Linking retail establishments reported in a household survey

Charles Mason, Bureau of Labor Statistics

Background: The Bureau of Labor Statistics' Point of Purchase survey collects information on the retail establishments where consumers purchase the goods and services covered by the Consumer Price Index. Among the data collected are the establishment names and addresses, which are used as a sampling frame for the selection of establishments priced in the CPI. Prior to selecting the sample of establishments, the raw data must be coded and collapsed so that multiple reports of the same establishment are combined. Additionally, address information is refined so that field economists can locate and initiate the selected establishments into the CPI. This paper reports on the recent change by the BLS from a computer assisted manual coding and collapsing procedure to a statistical linking processes that utilizes Soundex and additional scoring considerations.

The U.S. Consumer Price Index (CPI) is the principal measure of information concerning trends in consumer prices and inflation in the United States. Policy makers and analysts in both the public and private sector study the index extensively. Changes in the index level have significant impacts on the finances of the Federal Government. It is used to adjust payments to Social Security recipients, to Federal and Military retirees, and for a number of entitlement programs such as food stamps and school lunches. In addition, individual income tax brackets and personal exemptions are adjusted for inflation using the CPI. According to the Congressional Budget Office in Fiscal Year 1999, a one-percent change in the rate of growth of the CPI directly changes Federal spending and revenues by a net total of almost $6.0 billion.[1]

One of the key components in the construction of the CPI is the Point of Purchase Survey. This survey is designed to provide a sampling frame of retail establishments and other outlets frequented by urban households from which prices can be collected and used to compute the monthly index of price change.

The Point of Purchase Survey is collected for the Bureau of Labor Statistics (BLS) by the Bureau of the Census as a computer assisted random digit dial survey. Three key pieces of data are collected in the survey: The name of the establishment from which the purchase took place, the address of the establishment, and the amount spent at the establishment for a specific category of goods or services during a specified time period.[2] Previous research has demonstrated that survey respondents are able to provide highly accurate data for establishment names but that the address and expenditure data are less accurate. [3]

The accuracy or completeness of the address information is critical to the CPI because without complete and accurate address information it is difficult for CPI price collectors (Economic Assistants or EA's) to locate the selected outlets in order to collect the monthly prices necessary for computation of the CPI. Historically, Census has undertaken an extensive manual address refinement process during which addresses were standardized and multiple reports of the same outlet were collapsed into a single address record. The process was effective in improving address quality but came at a high resource cost. In 2000, BLS determined that the manual process was not cost effective and decided to test a probabilistic linking address refinement process.

Probabilistic linking is a subset of record linkage and is an effort to identify records on two different electronic files that contain information about the same entity and it varies from exact record linkage in the accuracy that can be assumed about the result. For exact matches, the goal is to match two or more records on the same individual based on unique identifying information. For example, Payroll records and employee training records can be matched to get a more complete employee record when social security numbers or employee identifiers are present on both records. Probabilistic Matches are essentially the same as exact matches except that identifying, or linking variable, may not be completely unique and as a result differing weight of evidence characterizes different matches. For example birth and death records can be linked using first and last name, but there is less certainty about the accuracy of the match since names are not unique and can change over time.

The CPI program did not have a history of using record linkage and first needed to develop a record linkage process. Our initial tests of record linkage were based on crude linking software but were sufficient to demonstrate the potential for the approach. Specifically, during the summer of 2000, a summer intern was able to modify a record linkage software tool obtained from Census and to evaluate TPOPS record linkage. The software obtained was designed for one-to-one matching. The first step for CPI was to convert the software to support many-to-one linking, since in our application, the same establishment could be reported with multiple spellings and addresses. After converting the software, we evaluated record linking based on data collected in Baltimore MD and Provo UT. Baltimore was chosen because we were familiar with the retail establishments and the city, while Provo was chosen because of its "letter-and-number" grid structure. The testing revealed that we could process up to 50,000 input records in about one minute of computer time after standardizing the address data. In addition we achieved varying match rates ranging from 21% up to 52% depending on the source of address data used in the matching. The highest rates were obtained using the previously collected and refined addresses from our own survey. Based on the results, a decision was made to proceed and to develop our own record linkage application.

We established the following goals for the record linkage application:

- 75 percent of the establishments reported in the TPOPS survey had to be matched to previously reported outlets or coded as new outlets. No more than 25 percent of the reported establishments could be output for manual review.
- Of the 75 percent coded establishments, 95 percent had to be accurately coded.

The completed application consisted of the following components:

Standardization of establishment and address data

Data file blocking. Essentially blocking is sorting and filtering the files to increase the efficiency of record linking. The efficiency is achieved by limiting the number of record pairs that must be examined since each record to be matched in a block is only compared to the corresponding block in the file of potential linkage records file. In our case since each file contained address data, the file was be blocked by City and State so the linking comparison was only performed among address records from the same reported geographic area.

