data only allow

estimates within a margin of error of 16.5 percent. Thus, given V_4 of 100,000, one could estimate sales, S_1, of the largest firm to be in the range:

$$\min q_1 V_4 \leq S_1 \leq \max q_1 V_4$$

$$33300 \leq S_1 \leq 49800$$

clearly a large range and one which does not compromise the assumed disclosure rules.

As previously mentioned, the analysis cannot stop here as any one of the four largest firms could subtract out their sales and derive G_3 and minimum and maximum estimates for q_1. Before publishing H_4, analysis of G_3 must be conducted to determine acceptable publication ranges.

B. Acceptable ranges through indirect release of G_3.

Figure 2. (Figure 2. in Greenberg) presents the same graphic as Figure 1. except the difference measured is the maximum and minimum estimates of q_1 using G_3. Again, pursuing our assumption, any point below .1 on the Y scale clearly represents a value on the X scale in the tail regions of the distribution. The acceptable range for the relative HHI,G_3, is .393 to .512. When derived values for G_3 fall within this acceptance range, the HHI, H_4, data are publishable for the given cell.

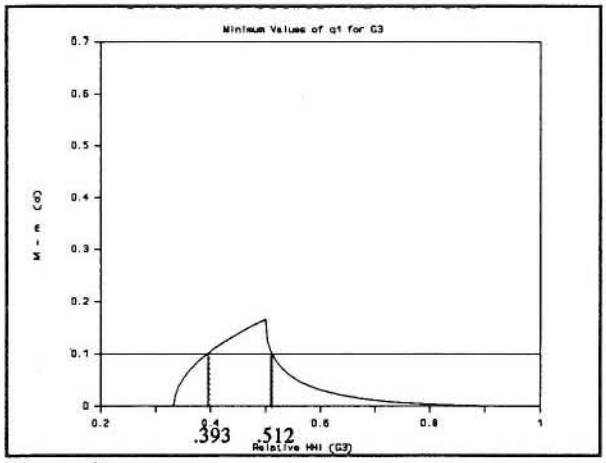


Figure 2

Continuing with examples 2 and 3 above, the following presents the analysis for G_3. We also present an example to illustrate unpublishable data based on G_3 even when G_4 falls within the acceptance range.

Example #5:

$C_4 = .722$ $H_4 = .1731$ $G_4 = .332$

Assume: $V_4 = 100{,}000$ $S_T = 138{,}500$

$S_1 = 49{,}500$ $s_1 = .357$
$S_2 = 20{,}500$ $s_2 = .148$
$S_3 = 16{,}000$ $s_3 = .116$
$S_4 = 14{,}000$ $s_4 = .101$

Since any firm potentially could derive G_3 by subtracting its own sales from V_4, we continue the analysis using the same assumption as Greenberg, where the fourth largest firm removes its sales. However, the analysis holds for any of the firms.

Then, $C_3 = C_4 - s_4 = .621$ $H_3 = H_4 - s_4^2 = .1629$
$G_3 = .422$

Max $q_1 = .577$ Min $q_1 = .455$

$d = .122$

Example #6:

$C_4 = .668$ $H_4 = .1945$ $G_4 = .436$

Assume: $V_4 = 100{,}000$ $S_T = 145{,}300$

$S_1 = 59{,}500$ $s_1 = .409$
$S_2 = 16{,}550$ $s_2 = .114$
$S_3 = 13{,}500$ $s_3 = .093$
$S_4 = 10{,}450$ $s_4 = .072$

Then, $C_3 = .616$ $H_3 = .1893$ $G_3 = .499$

Max $q_1 = .666$ Min $q_1 = .499$

$d = .167$

Example #7:

$C_4 = .798$ $H_4 = .1944$ $G_4 = .305$

Assume: $V_4 = 100{,}000$ $S_T = 125{,}300$

$S_1 = 40{,}000$ $s_1 = .319$
$S_2 = 32{,}750$ $s_2 = .261$
$S_3 = 15{,}750$ $s_3 = .126$
$S_4 = 11{,}500$ $s_4 = .092$

Then, $C_3 = .706$ $H_3 = .1860$ $G_3 = .373$

Max q_1 = .496 Min q_1 = .415

d = .081

In examples 5 and 6, H_4 is publishable because the difference between the minimum and maximum estimates of q_1 given G_3 is sufficiently large to prevent the derivation of close estimates of the sales of the largest firm, S_1. Thus, the rule that both G_4 and G_3 fall within their respective acceptance ranges is met for these two examples and H_4 is publishable as the actual computed value.

Conversely, example 7 shows that while based on G_4, the H_4 data would be publishable; however, the G_3 data suggest otherwise. Since G_3 of .373 is well below the lower acceptance range limit of .393, a disclosure risk results. Therefore, publishing H_4 for this example would violate disclosure rules by allowing users to develop close estimates of individual respondent data.

In the next section we develop the rules for what is publishable when the computed values for G_4 or G_3 fall outside the acceptance ranges.

C. Other disclosure risks and considerations.

To publish H_4 when G_3 or G_4 is outside the acceptance range would clearly violate Census Bureau disclosure rules. What then could be published in those instances where either G_3 or G_4 are outside the acceptance ranges?

To maximize data content while at the same time minimizing risk of disclosure, we decided on publishing the HHI data as a range in these instances. By doing so this introduced a new threat of disclosure as one easily could derive Census Bureau disclosure rules. To illustrate this point, we return to example 1 from section II.A.

The actual H_4 data are unpublishable for this example as G_4 falls outside the acceptance range, .280 to .520. And since publishing a range for H_4 was more desirable than publishing nothing at all, the range would have to be published as greater than the H_4 value which corresponds to G_4 of .520, since .610 > .520. This presented an undesirable situation, however, since always publishing > H_4 which corresponds to G_4 of .520 would reveal the assumed disclosure policy of 10 percent. Thus, we built into our disclosure model a random multiplier which prevents determination of the actual cutoff. Again we use for illustrative purposes a random range multiplier of up to 10 percent. Thus, in this case we would randomly select a multiplier between .900 and 1.000 and derive a new value for H_4 to present as a range.

Example #8:

C_4 = .893 H_4 = .4861 G_4 = .610

Since .610 > .520,

then publish H_4 as > .520 * $(.893)^2$ * .95 or
H_4 > .3939

where .95 = random multiplier between .900 and 1.000.

Thus, the publication rule, based on the 10 percent criterion, for publishing H_4 when G_4 or G_3 are outside the acceptance range is as follows:

when G_4 < .280 or G_3 < .393,

then publish $H_4 < .280 * C_4^2 * X$

where X = random multiplier

and

when G_4 > .520 or G_3 > .512,

then publish $H_4 > .520 * C_4^2 * Z$

where Z = random multiplier.

D. Methodology applied to the release of H_8 and H_{20}.

The G_4 and G_3 analysis above can be generalized to G_K and G_{K-1}. Hence, applying the analysis to G_8 and G_{20} is rather straightforward. However, because users can subtract intra-cell (e.g., V_4, C_4) data from other intra-cell (e.g., V_8, C_8) data and derive NG_4 and NG_3, the analysis also must be completed for G_{K-L} and G_{K-1-L}, where L < K and corresponds to the previous intra-cell data. In other words, if analyzing for H_8, L = 4.

By subtracting the four largest firms' data from the eight largest firms' data, one can easily apply the G_4 and G_3 analysis to the remaining four firms' data. That is, the fifth largest firm now becomes the largest firm for that group, the sixth the second, and so on. By applying these same disclosure avoidance procedures to the remaining data, the Census Bureau can ensure confidentiality for the fifth through the eighth largest firms. But the analysis on NG_4 and NG_3 is not

sufficient. The analysis also must be conducted on G_8 and G_7. Of course, the ranges are different for G_8 and G_7, but the methodology remains consistent.

Thus, publication rules for publishing H_8 are as follows:

if G_8, G_7, NG_4, and NG_3 are all within the acceptance ranges,

where $NG_4 = (H_8 - H_4) / (C_8 - C_4)^2$

$$NG_3 = (H_8 - H_4 - s_8^2) / (C_8 - C_4 - s_8)^2,$$

then publish H_8 as the actual computed value.

Otherwise, if any of G_8, G_7, NG_4, or NG_3 are outside the acceptance ranges, then:

for computed G_K values in the lower tail of the distribution, present data as $H_8 < L_R * C_8^2 * X$

where L_R = lower range value for G_8
X = random multiplier,

and for computed G_K values in the upper tail of the distribution, present data as $H_8 > U_R * C_8^2 * Z$

where U_R = upper range value for G_8
Z = random multiplier.

Like the analysis for H_8, for H_{20} the analysis must be applied to G_{20} and G_{19}. Further, to prevent disclosure of the ninth through the twentieth largest firms' data, the eight largest firms' data must be subtracted and the analysis conducted for G_{12} and G_{11}. Thus, it should be clear that our disclosure rule for H_{20} is that each of G_{20}, G_{19}, G_{12}, and G_{11} all must be within their respective acceptance ranges before H_{20} can be published. Otherwise, H_{20} must be published as a range.

