The Philippine Statistician, 2006
Vol. 55, Nos. 1-2, pp. 41-54

Empirical Evaluation of Statistical Matching

Zita VJ Albacea, and Glenn Melvin P. Gironella, Salve Kristina T. Mogul and Jezaniah Kira S. Tena [1]

Received: August, 2005; Revised: May, 2006

ABSTRACT

This paper aims to empirically evaluate statistical matching as a method to create panel data sets. Two methods of statistical matching, namely: constrained and unconstrained statistical matching were studied. The properties of the distribution of distances measured using a distance function were examined. In particular, the 1997 Family Income and Expenditure Survey (FIES) and 1998 Annual Poverty Indicators Survey (APIS) data of Southern Tagalog Region in the Philippines were statistically matched after removing the exact matched records. To identify the "best" form of the distance function to use in matching, four statistically matched files (STM_A, STM_B, STM_C_UN, STM_C_WE) using different forms of a distance function were obtained. Descriptive statistics, histograms, and box and whisker plots of distances measured show the closeness of the values of the matching variables in the performed statistical matching. The cumulative percentage distributions show that at some point almost 95% of the data are already matched having small calculated-distances. The results indicate that the form of the distance function used in generating STM_A is the "best" since the distribution of the distances measured in STM_A clusters near zero with 94.95% of the records that were matched having calculated distances at most equal to one. Also, this form of the distance function resulted to a unit less measure of distances. With the distance function used in STM_A, unconstrained and constrained statistical matching procedures were applied to generate the unconstrained (UNC_STM) and constrained statistical matched data sets (CON_STM), respectively. The statistics of test variables in the exact matched data set (EXM) are found to be similar with those of the statistically matched data sets. However, the average computed distances, which indicate the degree of similarity of the matched records of the UNC_STM data set is found significantly lower than that of the CON_STM data set at 1% level of significance.

Keywords: unconstrained and constrained statistical matchings, distance function

I. INTRODUCTION

Matching is a linkage of records from two or more files containing units from the same population. Two methods of performing such a linkage are exact and statistical matching. For exact matching, the variables linked were observed from the same unit. This technique requires unique identifiers such as name, address and social security number (Barry 1988). However, in most cases exact matching is not possible, either because there is no overlap, i.e. no individual, household or firm can be found in more than one data set, or there is no unique identification variable, or the use of this identifier is prohibited in order to protect personal privacy (Klevmarken 1983). In cases where the information available is not enough for an exact match, statistical matches may be constructed instead. Statistical matching links records that are similar, but do not

[1] First author is an associate professor of Statistics and the rest are BS Statistics graduates, all from the Institute of Statistics, University of the Philippines, Los Baños, Laguna. Email address of first author: zvjalbacea@yahoo.com

necessarily belong to the same unit. As mentioned in Cassel (1983), statistical matching is making the best of what is available in order to create artificial but hopefully realistic relations between variables not observed in both data sources. Statistical matching, however, is a risky process that needs to be conducted with caution.

Statistical matching techniques can be broadly classified into two types, namely: the distance-function statistical matching and random-draw statistical matching. Distance-function statistical matching is probably the most commonly used method. This technique involves calculating discrepancies or 'distances' between the values of matching variables in the two source files. The least distance of matching variables between the receiver and the donor file is considered when matching. When each donor record can be matched on an unlimited number of times, the procedure is known as the unconstrained matching. The alternative, which has usually proven more accurate in simulations, is constrained matching. With this method, each donor may be used as a match only once, after which it is removed from the donor file. If there are several matching variables measured in different units or on different scales, whatever form of distance function is used, the matching variable distances should be standardized by some form of weighting. In practice, the choice of weighing has been almost arbitrary and no optimum procedure has been found.

Vast majority of statistical matching work has been in the field of economics. The Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce did a constrained statistical match between the March 1965 Income Supplement of the Current Population Survey (CPS) and the individual tax returns. The purpose of the match was the improvement of the accuracy of CPS income amount and the addition of the details from the income tax return to the CPS observations. An unconstrained statistical matching was also carried out by BEA in 1969 between the 1964 CPS-TM file and the Survey of Financial Characteristics of Consumers (SFCC). Statistics Canada also performed an unconstrained matching, the SCF-FEX Match, between two Canadian microdata files, the 1970 Survey of Consumer Finances (SCF) and the 1970 Family Expenditure Survey (FEX). Its purpose is the addition of expenditure data to the SCF. The Office of Tax Analysis (OTA) of the U.S. Department of Treasury performed a statistical matching that is a logical extension of the constrained method first used by BEA. A more recent statistical matching was done by Yoshizoe and Araki (1999) in Japan. It was carried out between the October 1995 Family Income and Expenditure Survey (base file) and the 1995 Family Savings Survey (reference file). A portion of the two files has the same households such that exact matching was done prior to the statistical matching of the files. The exact matched data set was used to evaluate the properties of the statistically matched data sets.