Scoring. Records were scored based on the likelihood of a match as determined by the scoring algorithm provided below. The scoring was based on a combination of Soundex and character string matching and included as many of the reported variables as possible. The table below provides the basic match scoring algorithm developed.

Table 1. Record Linkage Scoring

Variable	Maximum Score	Scoring Factors
OUTLET NAME	120	(100% if SOUNDEX match 94% if SOUNDEX match 3-digits 75% if SOUNDEX match 2-digits) times (# of character matches in the outlet names [upto the # of characters in the shortest name]/length of shortest name)
STREET or | Cross STREET	40	100% if SOUNDEX match 75% is SOUNDEX match 3-digit
SHOPPING CENTER	40	100% if SOUNDEX match and exact character match 50% if SOUNDEX match
CITY	25	100% if SOUNDEX match 60% if SOUNDEX match 3-digit
STATE	15	100% if STANDARDIZED STATE CODE match character string
TOTAL	240	(less maximum score for any missing field)
Normalized Linkage Score	100	Sum of Variable Scores / TOTAL

For each variable, the score assigned was the maximum score times the scoring factor

described in the third column. For example, if the reported outlet name had a perfect SOUNDEX match with a previously reported outlet and had complete character matches, then the score assigned would be 120 * 100(times 1) or 120.

Matching. Based on the best score, each record was scored against each existing record, the reported retail establishments were assigned to one of three categories – Matched establishments, New reports, or Manual review needed for assignment. Based on the test cities, the initial matching thresholds were set at 83 and 0. Retail establishments with a normalized linkage score (NLS) exceeding 83 were assigned the outlet code of the matched outlet. All other outlets were assigned for manual review except for those outlets where the NLS was below 40, multiple matched outlets scored the same, and there was no character string match beyond the first character. In these cases, the reported outlet was assigned as a new outlet. Based on the calibration step, the lower threshold was eventually raised to 40. Outlets with NLS's at or below 40 were then assigned a new outlet code and only those outlets with NLS greater than 40 and less than 83 were output for manual review.

Calibration. A review process through which the scoring and matching were optimized.

Results: The record linkage application was run on data collected between April and September 2002 in the TPOPS survey. All 87 geographic areas in the CPI were included as were all item categories in the index. The results of the record linkage experience are summarized in table 2 below. On average nearly 45,000 outlet reports were obtained in each quarter of the survey. Of those reports, approximately 7,000 were not eligible for record linkage because they were internet or mail order purchases. These outlet reports were excluded from the initial runs of the record linkage because we felt that special scoring rules would be needed.[4] During the initial record linkage runs, two-thirds of the outlets were either automatically linked (or coded to existing outlets) or were identified as new outlets. One-third of the reported outlets were assigned for manual review. Based on this result, the first goal of 75 percent of outlets coded without manual intervention was not met at the initial thresholds.

Table 2. Results from AUTOMATED Linkage

	2002 Q2	**2002 Q3**
Total Records for Matching	44194	47261
Ineligible for linking	7245	6144
Eligible for linking	36949	41117
Matched to existing Establishment	11411	12329
No Match (new outlet)	10324	10818
Manual Review needed	15214	17970

Table 3 reports the results achieved with respect to the second goal that 95 percent of the automated records need to be assigned correctly. The decision table presented compares the automated decision against the "correct" or true result which was determined based on a manual review of the data. Because the address information is incomplete,[5] the manual review could also have resulted in errors but this outcome has been ignored in the current analysis. As can be seen, the automated process was very successful in correctly matching outlets. Successful automated matches were achieved in over 98 percent of the instances when automated matches were made and 100 percent of the new outlets were similarly found to be correct when reviewed manually. Finally, manual review demonstrated that there was room for improvement in the application because nearly half of the outlet reports assigned to manual review resulted in a match to an existing record.