III. CONCLUSION

In this paper, we have applied the theoretical model described in (Greenberg, 1993) to empirical data and developed a disclosure release strategy which maximizes the amount of publishable data while minimizing the risk of disclosure of individual respondent data.

To this end, the special tabulation for the Food Policy Marketing Center of the University of Connecticut was a success for both the Census Bureau and the FPMC. The Census Bureau, because it is now able to provide data which in the past has been deemed unpublishable, and the FPMC in that it now may obtain previously unavailable data for use in its studies of the grocery stores and supermarkets industry.

To summarize then, the Census Bureau was able to provide to the FPMC the following data where disclosure rules otherwise did not prevent the release of the data: number of firms, number of establishments, total sales for the cell total by metropolitan area, and sales, concentration ratio, and the partial Herfindahl-Hirschmann index for the four, eight, and twenty largest firms in each metropolitan area. The HHI data were published as actual values when all associated G_K computed values fell within their prescribed acceptance ranges. Otherwise, the HHI data were published as ranges.

It also should be noted that the foregoing analysis may be applied not only to grocery stores, but to any industry for which data are collected on a firm-by-firm basis. Hence, the Census Bureau could provide the HHI data as part of its standard publication series.

References

Bureau of the Census (1992), "Concentration Ratios in Manufacturing," 1987 Census of Manufactures, Publication MC87-S-6, U.S. Department of Commerce.

Title 13, United States Code, Section 9 (1954).

Department of Justice and Federal Trade Commission (1992), "Horizontal Merger Guidelines."

Greenberg, B.V. (1993), "Characterizing the Disclosure Risk for the Hirschmann-Herfindahl Index," Proceeding of the International Conference of Business Surveys.

Herfindahl, O. C. (1950), "Concentration in the Steel Industry," Unpublished Ph.D. dissertation, Columbia University.

Hirschmann, A. O. (1945), "National Power of Structures of Foreign Trade," University of California Press, Berkeley, CA.

Monsour, N.J. (1980), "Concerning Herfindahl Indices and Confidentiality," Internal Census Bureau Memorandum.

METADATA FOR ESTABLISHMENT SURVEYS: CORPORATION SOI

Martin H. David, John L. Czajka

Martin David, Economics, University of Wisconsin, 1180 Observatory Dr., Madison WI 53706

KEY WORDS: Documentation, aggregates, time series

The ongoing collection of a sample of tax returns from corporations (Csoi) is a major source of information about corporate enterprises. The sample is produced by the Statistics of Income Division/IRS. The sample has been used extensively as microdata by the Office of Tax Analysis, US Treasury and the Joint Committee on Taxation of the Congress.

IRS/SOI has published aggregated data from Csoi as the *Source Book of Corporation Income Tax Returns* (SCsoi) for more than thirty years. The SCsoi is used by tax policy analysts, the Bureau of Economic Analysis of the Department of Commerce, and public accountants. The SOI Division seeks to increase statistical use of SCsoi. To assist users of the SCsoi, the data product and documentation, metadata, should fulfill several objectives: a) enable the analyst to calculate unbiased statistics and measures of their error; b) minimize disclosure risk, and c) direct the analyst toward appropriate interpretation of the measures available. These simple objectives appear self-evident, yet they have not been implemented in an integrated way. The purpose of this paper is to relate the problems of documenting these aggregate statistics to a generic theory of metadata and provide some examples of the benefits of more systematic metadata.

Compilation of the SCsoi in machine-readable form for the period 1966-1987 highlights the need for metadata systems. The data product lends itself to study through conventional statistical packages, a capability not present in print data. Statistical analysis of the aggregates is perilous because the data do not conform to widely used assumptions in standard software, namely i) identically and independently distributed errors, ii) absence of truncation on particular attributes, and iii) sampling by known exogenous rates. All of these assumptions can be violated. Our analysis will show how users of SCsoi can be better informed through the use of metadata, information about the data being analyzed.

Sections 1-2 describe some sources of difficulties in interpreting data from SCsoi and the Csoi. Sections 3-4 survey the datasets and documentation in use. Section 5 sketches desiderata for metadata systems. Implications for aggregated data appear in Section 6.

1. Interpreting a cross-section of corporate tax returns

Public use data from corporate tax returns take the form of tabulations. The SCsoi contains aggregates classified by Enterprise Standard Industrial Classification (ESIC) and by size of asset classes. (The sample, first introduced in 1951, is stratified on those classes, assuring stability of the aggregates presented.) The conceptually complex content of SCsoi creates a substantial risk of misinterpretation by tax policy analysts.

Heterogeneity. The population of tax returns included in the universe pertains to entities with substantially different treatment under the tax law, ranging from the Sub-chapter S corporations who have no liability for corporate income taxes, to corporations controlled by foreign entities, to the insurance industry whose "cost" of business includes large reserves for paying conditional liabilities, to the simple fabricator that one might imagine as a typical corporate entity.

Aggregation. Because the data are presented as aggregates, understanding covariation among the variables tabulated is foreclosed, and estimation of moments of the underlying data is not possible. This tends to focus attention on differences in the mean aggregate. Those differences can not be tested statistically. For example, an analyst might be interested in the relationship between retained earnings (r_{kis}) and net income less tax due (x_{kis}) where i indexes industries, s indexes size of asset classes, and k indexes enterprises within each size and industry class. The SCsoi reports the set of cell entries $\{ r_{.is} , x_{.is} \}$ where the "." subscript represents an aggregate over all k_{is} enterprises within the cell. Despite apparent regularities over size classes the correlation of (r_{kis} , x_{kis}) could be extremely low.

Entities tabulated. Interpretation of the data is further complicated by the fact that the entities tabulated are themselves aggregates. Tax law permits the filing of a consolidated return covering subsidiaries within a corporate family. Because consolidation can be elected by the taxpayer (just as married individuals can elect to file a joint return), the number and structure of consolidated returns depends on strategies for minimizing taxes. Where separate filing is perceived as tax-minimizing, each return will represent one

corporation, and the ownership structure of the corporation does not influence the number of entities. Otherwise, the number of entities reflects the reduction in taxes possible by netting the dividends paid within the corporate family and combining net income from several corporations.

The ESIC depends on the principal products produced within the consolidated family. Clearly, that classification varies with the extent of consolidation of returns for related corporate entities.

The universe and accounting period. The scope of returns covered by the sample also confounds simplistic interpretation. Corporation returns for 1992 may include business activity that relates to as much as six months of activity in 1991, or up to six months of activity in 1993. This situation is further complicated by the existence of part-year returns. Such returns arise for entities that change accounting years, begin, or cease operation. (The frequency of part-year returns increased dramatically after 1985 as firms adopted calendar-year accounting periods.) While all returns serve as a record of income of legal entities, it is often more interesting to tabulate income of full-year returns or annual rates of income for all returns to gain perspective on the financial flows.

Some corporations delay the filing of a return beyond the normal date. For that reason returns of some corporations are absent. Missing returns in the current year are imputed for the Csoi, reducing the bias in estimates of means and aggregates but also contributing to a shrinkage of conventional variance estimates (Little and Rubin 1987).

To minimize the problem of delayed returns, the business master file is sampled for 24 months. Thus returns for accounting periods ending in the twelve months before 30 June 1991 are sampled from 1 July 1990 until 30 June 1992. Imputations of rows are restricted to "critical cases", very large firms within a particular ESIC. The prior year is then substituted for the missing return, less than 20 firms in recent years. Imputed returns are idiosyncratic. Delayed filing is more probable for complex corporate entities and entities undergoing reorganization.

2. Interpreting repeated cross-sections dynamically

Repeated cross-sections of the SCsoi can be interpreted in several ways. The most obvious interpretation ignores temporal correlation in the sample and interprets the information as change in particular economic niches (i.e. minor industry groups). Alternatively, if the population of enterprises in a particular cell is stationary, change can be interpreted as comparative statics related to other changes in the economy and economic policy. The most questionable interpretation is to treat the data as a synthetic cohort and assume that some information about the dynamics of firms can be inferred from the aggregates. Each of these interpretations is clouded by dynamics in the statistical design, the tax law, and interactions between the tax law and enterprise activities.

Real reclassification. The ESIC of the enterprise is determined by the ESIC of the business activity generating the largest volume of sales. Firms whose two largest activities generated expected sales of similar magnitude can be classified in different ESIC categories in successive years. E.g., a firm engaged in paper-making and printing activities that generate the same level of sales on average, can be classified in either printing or paper in any year. Because of year-to-year change in the assigned ESIC, a small decline in the dominant sales activity can result in a large apparent reduction, as all sales in that ESIC class moved to another ESIC.