Little is known about the nature and extent of the errors present in a data set resulting from a statistical match. Most of the literature on statistical matching consists of descriptions of matches performed with little evidence presented on the errors in the matched results. This is basically because these errors are difficult to estimate. Thus, given what is known at this time, statistical matching is not a satisfactory substitute for exact match in most cases. A substantial amount of research on statistical matching is needed, regarding both the optimal methods of matching and estimation of errors present in the matched results. (Federal Committee on Statistical Methodology, 1980).

With the aim of contributing to the study of statistical matching, this paper reports the results of doing statistical matching on 1997 Family Income and Expenditure Survey and the 1998 Annual Poverty Indicators Survey for Southern Tagalog Region in the Philippines. But unlike other studies, this paper empirically evaluates the properties of the distribution of the distances measured by applying the distance function in statistical matching.

II. METHODOLOGY

The Family Income and Expenditure Survey (FIES) provides information on the levels of living and disparities in income among various groups of Filipino families. It gives a general idea of the spending patterns of families belonging to different income groups. Another nationwide survey is the Annual Poverty Indicators Survey (APIS). It is designed to provide access and impact indicators that can be used as inputs to the development of an integrated poverty indicator and monitoring system for the assessment of the government program on poverty alleviation. Specifically, the survey aims to identify the poor families through the use of non-income indicators.

The base file is the 1997 FIES and the reference file is the 1998 APIS. Both data files are data from Southern Tagalog Region only. The unit of observation considered in FIES is the household and specific information on the household head is included in the survey. These are the age as of last birthday, gender, marital status and highest grade completed. All of these variables are also included in the APIS. The geographic variables that are common to both data files include: (1) province; (2) municipality; (3) urbanity; (4) barangay; (5) barangay stratum number; (6) household control number, and (7) enumeration area. These geographic variables and the other common variables in the two surveys comprise the set of possible matching variables.

Some of the households included in the two surveys are the same so that exact matching is possible. The calculated overlap between the two data sets is at most 75%. The number of records in the FIES and APIS data sets, are 6162 and 6063, respectively. These two data sets were exactly matched using the geographic characteristics of the household, namely; province, municipality, barangay, enumeration area, urbanity, stratum, and household control number. In addition, the interviewer status of the household and the characteristics of the household head such as gender, marital status, and age were also used in the matching. All of which except for age and interviewer status should be equal in both files. To be considered an exact match, the interviewer status should be equal to 1, which indicates that a particular household is the same sample household in both surveys. In the case of age, a household head in APIS should have the same or one year older than the one reported in the FIES since APIS was conducted a year after FIES. The generated data set is denoted as EXM data set.

The remaining FIES and APIS records after excluding the exactly matched records were subjected to statistical matching. To study the performance of the distance function used, four statistically matched data sets were created using constrained statistical matching. Four matching variables; household head's gender (coded: 1 = Male, 2 = Female), marital status (coded: 1 = Single, 2 = Married, 3 = Widowed, 4 = Divorced or Separated, 5 = Others), age, and family size, were used.

One data set created, denoted as STM_A, used the following form of the distance function

$$d(x_i, x_j) = \frac{(x_{i,A} - x_{j,A})^2}{s_A^2} + \frac{(x_{i,F} - x_{j,F})^2}{s_F^2} + \frac{(x_{i,SX} - x_{j,SX})^2}{s_{sx}^2} + \frac{(x_{i,M} - x_{j,M})^2}{s_M^2}$$

where s_F^2, s_A^2, s_{sx}^2 and s_M^2 are the variances of family size and household head's age, sex, and marital status in FIES, respectively. Another form of the distance function that was used to generate a statistical matched data set is the absolute difference of the matching variables instead of the square of the difference. The resulting matched data set is denoted as STM_B. A form of the distance function used in Cassel (1983), given as:

$$d(x_i, x_j) = \lambda_A (x_{i,A} - x_{j,A})^2 + \lambda_F (x_{i,F} - x_{j,F})^2 + \lambda_{SX} (x_{i,SX} - x_{j,SX})^2 + \lambda_M (x_{i,M} - x_{j,M})^2$$

where $\lambda_A, \lambda_F, \lambda_{SX}$ and λ_M are weights used for variables age, family size, sex and marital status, respectively, was used to create a data set denoted as STM_C_WE. Several weighing schemes were tried in this form of the distance function. Arbitrarily, the weights are set as λ_A = 0.15, λ_F = 0.15, λ_{SX} = 0.35 and λ_M = 0.35. In a preliminary analysis, this set of weights gives the most number of zeros in the calculated distances. The form of the distance function of Cassel (1983) without weights was also used to generate the fourth statistically matched data set, denoted as STM_C_UN.