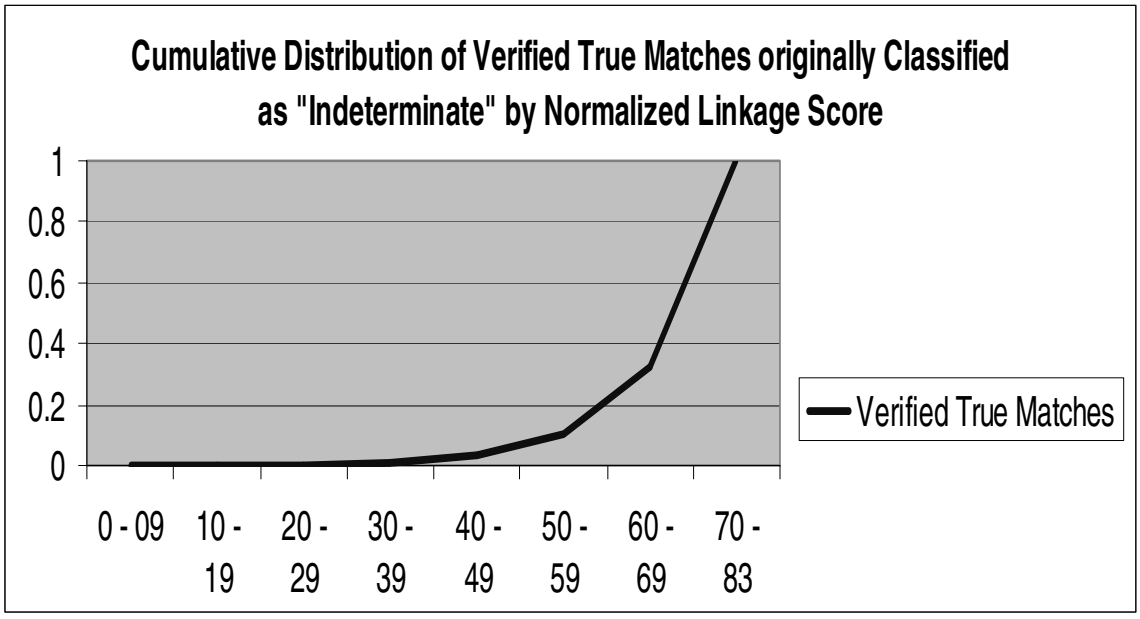
Based on the results of the manual review the system was recalibrated to determine if additional outlets could be matched in the automated phase of record linkage thereby reducing the need for manual coding. First we examined the distribution of matches and new outlets by NLS. The results of that review are shown in chart 1.

Table 3		**Correct Result**	
		No Match	**Match**
Automated	**Match assigned**	226	11184
Outcome	**New Outlet assigned**	10324	0
	manual review	7968	7246

From this chart it was clear that nearly all outlets with NLS scores below 40 were new outlets without a corresponding match in the previously reported files. Similarly, 50 percent of the outlets with NLS scores between 70 and 83 were associated with a previously reported outlet and should have been treated as a match during the automated record linkage. Based on these findings, it was decided that the automated thresholds should be modified. The lower threshold was raised to 40, resulting in 4,050 more new outlet assignments and a related reduction in the number of outlets requiring manual review.

Raising the threshold however, resulted in an increase in the number of errors, and increased the error rate to 2.6 percent. Four hundred and fifty of the newly assigned new outlets in actuality had a match in the previously reported outlet file. This was determined to be an acceptable error rate, since it was still well within the goal of a 5 percent error rate and equally as importantly, this error could be corrected in the field. Economic assistants, who were assigned to collect prices in one of these 450 outlets coded as new, would find that they were already collecting prices in these same outlets under a different outlet code. In this case, they would be able to merge the two outlets into a single collection entity.

Chart 1.



Although the data review suggested that manual intervention could be reduced by lowering the upper threshold, it was determined to keep the original value in place for the time being. Lowering the threshold to 80 for example would increase the number of automated matches by over 2000, substantially reducing the manual process. However, it would increase the number of erroneous matches by 400. While we would still be within our stated 5 percent error goal, the errors introduced would be non-correctable in the field. Unlike erroneous "new outlets", erroneously matched outlets could not be detected by the EA's. To them it would appear as if an outlet had been reported by a survey respondent by mistake. If the outlet did not sell the item for which it had been selected, a price quotation needed for the CPI's computation would be lost. Without additional research the CPI program was unwilling to accept higher rates of this error type. Table 4 shows the final outcomes achieved by the automated record linkage application.

Table 4. Results based on final Thresholds

		Correct Result	
		No Match	**Match**
Automated Outcome	**Match assigned**	226	11184
	New Outlet assigned	13924	450
	manual review	4368	6796

Conclusions: The CPI program was able to develop an automated record linkage application

that replaced a resource intensive manual process. The initial results of the application met the goals established and demonstrated the worth of this approach. There remains additional work to expand the applicability of record linkage to include mail order and internet outlets and to fine tune the thresholds in order to further reduce manual intervention and to minimize erroneous matches which cannot be correct by the EA's in the field.

---

[1] The Economic and Budget Outlook: Fiscal Years 1998-2007, A report to the Senate and House Committees on the Budget, Congressional Budget Office, January 1997.

[2] For additional information on the Point of Purchase Survey see *Handbook of Methods*, Bureau of Labor Statistics, "Chapter 17, The Consumer Price Index."

[3] Mason , C., U.S. Bureau of Labor Statistics, (2000) " Results from a Random Digit Dialing Survey of Where Consumers Purchase Goods and Services ", *Proceedings of the Section on Government Statistics*, American Statistical Association

[4] Additional testing has revealed that the record linkage application developed can be used for internet and mail order outlets as well with comparable results.

[5] If the survey had been more successful in obtaining complete and correct addresses, there would have been no need for the record linkage application.