Dynamics of the tax code. The Internal Revenue Code has been amended with increasing frequency since 1976 (Witte 1991); since 1981 substantial alterations in the definition of corporate net income have been made biennially and "technical corrections" to those major legislative changes in intervening years have occasionally had significant effects in particular industries. As a result the relation of tax constructs, such as corporate net income, to underlying economic values is likely to shift from year-to-year. Interpreting change in means of tax constructs as changes in mean economic values is likely to be flawed. The consequences of tax law for corporations with net income differ from the consequences for those with net loss (Altschuler 1988); differences arise between firms with little inventory and those with large inventories (Feldstein and Summers 1979), between firms with reserves for contingent liabilities and those with none, and so forth. For all of these reasons change in magnitudes on tax returns requires differential interpretation conditioned on logic embedded in the tax law (and resulting entries on the return).

Elective consolidation of returns. The ability of the enterprise to alter the scope of its consolidated return also affects the meaning of changed levels of means, quite apart from the change in tax law. Endogenous shifts in the extent of consolidation of returns within corporate families will lead to different numbers of entities being counted, so that means will change, even when the aggregates are constant.

Change in entity. Merger, acquisition, and spin-off

of activities within the corporate family further complicate interpretation of year-to-year change in aggregates. The size of assets classification and the ESIC may change as a result of these legal reorganizations, though underlying economic activities need not change. An enterprise with 40% of its sales in paper-making and 60% in printing services would be reclassified from printing to paper-making, if it acquired a paper mill whose sales exceed 20% of current sales. (When a larger entity absorbs a smaller entity, sales will be reclassified to the ESIC dominating the activities of the merged firm, but a smaller dollar value moves across the industry boundary.) The effects of legal reorganizations are clearly the greatest in tabulated cells that have few entities, particularly in those ESIC in which economic activity is highly concentrated.

Change in the ESIC. Alterations in the ESIC create discontinuities in the membership in particular cells, even when business organization is unchanged.

3. Computational capabilities

Concatenated SCsoi. In 1991 the Office of Tax Policy Research at the School of Business of the University of Michigan released the aggregate data in SCsoi for tax years from 1966 to 1987 in machine-readable form. Similar entries on the tax return were included under a single variable label and the repeated cross-section samples were concatenated to facilitate study of trends. The documentation included with the release explains change in the ESIC that occurred during the period.

Panel data in use. Users of the SCsoi are aware that large corporate entities are likely to occupy particular asset and ESIC classes, because data on those corporations is available through public records (the Securities Exchange Commission's Form 10-K), private data collection compilations (Dun and Bradstreet, Standard and Poors), and microdata (Standard and Poors COMPUSTAT).

Therefore, aggregates available in the SCsoi refer to particular groups of large firms, and changes in those aggregates may be interpreted (erroneously) to reflect change in the tax picture for fixed portions of an industry. Because of disclosure limitation, year-to-year change in class membership, and elective changes in the composition of the filing entity, it is clear that interpretation of change as a trend pertaining to a fixed population is in error. However, it is not always in error, and this possibility clearly creates a need for appropriate metadata.

4. The state of documentation

Background. The IRS/SOI (1992, 9-17) *1989 — Corporate Income Tax Returns, Publication 16* contains an excellent overview of the scientific design for the data collection. An overview is provided under nine headings: Background, population, sample design, sample selection, data capture, data cleaning, data completion, estimation, data limitations and measures of variability. Two tables are provided to show sampling rates and approximate coefficients of variation according to the number of returns represented. Samples of forms are also included.

SCsoi diskettes. The diskette data issued by the Office of Tax Policy Research include a data dictionary that provides labels for each data element and information on the several versions of ESIC in use over the repeated cross-sections.

Needed extensions. To understand why neither set of documentation is complete, and to appreciate the difficulties of an analyst undertaking to model change in the aggregates, a conceptual model of data and its documentation is required. David (1991, 1993) supplies that model, and explains why metadata exist for each entry in a dataset, for the attributes, and for the entities in the dataset.

Sections 1-2 have discussed aspects of the SCsoi data that require special attention from analysts, and which require information that is not available in existing published documentation. In addition documentation is not conveniently accessible, since details are included in *Publication 16* for each year. The dominant problems for documentation identified are: The data are aggregates of entities; the entities change over time; the content of particular items changes as tax law changes; and representation of entities in the sample is affected by selection associated with some rare and interesting phenomena — large asset size, entity birth, death, and reorganization, and potential disclosure risks.

To assist analysts of the SCsoi and the statutory obligations of the IRS, documentation should fulfill the objectives listed in paragraph 2. We note that available documentation and data afford less capability for understanding and measuring variances than to levels of aggregates. The extent of imputation and truncation in particular aggregates is not quantified. And no work has been done to create an understanding of the relationship between micro-dynamics and changes in aggregates over time.

What is implied for the SCsoi and its documentation? We explore that question after discussing a model for the metadata that provides necessary support for data.

5. A model of documentation — microdata

David (1993) defines necessary support for microdata to include documentation, retrieval capabilities, and adaptation to continuing discoveries about data quality and content. Documentation describes entities, attributes, and values recorded in data arrays and is organized by the temporal process of data production and discovery in data retrieval and statistical analysis. Retrieval requires that every element of necessary support must be uniquely labelled and cataloged.

Necessary support consists of three support-types:

- Literal support consists of data types that can be manipulated in an appropriate computer environment.
- Meanings are interpretation of the literal support in natural language.
- Meta-rules define the processes used by both the data producer and the respondent to complete the data collection and processing.

Furthermore necessary support can be partitioned according to its timing in the process of creating knowledge from scientific measurement:

- Design creates process-related support.
- Execution of the design creates outcome-related support.
- Analysis of the data produced (the outcome) creates analysis-related support.

6. Necessary support for aggregate data

David (1991) indicates that the operations required to document data apply to the documentation of aggregates. The aggregation function is applied to a summand, or target attribute, and defines cells that may be identical to values in the original data (e.g., ZIP code associated with the entity) or they may represent a transformation of underlying data (e.g., first 3 digits of the ZIP code associated with the underlying entity).

The aggregation operation. **A** is defined as an aggregation function that operates on the K rows of the data array **X** including J attributes on K entities. The summand is a subset **X*** of **X**. The aggregation function defines N cells ($N > 0$) using attributes $\mathbf{X}^{**} \subset \{\mathbf{X} - \mathbf{X}^{*}\}$. It is convenient to think of **AX** as including each row once. Transformations of **X** required to obtain definitions of cells (and the summand) must be documented to the same standard as all other operations on the microdata. The rules for metadata presented in David (1993) apply; letters in "( )" and italic expressions in the tables below refer to that article. **A**ι is defined as the aggregation of the unit vector, or the count of the K entities. Disucssion of metadata required for the target attribute **X*** follows.

Inherited metadata. Some metadata pertaining to **X** apply directly to **AX**. The aggregate has the same node in the flow of information as the underlying data. That is, the *parent* (d) and *child* (i) of the summand remain unchanged. The relevant part of the protocol (j), graph of pathways (k), instructions to the editor (m) and respondent (n) are also inherited. Meanings (o) of the classifiers used in forming aggregation cells are the same as meanings of microdata.

Support for aggregates. Much of the necessary support for aggregates is not inherited. Indicators for K_n entities must be aggregated to describe cell *n* in the matrix of aggregates. Indicators describe imputation, truncation, censoring (a-c), inconsistencies (e) and response (r) in the microdata. Applying the identical aggregation operation to indicators and the summand is sometimes useful. For example, if $\mathbf{AX}_j$ is the *average* aggregation function for the summand $\mathbf{X}_j$, **A** can be applied to the indicators for $\mathbf{X}_j$ to obtain the proportion of imputation, truncation, censoring, inconsistency (denoted as $\mathbf{AI}(\mathbf{X}_j)$, and response for the attribute.

Aggregate inconsistencies. In addition to inconsistencies detected in microdata (denoted as **I**(**X**)), inconsistencies occur in aggregates (denoted as **I**(**AX**)). E.g., the aggregate of sales reported by *establishments* should be identical to the aggregate of sales reported by *enterprises*, for a particular accounting period. Aggregate consistencies have historically played an important part in the estimation of the national accounts where measures of the same transactions cumulated over buyers (e.g., wages paid by employers of labor) should aggregate to the same value as when the transactions are measured over sellers (e.g., wages received by workers).

An example. The contributions of corporations are deductible from total income. Documentation for that field in the Csoi microdata illustrates the concepts in David (1993). See Table 1. (We name the table by the line on which charitable deductions appear in the SCsoi, x_{54} to assist in our later illustration of that aggregate.)

The documentation corresponding to the aggregate contributions reported in SCsoi is displayed in Table 2. To avoid repetition, inherited information about microdata on contributions is not repeated. The cells used in the SCsoi represent a sampling on the underlying data, because cells with small numbers of enterprises have been censored to preserve confidentiality of data on the returns for particular enterprises. Thus metadata for response (**P**) must be provided with respect to the aggregates.