The properties of the distribution of the distances measured using four forms of the distance function in creating constrained statistically matched data set were evaluated. Based on these properties, the "best" form of the distance function was chosen. The "best" form is the one that resulted to distances, which are mostly zeroes or near zero. Also, the "best" form will preferably result to a unit less measure of distance. The data set created with the "best" form of distance function will then be renamed as CON_STM. Using the same form of distance function, another data set was generated but this time using the unconstrained statistical matching procedure. The generated data set will be named as UNC_STM.

The characteristics of the data sets created by statistical matching (UNC_STM and CON_STM) were compared with those of the exactly matched data set (EXM). The comparison was made using the characteristics of the test variables. The test variables identified include total income, total expenditure and national income decile group. Likewise, the differences between the values of these test variables in FIES and in APIS were used in the comparison. Descriptive statistics were computed and scatter diagrams and histograms of these variables were also constructed and compared. In addition, comparison of the mean values of the test variables in the matched data sets was also conducted.

III. RESULTS AND DISCUSSION

From the original sizes of FIES and APIS, which are equal to 6150 and 6045, respectively (records with missing values excluded), it was reduced to 3076 and 2971 after exact matched data set was deleted. The total number of exactly matched records is 3074. With the calculated overlap of at most 75%, only 50% was attained. This may be due to some unsuccessful interviews and replacements of some households.

There are four (4) variables used in the matching procedure. Table 1 shows the descriptive statistics of these matching variables as observed in FIES and in APIS.

Table 1. Descriptive Statistics of Matching Variables in FIES and APIS Data Sets

Variable	Mean	Std Dev.	Range	Median	$Q_3 - Q_1$
FIES data set					
Sex	1.2	0.4	1	1	0
Marital Status	2.2	0.6	4	2	0
Age	47.3	14.8	82	45	21
Family Size	5.0	2.4	18	5	3
APIS data set					
Sex	1.2	0.4	1	1	0
Marital Status	2.2	0.5	4	2	0
Age	46.9	14.6	87	45	22
Family Size	4.8	2.2	15	5	3

Comparing these two sets of statistics, the medians of all the matching variables are the same for both data sets. The mean, as another measure of center of the distribution are the same in 3 matching variables except in the variable age of the household head with a difference of 0.4. Likewise the dispersion of the values in both data sets are almost the same as indicated by the similarities in the values of the standard deviation, range, and interquartile range.

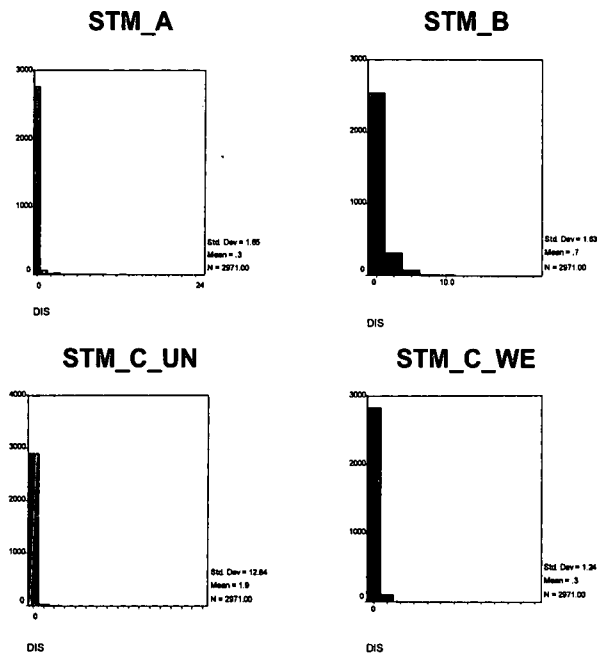
These four matching variables were used in the distance function for matching. The form of the distance function used in generating STM_A resulted to a unit less measured distance. Table 2 indicates similarities in the distances measured. It clearly shows that all distances measured are small and positively skewed. Almost 75% of the distances calculated are at most one in all matched data sets. This indicates closeness of the values of the matching variables as shown by values of distances mostly equal to zero. All statistically matched data sets have the same median, mode, minimum value and first quartile, which are all equal to zero. It was also observed that the distributions are positively skewed because of the presence of some extreme high values. STM_C_WE has the lowest mean and standard deviation with 0.2751 and 1.2359, respectively. On the other hand, STM_C_UN has the highest mean and standard deviation of 1.8859 and 12.555, respectively. The maximum distance of STM_C_UN that is equal to 405 was observed. This is partially expected since there is no weighing scheme incorporated in this form of the distance function. Nevertheless, it does not necessary lead to poor results compared to the form of the distance function that uses weights like the one that was used in statistically matched data set STM_A which shows relatively small mean and standard deviation.

Table 2. Descriptive Statistics of the Distances Measured in STM_A, STM_B, STM_C_UN and STM_C_WE data sets.

	Mean	Median	Mode	Std. Dev.	Skewness	Min	Max	Q1	Q3
STM_A	0.3360	0	0	1.6533	7.1700	0	25.33	0	0.004
STM_B	0.7405	0	0	1.6310	5.6419	0	23.00	0	1.000
STM_C_UN	1.8859	0	0	12.8448	20.5551	0	405.00	0	1.000
STM_C_WE	0.2751	0	0	1.2359	11.1180	0	24.65	0	0.150

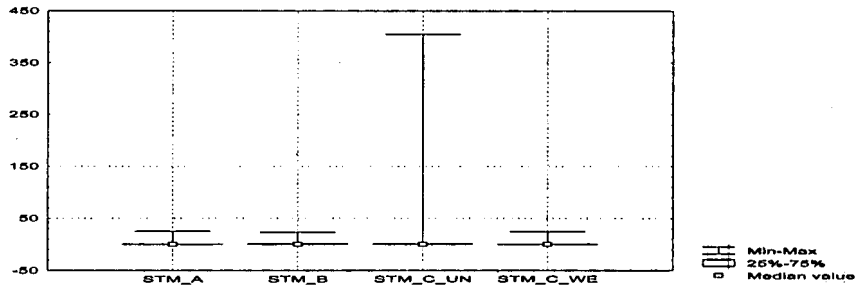
Figure 1 shows that the distances measured in all statistically matched files are positively skewed as noted in Table 2. Histograms of measured distances of STM_A, STM_B, STM_C_UN and STM_C_WE show that the measured distances cluster near zero. In the generated statistically matched data sets, there were at least 2500 records that have distances measured less than or equal to one. This indicates the closeness of the values of the matching variables in the statistically matched data sets.

Figure 1. Histograms of Distances Measured in STM_A, STM_B, STM_C_UN and STM_C_WE data sets



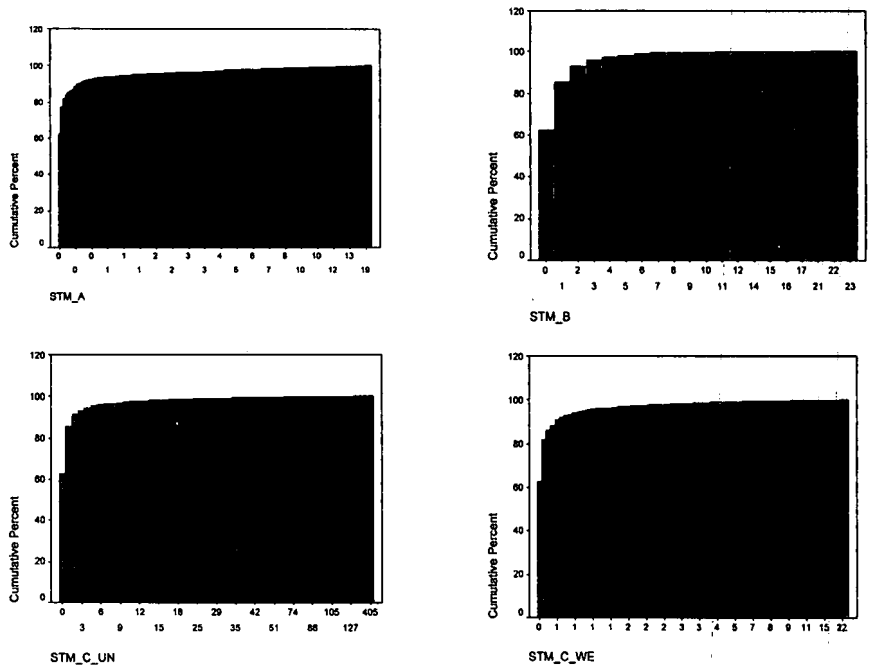
In addition to the observed characteristics of the histograms of distances, Figure 2 showing the box and whisker plots exhibit that the distances measured in STM_A, STM_B, STM_C_UN and STM_WE cluster near zero. This indeed shows the closeness of the values of the matching variables in the statistically matched data sets. An extreme value in the STM_C_UN where a maximum distance of 405 was observed indicating a wide range of values. It was also determined that these extreme values of the distances were observed when there are few or no more records to choose from to make a match.