The generation function that defines the aggregates is documented as the entry *A3* in Table 2. (Because the aggregation function transforms the underlying return information for particular entities, it is clear that the

logic of the form, including integrities and conditioning is inherited by the aggregate and need not be repeated.) The new data matrix is denoted as Y. It is defined on N industry - asset size classes. Because the cell definition is a transformation of reported ESIC, Year, and Asset size classes, it is clear that documentation of the necessary transformations must be included in the metadata for X**; that metadata properly belongs with the Csoi metadata as a variable transformation. Properties of X** will determine meanings of Y.

At least one aggregate consistency needs to be reported for aggregate contributions. The proportion of enterprises whose contributions are truncated by the statutory limitations is useful. Information on the distribution of truncating thresholds is necessary to analyze and interpret this aggregate.

7. Extensions of the SCsoi

Moments. IRS/SOI (1992, 30-31) has already published coefficients of variation for selected items within minor industry groups. A generalization, with high utility for users is to supply all second moments for a set of attributes on the return. This would make possible a more rigorous testing of relationships among attributes within the cross-section. Two costs limit the extent to which this extension can be realized. The number of entries for J attributes expands to J(2J + 1)/2; this would be prohibitive for printed tabulations, but is quite feasible for data released in diskette form. The more serious cost is that additional information on the joint distribution of J variates has the potential to disclose confidential data. Thus the number of variates J must be considered in relation to the number of cases in each cell and the shape of the joint distributions, a non-trivial task. (NRC, 1993, discusses problems of releasing data on organizations.)

Dynamics. The most pressing extension for supporting data is that SCsoi provide information on the temporal aspects of the underlying data. To what extent does the membership of particular cells remain stable over time? What are the characteristics of units that did not change? This kind of information will be generated as Csoi becomes a panel (Silverstein 1992), and as the underlying information about the structure of families of related corporate entities is captured for statistical purposes (Greenia 1991).

8. Conclusions

Five lessons can be drawn from this discussion. Metadata play an essential part in using data to make correct inferences. Metadata are required for aggregates as well as for the underlying microdata. Necessary metadata to provide for scientific analysis can be generated from the ideas displayed in Table 1 and Sections 5-7. Sections 1-2 make clear that a great deal must be known about the measurements presented in the Csoi before an appropriate statistical tool can be chosen for analysis. Study of the aggregates available in the SCsoi requires metadata on Csoi and additional metadata on the aggregates.

References

Altschuler, Roseanne. 1988. Corporation taxation — Dynamic issues. (Unpublished Ph.D) New York: Columbia University.

David, Martin H. 1991. The science of data sharing: Documentation. In Joan E. Sieber (ed.) *Sharing social science data: Advantages and challenges.* Newbury Park CA: Sage Publications.

David, Martin H. 1993 Systems for metadata: Documenting scientific databases. *Proceedings of the IEEE 26 Hawaii International Conference on Systems Sciences*, January.

David, Martin H. 1992. Review of Michalewicz (1991) *Journal of the American Statistical Association.*

Feldstein, Martin S. and Robert Summers. 1979. Inflation and taxation of capital in the corporate sector. *National Tax Journal* 32:445-70.

Greenia, Nick. 1991. Report to the IRS Consultants Panel on capturing information about related corporate entities.

Internal Revenue Service/Statistics of Income. 1992. *1989 — Corporate Income Tax Returns, Publication 16.* Washington DC: US GPO.

____. (annual) *Sourcebook of Corporation Income Tax Returns.* Washington DC: US GPO.

Little, Roderick A and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

National Research Council. 1993. *Private lives and public policy: Confidentiality and accessibility of government statistics.* Washington DC: National Academy Press.

Silverstein, Jerry. 1992. Multi-year dynamics of the corporate alternative minimum tax. *Proceedings of the NTA-TIA.*

Witte, John F. 1991. The Tax Reform Act of 1986 — A new era in tax politics. *American Politics Quarterly* 19:438-57.

Table 1: Principal items in the metadata system — Charitable contributions in the Csoi (x_{54})

Type of information required Source: *Publication 16*

DESIGN

*A Title:**

Corporation Income Tax Returns — Statistics of Income

Entity specific

A1 Sampling probabilities Figure D p. 10

Meaning of the sample 10-11

Objective of stratification; logic of 100% sampling, and its implications; logic for including only active corporations

Rules for processing the sample

Relation to processing of Business Master File, timing

Attribute specific

A2 Conditioning rules

None. L19, 1120, follows unconditionally from line 18. The rules establish the *parent* of the attribute . Because no conditioning applies, the information is a *leaf* in the datastructure.

A3 Generating function 123

An image of the form is needed to display cues to the taxpayer. In 1120-PL code section references are embedded in the form; in 1120 taxpayers are cautioned about the 10% limitation.

Rules for the generating function 129

Instructions for the taxpayer are essential to prescribing relations to other amounts. Code references (Section 832(c)(9) etc.) also need to be accessible. (If data are manipulated for data protection or other reasons, the applicable rules need to be included here.)

Meaning of the rules

The meaning of the rules often needs explanation. Contributions are truncated in some cases; in other cases the amount includes cash payment from prior years, so tax amounts do not necessarily reflect current flows.

A4 Integrities 123, 129

The amount of contributions is conditioned by the total income, without regard to certain deductions (Line 30 plus line 19 plus line 29b on form 1120). Because the statute limits deductibility, the entry must be tested for conformity to tax accounting rules.

Rules for integrities

The applicable formula must be given in code: $x_{54} < 0.1\ (x_{76} + x_{54} + x_{\_})$ where $x_{\_}$ represents other disallowed deductions. The disposition of cases where deductions exceed the limit must be documented. Because SOI data are captured directly from the tax form, the possibility exists that this integrity was not tested. In that case, the absence of implied integrities must be documented and consistencies reported below.

Meaning of integrities

Because many of the formulae in the tax code are not self-evident, explanation of purpose and interrelationship of code provisions is important. The disposition of cases that fail the integrity must be explained to the user. Is a calculated and truncated valued substituted for the taxpayer entry?

Child of the attribute

Completion of line 20, following line 19 on 1120, requires submission of Form 4562. Hence line 1, 4562 is the child of the entry in contributions.

OUTCOMES

B1 Label for data; meaning

CNTRBTNS; contributions (1120, line 19)

B2 Response

Because blank fields are treated as a default value of 0, response is 100% to this item. Unit non-response may be problem and is handled by weighting or imputation. Details of the weighting calculations must be provided.

B4 *Consistencies---*

If erroneous taxpayer entries are not removed by testing the integrity of the information, indicators must be supplied when $x_{54} > 0.1\ (x_{76} + x_{54} + x_{\_})$.

ANALYSIS

C1 Citations

References to publications where distributions and other findings about contributions need to be included here.

C2 Reports (archival electronic copies)

Table 2: Metadata system for aggregates — Charitable contributions in the SCsoi (x_{54})

Type of information required Source: *Publication 16*

A Title

Sourcebook of Corporation Income Tax Returns — Statistics of Income, 1966-87

Entity specific

A1 Sampling probabilities

Design probabilities Various editions of *Publication 16*

Censored data cells

Rules for processing the sample

The design probabilities are affected by disclosure rules censoring cells. (In addition, as Slemrod notes, archival copies do not exist for some of the historical data.)

Meaning of the sample

Because of the varying concentration of firms the effective sampling rate or mean weight will vary substantially by minor industry classes.

Attribute specific

A3 Generating function: x_{54}

$x_{54} = A(CNTRBTNS)$ The aggregation function

$A: X^* \rightarrow Y$ on the summand $X^* \subset X$.

$(K \times J^*)\ (N \times J^*)$

For weighted sums X^* includes both a vector of amounts and the weighting factor w for each enterprise.

$N = g(X^{**})$ Cell definition on $X^{**} \subset X - X^*$

N is the cartesian product of minor industry classes, year, and size of asset classes. The generating function for the count A_t is defined in the text. A corresponding function is required for the weighting factors applicable to each cell in N, $W = Aw$, where w are the weights associated with corporate entities: $\mu(X^*) = AX^* / Aw$

B1 Label for data; meaning

Y; Sum of X^*

$\mu(X^*)$; Average of X^*

A4 Aggregate integrities

The proportion of enterprises whose contribution is affected by the truncation rule should be calculated and reported here. In addition, the distribution of the values of the truncation points needs to be reported, since the truncation thresholds depend on other characteristics of the reporting enterprise.

*Italicized classification and headings (e.g. *A Title*) are explained in David (1993).

RECENT ADVANCES IN MATCHING AND RECORD LINKAGE FROM A STUDY OF CANADIAN FARM OPERATORS AND THEIR FARMING PRACTICES

Martha E. Fair, Statistics Canada
Statistics Canada, CCHI, RH Coats Bldg., 18-Q,
Tunney's Pasture, Ottawa, Ontario, Canada

Introduction

Record linkage of existing administrative and survey files for statistical purposes is an extremely useful tool. Over the past fifteen years numerous health studies have been carried out at Statistics Canada using probabilistic matching techniques incorporated in a generalized record linkage system (Statistics Canada 1992). There have been a series of workshops describing the methods, software and results of such research (Howe and Spasoff, 1986; Carpenter and Fair, 1989), as well as a number of earlier papers and books published on the methods used (Fellegi and Sunter 1967; Newcombe, 1988; Newcombe, Fair and Lalonde, 1992). Details of the generalized software are described in two other papers presented at this conference (Nuyens 1993; Miller 1993).