Figure 2. Box and Whisker Plots of Distances Measured in STM_A, STM_B, STM_C_UN and STM_C_WE data sets



Cumulative percentage distributions of distances measured in STM_A, STM_B, STM_C_UN and STM_C_WE are presented in Figure 3. In the statistically matched data sets STM_A, STM_B, STM_C_UN and STM_C_WE, 94.95%, 85.26%, 85.39% and 95.32%, respectively, of the records have distances at most equal to one. The distribution shows that at some point almost 95% of the data are already matched with small calculated-distances.

Figure 3. Cumulative Distribution of Distances Measured in STM_A, STM_B, STM_C_UN and STM_C_WE data sets



The individual contribution of each of the matching variables, namely: age, sex, marital status and family size, to the generated distances were obtained. The statistics of the contribution of each matching variable in the measured distances is presented in Table 3. It can be observed that family size has the largest mean contribution in the

distances measured in the four forms of the distance function used in statistical matching procedure.

Table 3. Contribution of the Matching Variables to the Distances by the Measured Distance Function

	Mean	Median	Mode	Std. Dev.	Skewness	Minimum	Maximum	Q1	Q3
STM_A									
AGE	0.05	0	0	0.31	11.41	0	6.28	0	0.01
SEX	0.07	0	0	0.65	9.48	0	6.26	0	0
MS	0.08	0	0	0.69	13.53	0	13.20	0	0
FSIZE	0.14	0	0	0.89	14.08	0	25.31	0	0
STM_B									
AGE	0.28	0	0	0.93	9.41	0	18.00	0	0
SEX	0.07	0	0	0.25	3.43	0	1.00	0	0
MS	0.10	0	0	0.35	3.76	0	3.00	0	0
FSIZE	0.29	0	0	0.81	4.48	0	10.00	0	0
STM_C_UN									
AGE	0.75	0	0	9.37	33.73	0	400.00	0	0
SEX	0.08	0	0	0.27	3.15	0	1.00	0	0
MS	0.14	0	0	0.56	7.57	0	9.00	0	0
FSIZE	0.91	0	0	7.37	24.73	0	289.00	0	0
STM_C_WE									
AGE	0.10	0	0	0.76	19.00	0	21.60	0	0
SEX	0.02	0	0	0.08	4.09	0	0.35	0	0
MS	0.03	0	0	0.14	6.72	0	1.40	0	0
FSIZE	0.12	0	0	0.69	11.54	0	15.00	0	0

Among the four forms of the distance function studied, the form given as the squared differences and standardized by its variance, as proposed by Yoshizoe and Araki (1999) and used in the statistical match data set labeled STM_A did perform well in getting the least measured distances. In addition, this form of distance function generates a unit less measure of distances. Out of 2971 matched records, 2821 records were matched with measured distance which is at most equal to one. In other words, 94.95% of the matched records in STM_A have calculated distances at most equal to one. Thus, this form is considered the "best" form of the distance function to use in statistically matching the two data sets.

STM_A data set was then renamed as CON_STM and using the "best" form of the distance function (the form of distance function used to generate STM_A), another statistically matched data set was created using unconstrained statistical matching and named as UNC_STM. Table 4 shows the descriptive statistics of the differences of the test variables as observed in matched data sets. In the exact matched data set, the values of total income, total expenditure and national income decile from the APIS records are lower than the values from the FIES records, that is why the differences are mostly negative. On the average, the household's total income, total expenditure, and national

income decile decreased by P69,314.95, P55,815.19 and 0.03, respectively, from 1997 to 1998.

Table 4. Descriptive Statistics of Test Variables in the EXM, UNC_STM, and CON_STM Data Sets

Variable	Mean	Std Dev.	Min	Max	Q_1	Median	Q_3	Skewness
Exact matched data file								
dif_inc	-69314.95	94870.3	-1670500	299700	-90190	-46501	-22937	-5.3
dif_exp	-55815.29	62953.5	-801679	382244	- 72296	-40826	-21162	-3.1
dif_nid	-0.03	1.8	-9	8	-1	0	1	-0.3
Unconstrained statistically matched data file								
dif_inc	-75329.58	185372.6	-5246320	1826960	-115185	-46552	-3847	-8.7
dif_exp	-56292.23	100466.6	-1249642	449050	-88068	-38605	-4619	-2.4
dif_nid	0.07	3.4	-9	9	-2	0	2	0.0
Constrained statistically matched data file								
dif_inc	-74260.04	195803.5	-5303840	2345730	-115352	-46480	-3382	-7.2
dif_exp	-56866.80	101746.7	-1222295	527154	-89570	-38652	-4000	-2.2
dif_nid	0.02	3.5	-9	9	-2	0	2	0.0

The difference in total income (*dif_inc*), total expenditure (*dif_exp*), and national income decile (*dif_nid*) between the APIS and FIES are negatively skewed. Almost 75% differences in total income and total expenditure between APIS and FIES are negative as indicated by the negative values of the 3rd Quartile. The maximum value of 8 in the difference in national income decile between APIS and FIES records indicates that there is an extreme disagreement in the values of national income decile. This maximum value was obtained from households with a value of 10 in national income decile for APIS while in FIES it has a value of 1.These households were separated and further studied. It was observed that on the average, the increase in the total income of these households is P154,441.20 which resulted to an increase in their classification based on the national income decile groupings. On the other hand, there are also households who had a decrease in their total income that resulted to a minimum value of –9 in the difference in national income decile.

In the data set generated using the unconstrained statistical matching procedure, the total income and total expenditure decreased as indicated by the negative average difference. However, the difference in national income decile increased by 0.07. The difference in national income deciles is symmetric while the differences in total income and expenditure are negatively skewed. The same observations hold for the distribution of differences generated using constrained statistical matching procedure. The differences of the test variables with negatively skewed distributions have at most 75% of its values negative. Likewise, the existence of an extreme disagreement in the values of national income decile between the APIS and FIES in some households as in the EXM data set was also observed in both statistically matched data sets.

Figure 4 shows the histograms of the differences of the test variables in the three generated data sets, namely; EXM, UNC_STM, and CON_STM. The difference in total income and total expenditure between the APIS and FIES records cluster near zero. Most of the values are negative indicating a decrease in total income and expenditure from 1997 to 1998. Such distributions are the same across the three generated data sets. However, the histograms of the difference in national income decile of the statistically matched data sets are different from that of the exact matched data set. The histogram in the exact matched data set is negatively skewed while those in statistically matched data sets are said to be symmetric.

Figure 4. Histograms of Test Variables in the Exact and Statistically Matched Data sets