Advances in the ease of carrying out computer matching have been achieved by:
(1) developing standard data collection and edit rules for use in the future;
(2) making adjustments or adding questions in the questionnaire design of new surveys (e.g. addition of full birth date rather than age range);
(3) using generalized software to code fields from text (e.g. occupation) and to carry out the linkage;
(4) preprocessing files to generate items such as postal code from addresses, phonetic codes from surnames, and to create alternate entries to search files (e.g. married and maiden surnames for women);
(5) including multiple levels of comparison and specific values in comparing and weighting items, rather than just simple agree/disagree outcomes; and
(6) using a file of NONLINKS created from a set of randomly matched pairs to assist in the generation of appropriate weights for the comparison of items.

These various points will be described in the context of carrying out a mortality and cancer study of Canadian farmer operators. Files with longitudinal follow-up of farms, as well as for individual farmers, were utilized to look at farming practices over time. Some of the methods have general application, whether one is looking at matching individuals or establishments, and these are discussed.

Background -- An Overview of Record Linkage

The emphasis in this paper is on the matching of records where no unique identifier is available. Both an internal linkage (e.g. to identify the same entity within the same file) and several two-file linkages (e.g. matching a file of farmers against a death file) were required. In probabilistic linkage, the comparison or matching algorithm yields for each record pair, a probability or "weight" which indicates whether the records relate to the same entity.

Quite briefly, when comparing values A_x from a Record A (e.g. the farm operators file) with value B_y from a record B (e.g. a death file which is being searched), the ODDS in favour of a correct LINK is estimated for the outcome $A_x \cdot B_y$ (i.e. the comparison pair of values). This may be written in terms of the relative probability of occurrence of the particular outcome in LINKS compared with NONLINKS.

$$\text{ODDS} = P(A_x \cdot B_y \mid \text{LINK})/P(A_x \cdot B_y \mid \text{NONLINK}).$$

As in information theory, the odds are usually expressed as logarithms to the base 2, and are often multiplied by ten and rounded to avoid decimals.

$$\text{Outcome Weight} = 10*\log_2 (\text{ODDS})$$

A rule is created to compare the fields (x,y) in the records. The comparisons can be straight comparison, cross comparisons (e.g. comparing first forename on record x with the second forename on record y), or specially written functions.

The total odds in favour of a match varies as the sum of a number of " outcome weights".

$$\text{Total Weight} = \text{Outcome Weight}(O_1) + \text{Outcome Weight}(O_2) \ldots + \text{Outcome Weight}(O_n)$$

where O_1, O_2 ... O_n are the outcomes for the rules 1 to n (including any used for blocking) used to compare the fields on the records. The outcomes are assumed to be statistically independent.

The total weight becomes an estimate of the "relative probability" that the potential link is in fact a definite link. Details of the calculation of absolute odds are described elsewhere (Newcombe, 1988, Fair, Lalonde and Newcombe, 1991; Newcombe et al. 1992). By comparing the total weight against two thresholds this estimate is converted into a decision as to whether or not the link is a "true" one. If the total weight is above the upper threshold, the link is assigned a temporary status of "definite link"; if it is below the lower threshold, the temporary status is

"unlinked"; if it is between the two thresholds the temporary status is "possible".

<----------(Lower -------Upper) ---------->

Unlinked i.e.	Possibly	Linked i.e.
NONLINKS	Linked	LINKS

Possible links are then examined in more detail, perhaps on a sample basis, to fine tune the setting of the thresholds. In smaller projects, manual resolution can be carried out on these links where reference can be made to source documents which have additional identifiers that were not available in machine readable form. In larger projects, one threshold value is sometimes chosen to classify records as links and nonlinks.

The validity of the linkage process can be expressed as sensitivity and specificity. This is similar to the terminology used in the outcomes of a diagnostic screening test. The sensitivity represents the proportion of true positive links and the specificity represents the proportion of true negative nonlinked records (Herings 1993).

It is important to mention here that all studies involving record linkage at Statistics Canada must satisfy a prescribed review and approval process. For example, the purpose of the record linkage activity must be statistical or research in nature and must be consistent with the mandate of Statistics Canada as described in the Statistics Act. The Statistics Act protects the confidentiality of all records. The various projects require ministerial approval. The record linkage activity should have demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option. It must be shown to be in the public interest.

The Canadian Farm Operator Study

As an occupational group, farmers have low overall mortality (Canadian Centre for Toxicology 1984). However a number of epidemiological studies suggest increased risk of certain cancers among farmers, including cancer of the stomach, lip, prostate, brain and skin, leukaemia, Hodgkin's disease, multiple myeloma, non-Hodgkin's lymphoma (Gallagher et al., 1989; Wigle et al., 1990).

A mortality and cancer cohort study of about 326,000 Canadian male farm operators enumerated in the 1971 Census of Agriculture is being conducted. This is a collaborative study of Health and Welfare Canada and Statistics Canada. The mortality and cancer patterns of 1971 farm operators are being examined in relation to farm practices (e.g. use of commercial fertilizers and herbicides over time) and a variety of socio-demographic variables. The prime concern is the association between specific farm variables which could serve as surrogates to develop exposure indices on individual farmers (e.g. use of commercial fertilizer, herbicides, and pesticides). Longitudinal follow-up of this cohort is being carried out in order to examine mortality and cancer incidence.

The Data Sources

Seven major files were linked to create the data required for the analysis in this study, namely: (1) the 1971 Census of Agriculture (AG); (2) the 1971 Census of Population (POP); (3) the 1971 Central Farm Register (CFR); (4) the 1981 Central Farm Register; (5) the Canadian Mortality Data Base; (6) the 1966-71-76-81-86 Census of Agriculture Longitudinal File; and (7) the Canadian Cancer Data Base. Further exposure and smoking data have been collected for a number of the farmer operators in the 1984 National Farm Survey.

Details of the major files have been described earlier (Jordan-Simpson et al, 1990; Statistics Canada 1992) and are summarized here. Each census year, every farm household in Canada receives both the Census of Agriculture and Census of Population questionnaires. The Census of Agriculture produces a snapshot of Canadian agriculture by providing statistics on such topics as the number of census farms, the use of commercial fertilizer and herbicides and the use of farm land. The 1971 Census of Agriculture included some questions applicable to 1970, such as the number of acres sprayed with herbicides and insecticides, and expenditure for pesticides and for fuel/oil for farm operators. Questions applicable to 1971 related to the size of the farm, the number of acres of crops planted, acres fertilized, amount of livestock holding, farm machinery owned, and other agricultural workers. The Census of Population, on the other hand, provides important information on the Canadian population (such as age, sex, education level, mother tongue and income). The Census of Population contains records for every individual in Canada and was collected using two kinds of questionnaires-Form 2A (short form) and Form 2B (long form). In 1971, the long form was provided at random to one third of private households in Canada. In addition to the 19 questions that were on the short form, it contained 20 housing and 50 socio-demographic questions. In census years 1971

through 1986, census representatives visited each household and dropped off a Census of Population questionnaire. If someone in the household operated a farm, the census representative also left a Census of Agriculture questionnaire.

In the 1971 Census, a census farm was defined as an agricultural holding of one acre or more with annual sales of at least $50 dollars. A farm operator was defined as the person directly responsible for the agricultural holding, whether as owner, tenant, or hired manager. The total number of farms reporting in 1971 was 366,128.

A Central Farm Register, which is a sampling frame used for agricultural surveys, is created from each census. It contains the names and addresses of all farm operators enumerated within a given census year.

The Census of Agriculture Longitudinal File (1966-71-76-81-86) contains data on individual farms for each census during which they were in operation, even if ownership changed hands. It has been created by linking each census of agriculture to that created five years earlier (e.g. 1986 census farms are linked to 1981 census farms).

The Canadian Mortality Data Base (CMDB) contains information for all deaths occurring in Canada dating back to 1950, including name, date, place and underlying cause of death. The death file contains 6.2 million records. The number of records used for this study were about 2.9 million death records relating to individuals for the period 1971-1987.

Canada is one of the few countries in the world with a cancer reporting system covering the whole population. This coverage is achieved through the cooperation of the various provincial/territorial registries which have provided data to Statistics Canada since 1969 (Band et al. 1993). The Canadian Cancer Data Base contains information on cancer cases in a form suitable for record linkage purposes. The cancer file contains about 1.7 million records. About 650,000 records relating to males up to 1986 were used for this study.