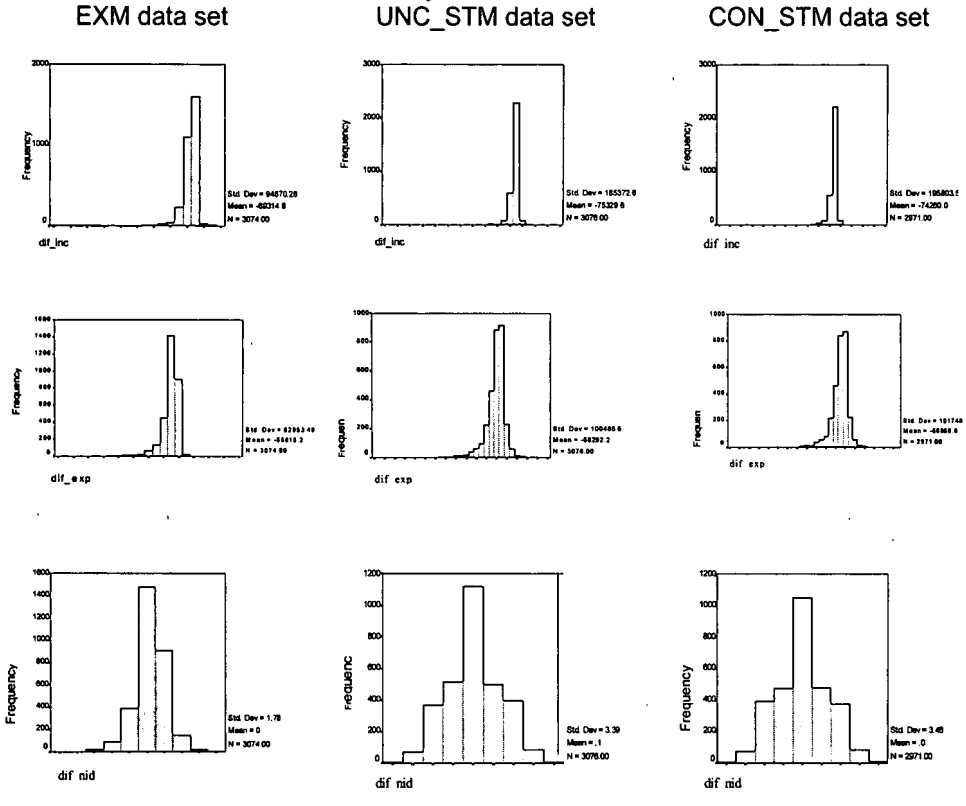


Table 5 shows that the means of the test variables between the exact and unconstrained statistically matched data sets are not significantly different. Also, the exact and constrained statistically matched data sets showed no significant difference when the means of the test variables were compared.

Table 5. P-values in the Tests used to Compare Exactly Matched Data Set to Each of the Statistically Matched Data Set

VARIABLE	EXM_UNC		EXM_CON	
	Pr > F'	Pr > |T|	Pr > F'	Pr > |T|
Dif_inc	0.0000**	0.1093	0.0000**	0.2140
Dif_exp	0.0000**	0.8234	0.0000**	0.6303
Dif_nid	0.0000**	0.1224	0.0000**	0.4587

Note: i. EXM_UNC is the comparison between EXM and UNC_STM; EXM_CON is the comparison between EXM and CON_STM,
ii. 1st column under each comparison refers to the test on equality of variances while the 2nd column to the equality of means

Table 6 shows the descriptive statistics of the computed distances obtained in statistically matched data sets. In unconstrained statistically matched data, the average computed distance is 0.03 with a standard deviation of 0.2. Its maximum value is 6 while most of its values are zero. On the other hand, the computed distance in constrained

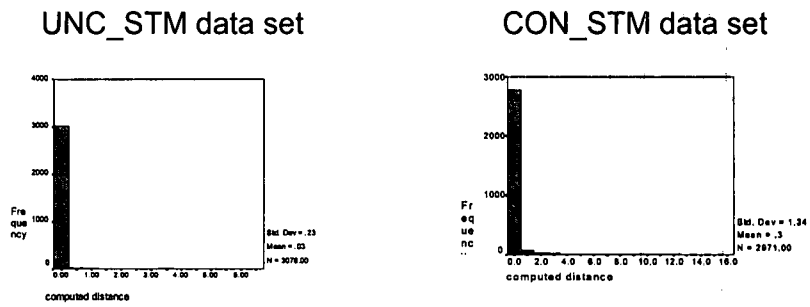
statistically matched data set has an average value of 0.29 with a standard deviation of 1.3. Its maximum value is 16 and likewise almost 75% of its values are zero.

Table 6. Descriptive Statistics of Computed Distances in the UNC_STM, and CON_STM Data Sets

Data Set	Mean	Std. Dev.	Min	Max	Q_1	Median	Q_3	Skewness
UNC_STM	0.03	0.2	0	6	0	0	0	17.3
CON_STM	0.29	1.3	0	16	0	0	0	6.3

Looking closely on its distribution, Figure 5 shows that more than 90% of its distribution has values equal to zero and the remaining values are mostly 1, with very few extreme values. Both distributions are negatively skewed.

Figure 5. Histogram of the Computed Distances in the Statistically Matched Data Sets



Comparing the means of the test variables in the two statistically matched files, the test showed no significant difference (Table 7). However, in the mean calculated distance (*dis*) the difference is said to be significant. The mean calculated distance is significantly lower in UNC_STM than in CON_STM.

Table 7. Probability of Significance in the Tests used to Compare Two Statistically Matched Data Sets

VARIABLE	CON_UNC	
	Pr > F'	Pr > | T |
dif_inc	0.0026**	0.8275
dif_exp	0.4683	0.8251
dif_nid	0.1249	0.5473
Dis	0.0000**	0.0001**

Note: i. CON_UNC is the comparison between CON_STM and UNC_STM,
ii. 1st column under each comparison refers to the test on equal variances while the 2nd column to the equality of means

V. SUMMARY AND CONCLUSION

The objective of this paper is to present an empirical evaluation of statistical matching as a method of creating panel data sets. Statistical matching links records that are similar, but do not necessarily belong to the same unit. Statistical matching techniques can be broadly classified into two types, the distance-function statistical matching and random-draw statistical matching. Distance-function statistical matching is probably the most commonly used method. This technique involves calculating discrepancies or 'distances' between the values of matching variables in the two source files.