In the early planning of this study, it was recognized that the historical files lacked some important information which caused limitations in the ability to link the files and for the analysis, particularly with respect to exposure estimation. Steps have been taken to try to overcome some of these limitations. The 1971 Census of Agriculture did not have the names of farmers in machine readable form - these however were available on the Central Farm Register. The Census of Agriculture did not have the exact date of birth - the year and month were available from the Census of Population. Additional questions were added to the 1984 National Farm Survey, such as complete birth date. The farm operator was asked whether he/she applied the chemicals themselves. Smoking data for farm operators were also collected.

Various Phases of the Study

To form the cancer and mortality analysis file, all the files above were linked together in several phases.

1. Follow-up. The 1971 and 1981 Central Farm Registers were linked to determine if farmers listed in the 1971 were still farming in 1981 on the same farm. If so, this information was added to the 1971 Central Farm Register.

2. Creation of the 1971 farm operators cohort. The 1971 Central Farm Register (CFR) was linked to the 1971 Census of Agriculture (AG) and to the Census of Population (POP) files to create the 1971 CFR/AG/POP farmer operator cohort file. The Central Farm Register unfortunately did not have gender code on the file. A routine was developed, using forename data from other files, which will look up forenames and assign a probability as to whether an imputed gender code would be male or female. Also, another program will examine surname spelling. Surnames are looked up in a dictionary and if not present, the surname is flagged.

In carrying out the linkage of the Central Farm Registry to the Census of Population, it was necessary to link the correct farm operator from a household. The sex code and the birth year and month were extracted from the Census of Population file.

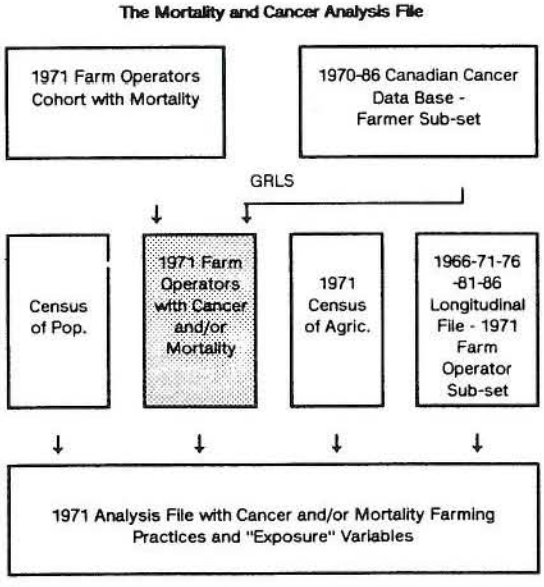
3. The Mortality Linkage. The CFR/AG/POP farmer operators cohort file was linked to the Canadian Mortality Data Base to determine those individuals who were deceased.

4. Farming practices and "exposure" estimates. A linkage of farms from one census to the previous census had been carried out earlier at Statistics Canada by the Agriculture Division and a longitudinal file created. It was necessary to select the information relevant for the 1971 farm operator cohort. Data on direct exposure to pesticides were not available from the 1971 Census of Agriculture. Surrogates used were: (a) number of acres sprayed for the control of insects; (b) number of acres sprayed for the control of weeds; and (c) cost of agricultural chemicals purchased. The 1971 and 1981 census contained questions which were applicable to 1970 and 1980, respectively.

5. Mortality analysis and evaluation of the mortality results. The data are being analyzed selecting the various items required from the Farm Operator file, the Census of Agriculture, the Census of Population, and the Longitudinal files. If necessary, the causes of mortality may be evaluated using data from cancer registries (e.g. cases of non-Hodgkin's lymphoma). Because of the size of the files, the linkages were first tested using Saskatchewan. The production runs were completed by geographic area, and then finally a file for all of Canada prepared.

6. The Cancer Linkage. The 1971 CFR/AG/POP farm operator cohort was matched against the 1970-86 Canadian Cancer Data Base to obtain a subset of records from the cancer file which might relate to farm operators. The cancer file contains incidence information relating to new cancers i.e. it is an event-oriented, not a person-oriented file. This file was then internally linked to produce a person-oriented cancer file potentially relating to farm operators. This cancer sub-set file, which was now person-oriented, was then again linked to the 1971 CFR/AG/POP file, with the setting of thresholds being set more precisely.

7. The Mortality and Cancer Analysis File. The cancer registry data are brought together with the mortality, farming practices, and other variables required for the analysis.



Results

The steps in the farm operators study illustrate the variety and number of linkages often required in a longitudinal study. The birth, longitudinal follow-up, and death of persons for health studies have many analogies that are similar to that for follow-up of entities over time. Some of the overall results that this and other similar studies have shown are the following.

1. Develop standard data collection and edit rules.

In the course of carrying out various record linkage studies, such as the farm operators study, we have found that there are often insufficient identifiers in machine readable form, particularly on older records, to identify an entity. A data collection package has been developed which outlines for long-term medical follow-up studies the type of information which should be collected (Carpenter and Fair, 1990).

2. Questionnaire design.

There were a limited number of identifiers available on the records to identify the individuals on the Census of Agriculture. It would have been particularly helpful to have had the sex code and complete birth year, month and day of the individual. These items were added to a sample of farm operators in the 1984 National Farm Survey.

3. Generalized software.

The generalized record linkage system and the automated coding by text recognition were the two generalized systems used throughout this project, and were found to be very useful. When the 1971 Farm Operator file was linked with mortality data, some possible links were generated. It was found that by referring back to the original source death registrations, that the occupation code was often helpful in resolving some of the doubtful cases. For British Columbia deaths, for some recent years, occupation information was available in the province. These data were obtained and the automated coding by text recognition system used to code the text, particularly if it was for a farmer or farm related occupations. This item was then helpful in the linkages. The iterative aspect of the generalized software enables one to modify the linkage rules and weights. Often additional rules can be applied that would be particularly helpful in separating out borderline cases, such as the use of name variants, truncations and so forth (Newcombe et al. 1992).

4. Preprocessing of files.

A routine was developed to impute a gender code where forenames were available. This is a particularly helpful routine to use when investigating new files to check for possible errors in gender coding. This routine examines each forename and then calculates a weight indicating whether the record relates to a male or female. Some caution is required for

forenames that can be used for either gender - for example Jean Marie. Surname spellings can be checked in the preprocessing of the files. If desired, alternate entries can be generated to facilitate the linkage. For example, it was once thought that it would be profitable to change the phonetic code routine (referred to as the NYSIIS code) to take into account particular problems which had been identified with French phonetics. We however found with various tests, and examining a number of phonetic codes, that it was preferable to create alternate entries, that is, to generate additional records with alternate spellings for the same individual.

5. Linkage rules and outcomes.

When examining records to see whether they relate to the same entity (e.g. farm) or person (e.g. farm operator), two things are usually done. First one must decide on the comparison rules and then a decision is made as to whether to accept, reject or further assess the linkages. In the past, comparison rules have often been unnecessarily simplified with emphasis on agreement and disagreement. Multiple comparison outcomes are possible and have been used to include common levels of discrepancies and common inversions. As an example, one can have the exact spelling of a name, a phonetic code of a name, or agreement only of a particular portion of a name. If items are correlated, these can be dealt with by concatenating these identifiers and recognizing further kinds of comparisons.

6. Use of NONLINKS.

A file of NONLINKS can be generated to get more precise weights for items (Lalonde, 1989). For example, to do this, both files being linked are sorted randomly so that they are in no particular order. The files are then blocked randomly into a thousand or more pockets or blocks within which records are compared. Each record is assigned a unique sequence number. The records are then matched and appropriate weights derived.

The overall results of the health study have been reported in a number of articles and further analyses are in progress. Initially, the study of non-Hodgkin's lymphoma mortality experience of male Saskatchewan farmers for the period 1971-85 indicated a significant trend in risk of non-Hodgkin's lymphoma among farm operators according to fuel/oil expenditures and herbicide spraying for farms < 1,000 acres (Wigle et al. 1990). Recently, two additional Canadian prairie provinces and two additional years of follow-up, along with data from the 1981 Census of Agriculture, have been analyzed. Based on the 1971 Census of Agriculture data, no excess risk was observed between herbicide spraying and non-Hodgkin's lymphoma for Alberta or Manitoba, and lower risks were noted in Saskatchewan with the two additional years of follow-up. However, a significantly increased risk of non-Hodgkin's lymphoma according to acres sprayed with herbicides was observed for the three prairie provinces combined, using 1981 herbicide spraying data (>= 380 acres sprayed, RR=2.11, confidence interval = 1.1-3.9) (Morrison et al. 1993).

Other analyses have been conducted looking at prostate cancer mortality (Morrison et al. 1993) among prairie farmers who were 45 years and older that were identified in the 1971 Censuses of Agriculture and Population. A weak, but statistically significant, association was found between the number of acres sprayed with herbicides in 1970 and risk of prostate cancer mortality. These findings encourage further detailed research to examine the effects of herbicides on prostate cancer. Brain cancer has also been examined using the records of male Alberta, Saskatchewan and Manitoba farmers (Morrison et al. 1992) as well as multiple myeloma mortality (Semenciw et al. 1993).

The Future

Future work with respect to record linkage development lie in the refinement of error estimates and threshold setting. The software that was used in the present study relates to the main frame version of the generalized systems. Effort is also being made in developing this software for smaller machines, both at Statistics Canada and at a number of centres. The postal code conversion file is being updated, particularly to refine rural postal code areas. This will aid in getting more precise geographic location and distances between areas.

Currently a further development is being planned in order to get more precise exposure data for the farm operators study. For example, a follow-up of the 1984 National Farm Survey cohort is anticipated. An extensive bibliography has been developed with respect to various publications regarding the health of farmers.

Acknowledgements

The author would like to thank Christine Poliquin for her contribution to the farmers study, and Pierre Lalonde who has designed many of the record linkage refinements used. The farmers study is a collaborative cost recovery study with Health and Welfare Canada. The cooperation of the various provincial/territorial vital statistics registrars and cancer registries is also acknowledged.

References

Band, P.R., Gaudette, L.A., Hill, G.B., Holowaty, E.J., Hutchcroft, S.A., Johnson, G.M., Illing, E.M., Mao, Y., and Semenciw, R.M. (1993), **The Making of the Canadian Cancer Registry: Cancer Incidence in Canada and its Regions, 1969 to 1988.** Copies available from: Bureau of Chronic Disease Epidemiology, LCDC, Health and Welfare Canada, Ottawa, K1A 0L2.

Canadian Centre for Toxicology (1984), **Agricultural Chemicals and Farm Health and Safety.** The Ontario Task Force on Health and Safety in Agriculture: Toronto.

Carpenter, M., and Fair, M.E. (1990), "A Standard Data Collection Package for Medical Follow-up Studies", **Health Reports,** 2, 157-173.

Carpenter, M., and Fair, M.E. (eds.) (1989), **Canadian Epidemiology Research Conference - Proceedings of the Record Linkage Session and Workshop.***

Jordan-Simpson, D.A., Fair, M.E., and Poliquin, C. (1990), "Canadian Farm Operators Study: Methodology", **Health Reports,** 2, 141-155.

Fair, M.E., Lalonde, P., and Newcombe, H.B. (1991), "Application of Exact ODDS for Partial Agreement of Names in Record Linkage", **Computers and Biomedical Research,** 24, 58-71.

Fair, M.E., Newcombe, H.B., Lalonde, P., and Poliquin, C. (1988), **"Alive" Searches as Complementing Death Searches in the Epidemiological Follow-up of Ontario Miners.** Report No. INFO-0266, Ottawa: Atomic Energy Control Board.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory of Record Linkage", **Journal of the American Statistical Association,** 40, 1183-1210.

Gallagher, R.P., Threlfall, W.J., Band, P.R., Spinelli, J.J. (1989), **Occupational Mortality in British Columbia 1950-1984.** Available from: Cancer Control Agency of British Columbia, Division of Epidemiology, Biometry and Occupational Oncology, 600 West 10th Avenue, Vancouver, B.C. V5Z 4E6

Herings, R.M.C. (1993), **PHARMO: A Record Linkage System for Postmarketing Surveillance of Prescription Drugs in the Netherlands.** Ronald Marinus Cornelis Herings. Utrecht: Utrecht University.

Howe, G.R. and Spasoff, R.A. (eds.) (1986), **Proceedings of the Workshop on Computerized Record Linkage in Health Research.***

Lalonde, P. (1989), "Deriving Accurate Weights Using NON-LINKS". In: Carpenter, M. and Fair M.E. (eds.) **Canadian Epidemiology Research Conference - Proceedings of the Record Linkage Sessions and Workshop.** pp 149-157.*

Miller, D. (1993), "Automated Coding by Text Recognition", **International Conference of Establishment Surveys.** (See this volume).

Morrison, H., Savitz, D., Semenciw, R., Hulka, B., Mao, Y., Morison, D., and Wigle, D. (1993), "Farming and Prostate Cancer Mortality", **American Journal of Epidemiology,** 137, 270 - 280.

Morrison, H.I., Semenciw, R.M., Morison, D., Magwood, S., and Mao, Y. (1992), "Brain Cancer and Farming in Western Canada", **Neuroepidemiology,** 11, 267-276.

Newcombe, H.B. (1988), **Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business,** Oxford, U.K.: Oxford University Press.

Newcombe, H.B., Fair, M.E., and Lalonde, P. (1992), "The Use of Names for Linking Personal Records", **Journal of the American Statistical Association,** 87, 1193-1204.

Nuyens, C. (1993), "Generalized Record Linkage at Statistics Canada", **International Conference of Establishment Surveys.** (See this volume).

Semenciw, R.M., Morrisson, H.I., Reidel, D., Wilkins, K., Ritler, L., and Mao, Y. (1993), Multiple Myeloma Mortality and Agricultural Practices in the Prairie Provinces of Canada. **Journal of Occupational Medicine,** 35, 557-561.

Statistics Canada, Occupational and Environmental Health Research Section (1990), **Data Collection Package.** Report No. OEHRS- No. 7.*

Statistics Canada, Occupational and Environmental Health Research Section (1992), **Studies and References Relating to Uses of the Canadian Mortality Data Base.** Report No. OEHRS-No. 11.*

Statistics Canada, Agricultural Division (1992), **Census Overview of Canadian Agriculture: 1971-1991.**Catalogue 93-348. Available from: Publication Sales, Statistics Canada, Ottawa, K1A 0T6.

Wigle, D.T., Semenciw, R.M., Wilkins, K., Riedel, D., Ritter, L., Morrison, H.I., and Mao, Y. (1990), "Mortality Study of Canadian Male Farm Operators: Non-Hodgkin's Lymphoma Mortality and Agricultural Practices in Saskatchewan", **Journal of the National Cancer Institute,** 82, 575-582.

* Copies available from Statistics Canada, Occupational and Environmental Health Research Section, RH Coats Bldg., 18th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6

DISTRIBUTED DATABASE SYSTEM: 3 CRITICAL ELEMENTS

Phillip L. Zellers, Rod DeSmet, C. Jenny Shiao, NASS, USDA
C. Jenny Shiao, Room 4839-S, NASS, USDA, Washington, D. C. 20250

KEY WORDS: Distributed, database, rightsizing, strategy, technology, implementation, re-engineering

Introduction

In the early information age in the 1960s, the U.S. government, scientists and corporations started to develop systems to reduce the mounting labor effort in the printing paradigm. It was a challenge to allow people access to those systems in an on-line environment. This early information experimentation demonstrated the many centralized data processing concepts in the manufacturing, airline, banking, and census industries. In the 1980s, the data entry mechanism became more affordable. The mass reproduction of systems and data flourished. Voluminous databases and their presentations established the foundation of today's proliferated information base. In the future, the quality, cost, and efficiency of the information and users' perceptions will be improved continuously.

Mainframes have traditionally provided the most effective and controlled utilization of information technology. With the introduction of PCs and workstations, information started to move from the centralized mainframe to local computers. End users are seeking more control and autonomy over their data. The replication of data and the parallel processing on various platforms involve risks of losing security, lack of data integrity and increasing problems with synchronization. MIS Departments face the challenge of implementing distributed database systems in a heterogeneous computing environment. No longer, is it sufficient for them to manage data; management and staff experts demand meaningful information. They will have to develop systems to turn data into useful information. The current environment, where data is incarnated by application systems, will have to end. To meet this demand, computer professionals must work cooperatively with end users. Executives need to empower computer professionals to produce quality systems in a much shorter time frame.

A Transition Plan is needed to ensure a smooth migration from the existing information system to the distributed environment which is currently required. The strategy is to assess the maturity level of the current information processing. The future system shall be clearly defined through business process re-engineering. The strategy should integrate desired business processes, data, implementations, operations, and advanced technology.

The National Agricultural Statistical Services (NASS) is the statistical agency in the U. S. Department of Agriculture. NASS conducts agricultural statistical programs through 45 State Statistical Offices (SSO). The agency has implemented Local Area Networks (LAN) in each SSO, and most recently in its headquarters in Washington D.C. Statisticians, mathematicians and computer specialists in all NASS offices have access to a full range of computer resources from a desktop PC to a large mainframe through LAN's. NASS has long been an advocate of advanced Data Base Management System technology and it continues to explore the most efficient mechanisms to distribute the

survey and estimation processing over both the LAN and the wide area network. This paper discusses three critical components of an effective and efficient distributed database system: strategy, technology and implementation, from an operational point of view.

Strategy

The first critical component of an efficient and effective distributed database system is **strategy**. Strategy does not rationalize the benefits or disadvantages of the distributed database system; neither does it address methodologies for utilizing resources, evaluating a specific technology, or selecting a suitable information system to be distributed. All these rationalizations and methodologies do not provide us a distributed database system. A strategy is needed for executives to make decisions. The direction of the distributed database system is a matter of organizational policy. The policy sets boundaries and rules, and the strategy dictates the decision of what has to be done, how, when, where and by whom.