In particular, the 1997 Family Income and Expenditure Survey (FIES) and 1998 Annual Poverty Indicators Survey (APIS) data in Southern Tagalog Region, Philippines were statistically matched after removing the exact matched records. Also, distance-function statistical matching is employed in this study. To evaluate the "best" form of the distance function, four statistically matched data sets are obtained (STM_A, STM_B, STM_C_UN, STM_C_WE). Descriptive statistics, histograms and, box and whisker plots of measured distances show closeness of the values of the matching variables in the performed statistical matching. The cumulative percentage distributions show that at some point almost 95% of the data are already matched having small calculated-distances. The results indicate that the distribution of the distances measured using the form of the distance function used in generating STM_A clusters near zero. Also, 94.95% of the records were matched with distances at most equal to one. Aside from the fact that this form of the distance function gives a unit less measure, it is also easy to use. It was also found that a stopping rule in the algorithm of the statistical matching procedure enhances the performance of the procedure in generating data sets. Such stopping rule depends on the percentage of distances measured with at most equal to one.

Using the "best" form of the distance function, unconstrained and constrained statistical matching procedures were used to generate the unconstrained (UNC_STM) and constrained statistical matched file (CON_STM), respectively. Results showed that the descriptive statistics of the test variables in UNC_STM, CON_STM and EXM files are statistically the same.

The distances measured in generating statistically matched data sets are mostly zero indicating similarities. However, the histograms of distance values calculated in UNC_STM and CON_STM are slightly different. It showed that there were fewer distances calculated greater than 1 in UNC_STM compared to CON_STM.

From the evaluated simple statistics and test on means, statistical matching was shown to be relatively a feasible method to extract fuller information out of FIES 1997 and APIS 1998. This is because the distribution of the variables were preserved in both statistically matched files and the result of the test on means show that means of the test variables are not significantly different between the exact matched file and the two statistically matched files. Unconstrained statistical matching was found to have a lower mean calculated distance than constrained statistical matching.

Acknowledgements

The authors are grateful to the National Statistics Office (NSO) and the Asian Development Bank for allowing them to use part of the data sets that were used in a project funded by the ADB through its Technical Assistance 3656 PHI: Improving Poverty Monitoring Surveys. Also, the authors are grateful to the project proponents, Dr. Ann Inez N. Gironella of INSTAT, UPLB and Dr. Eliezer A. Albacea of ICS, UPLB for inspiring them to pursue this study. Likewise, the authors are grateful for the comments given by anonymous external reviewer of this paper.

Specifically, the authors appreciated the suggested alternative method of finding the "best" distance function given by the external reviewer. This valuable suggestion can be pursued in future studies regarding statistical matching.

References

BARRY, J.T. An Investigation of Statistical Matching. Journal of Applied Statistics. 1988. 15:3:275-283.111

CASSEL, C.M. Statistical Matching - Statistical Prediction. What is the Difference? Statistical Review 1983:5 pp. 55-66.

GIRONELLA, G.M.P. Unconstrained Statistical Matching of 1997 FIES and 1998 APIS Region 4 Data Sets, Unpublished Undergraduate Special Problem, INSTAT, UPLB, 2002

KLEVMARKEN, N.A. Pooling Incomplete Data Sets. Statistical Review 1983:5 pp. 67-79.

MOGUL, S.K.T. Distributional Properties of Distances Measured in a Statistical Match, Unpublished Undergraduate Special Problem, INSTAT, UPLB, 2002

PAASS, G. Statistical Record Linkage Methodology. 45th Session of the International Statistical Institute. pp. 33-48.

STERN, S. 2000. Valuing Housing Subsidies: A Revised Method for Quantifying Benefits in a New Measure of Poverty. Prepared for the Annual Conference of the American Statistical Association.
www.census.gov/hhes/poverty/povmeas/papers/jsm00.pdf

Subcommittee on Matching Techniques. Federal Committee on Statistical Methodology. 1980. Report on Exact and Statistical Matching Techniques. http://www.fcsm.gov/working-papers/wp5.html.

TENA, J.K.S. Constrained Statistical Matching of 1997 FIES and 1998 APIS Region 4 Data Sets, Unpublished Undergraduate Special Problem, INSTAT, UPLB, 2002

YOSHIZOE, Y. and M. ARAKI. Statistical Matching of Household Survey Files in Japan. 1999. pp. 1-9.