We need to understand the current status of the information system and the involved processes. By developing a new vision of the desired business processes, the necessary actions can be defined. For a distributed database system, software quality becomes more critical. We need to assess the maturity level of the system, and identify risks involved in situations where the business stays as usual and where changes are made. The software system's quality is relative to its level of maturity. The higher the maturity level, the fewer risks the system anticipates.

The Software Engineering Institute (SEI) Process-Maturity Framework Model is useful in depicting the need of a quality distributed database system. SEI, founded by DoD, promotes quality software processes as its primary mission. The empirical model has five maturity levels, Initial, Repeatable, Defined, Managed, and Optimized. The following describes each maturity level, its characteristics, and the key improvement area for the SEI model.

Initial - An ad hoc system is developed as requested. End users rarely involve in the development process. There is no business plan and neither a development plan. People are the key to the success of the system.

Repeatable - A proprietary distributed database system is developed for a specific need. It faces major risks when requirements change. The system quality presents little risks as long as it is within the pre-determined plan and configuration.

Defined - Business plan specifies business processes and data models. A quality distributed database system is developed according to the plan, which uses open standards and advanced technology. Training and reviews are success factors for this level. Computer professionals and end users are given incentives and take great pride in implementing the system.

Managed - A measurable and controlled distributed database system is developed when the development plan and system processes can be measured quantitatively. The system quality is enhanced with the strong commitment from management.

Optimized - An effective and efficient distributed database system is developed and continuous quality review is conducted. The automated process improves human productivity. System quality is achieved through the Defect Prevention Process (DPP). The organization's performance is judged by

meeting the planned objectives.

The process of assessing system maturity level reasonably represents the evolutionary improvement of the software development from the past. The model provides guidelines for improving system quality. The maturity level assessment helps to define actions needed for enhancing the quality of the distributed database system.

A system development and improvement plan is required to accommodate actions which need to be prioritized and assigned appropriate resources. Furthermore, executives need to commit resources to execute the plan. A measurement process is mandatory to monitor and control the system maintenance and operation. From an operational point of view, system quality is achieved by preventing problems and by continuously improving the system. Quality is the key for an effective and efficient distributed database system.

The process of developing the distributed database system is referred to as rightsizing. NASS System and Information Division initiated a number of rightsizing development activities. It is out of the scope of this paper to assess the maturity level of NASS systems. However, the maturity framework could be applied to improve NASS Standard Processing Technology and its LAN-base general purpose systems.

Technology

The second critical component of the distributed database system is **technology**. Included in the technology are data base management systems, micro processors, network facilities, development tools, and user interface tools. All of these technological aspects must be considered.

The fundamental technology is the Data Base Management System (DBMS). A true distributed DBMS system provides users with simultaneous update capability of multi-vendor databases residing on different platforms. DBMS vendors incorporate this feature into their product lines at various levels. Relational DBMS systems encapsulate the flexibility and transparency of data access. Moreover, the distributed DBMS needs to be enhanced with the object oriented technology which provides a high level integration of complex information; i.e. images, documents, video and audio with the advanced hardware.

CPU, memory, I/O throughput and disk storage are essential items for capacity planning and configuration management. A low end Reduced Instruction Set Computer (RISC) is three times faster than Intel 386 microprocessors. A configured RISC computer or symmetric multi-processors can be as powerful as an IBM 360 mainframe computer. In the last ten years, the computing price performance has dropped continuously from $10,000 per Million Instructions Per Second (MIPS, a way to measure computer performance) to less than $1,000 per MIPS. The openness of the computer architecture strengthens competitiveness among hardware vendors.[1]

The local area network is a cost effective network design, which allows users to share resources; i.e. printers, software, and disk storage, at a local level. The wide area network, such as FTS2000, preserves users' access to the mainframe and computers at other locations. Electronic mail messages, network file sharing, and the access to the heterogeneous computing environments can be implemented with appropriate communication gateways, routers, and bridges. Conforming to the standard link, transport and inter-connect communication protocols, 10 BaseT, x.25, x.400 and TCP/IP standards, is

mandatory for an efficient network arrangement.

The high performance computers and network facilities contribute to high quality distributed DBMS. Operating Systems are the major ties between a quality distributed database system and a high performance computer. It is essential that the operating system is open and portable on various platforms. IEEE's POSIX standard clearly specifies the requirements of a portable operating system. All operating system vendors strive to conform to this standard with a set of common frameworks.

Security is one of the major problems in an open and distributed computing environment. Per the DoD Orange Book, security practices (in order of increased security, C, B, and A) must be considered in the distributed computing environment. Adequate security administration prevents a malicious user from purposely locking up the system.[2]

Graphical User Interface (GUI) is one of the most revolutionary changes to the human/computer interface. It has changed from a terse, character orientation to familiar windows, icons and menus interfaces. WYSIWYG (What You See Is What You Get) and X Windowing are the standard representations for user interfaces, which allow users to retrieve and manipulate complex data types and large databases in an appealing and comprehensive manner.

The major risks involved in the development of a distributed database system are the dependencies on the existing information systems, commercial software, hardware and communication products. It is especially critical in the area of rapid changing technology, demanding network management, complex re-engineering effort and incompatibilities among products. The software packages in Remote Procedural Calls (RPC), Application Programming Interface (API) techniques, Standard Query Language (SQL), and application and system administration tools provides a degree of independence. Without step-by-step refinement and analysis, the integration and testing of software packages can throw the schedule and budget off tract.

As might be expected, with the 45 State Statistical Offices, the planned distributed database environment in NASS will combine FTS2000, advanced client-server solutions, PC-mainframe connectivity, and WINDOW software. NASS continues to improve these tools by refreshing software, hardware, and communication components.

Implementation

The third critical component of the distributed database system is **implementation**. To fully exploit the new technology, re-engineering systems are developed for re-designed business processes. Thus, the new technology drives the new business processes and vice versa. Most of today's development work centers on re-engineering of existing systems. It requires a comprehensive abstract and analysis of the data environment, processes and interfaces with associated systems. The implementation of any new system requires the understanding of the artifact of the existing system's environment.

Traditionally, the system development life cycle distributes its efforts as follows: 30% in design, 30% on coding, and 40% in testing. The testing effort of some NASS systems is as high as 80%. The purpose of DPP is to increase the system quality, reduce the ratio of the testing and maintenance effort, and to enhance the productivity of developers and end users. A typical defective preventive process in the system life cycle is to have an

analysis team perform the code inspection and to have an action team formulate the improvement actions. With the early visibility of the possible defects in the implementation, DPP ultimately seeks preventive actions and cuts down the time and schedule in testing, integration, and maintenance thus reducing life cycle cost. In 1990, Hughes Ground System Group, by adopting DPP process, experienced a 50% increase in Cost Performance Index (CPI) and a turn over rate below 10 percent.[3] The effect of implementing a distributed database system decreases is that the cost of the system administration, data processing and the management of multi-vendor computing environments is under controlled, reliable and predictable.

It takes proper plans and strategies to effectively move the information among various platforms. The maturity framework discussed in the Strategy section defines optimization as the highest maturity level. Optimization does not exist in the real world. Progressive improvement of standards and procedures, database administration, configuration management, and security administration are achievable.

- Standards and procedures are improved with built-in human intelligence. They are defined, documented, and can be measured in a quantitative way. The automation of processes and procedures requires little human intervention.

- The database administration (DBA) functions include establishing procedures, setting standards, and educating users. Two most important DBA functions are data dictionary and data recovery. Data Dictionary contains process information of metadata, where the active reference of data is created automatically. For the re-engineering work, it is essential that the data repository is clearly defined and that they are mapped to each other in the existing system and the new system.

- Another determinant factor of a quality distributed database system is configuration management. Only when the configuration of baseline products and processes is manageable and controlled, can the autonomy of resources, facilities, vendors, systems, data, and procedures be obtained. Thus, risks start to decrease and productivity increases.

- Implementing a good system requires not only technology but also a sound and secure environment. Security administration is an example of the defect preventive process for secured business practices. It must be considered.

NASS implemented PC Summary, County Estimate, AGI (agricultural information), PC Ole (list overlap), ELMO (Enhanced List Maintenance and Operation), and other LAN-based systems. They provide interfaces to the existing processes, data, and systems residing on the mainframe. NASS enforces change control and security administration on its LANs and mainframe as part of the configuration management.

Summary

Procedures, data and application systems continuously evolve. In order to meet end users' requirements in today's distributed computing environment, continuous quality improvement is required. The maturity level assessment and the defect prevention process help to accommodate these requirements. For an effective and efficient distributed database system, the three critical elements of strategy, technology, and implementation must be thoughtfully considered.

REFERENCES:

[1] "A RISC Tutorial", Sun Microsystems, Inc.

[2] "An OEM's Guide to a Secure Unix System", UNIX World, February. 1988

[3] "Software Process Improvement at Hughes Aircraft", IEEE Software, July, 1991