MEMORANDUM

FROM: John O'Hare
TO: Qiang Ma
DATE: January 29th, 2010
June 28th, 2010 (Revised)
SUBJECT: Statistical Matching Documentation

This memorandum documents the procedures, programs and results from performing a constrained statistical match between the 2004 SOI Public Use File (PUF) and the March 2005 Current Population Survey (CPS). First, we provide an overview of statistical matching and how this relates to the specific challenges of combining information from the SOI and CPS. Second, we describe the SAS programs we developed in the course of implementing the match. Finally, we present the results of the match performed in November 2009.

Overview of Statistical Matching

In the standard statistical matching framework, one has observations from two data sets (File A and File B) on a set of common variables (X-variables). Additionally, records from File A contain information on another set of variables (Y-Variables) that are not available on File B. Similarly, File B contains information on a third set of variables (Z-Variables) that are not available on File A. Statistical matching involves creating a new data set (File C) containing information on X, Y and Z.

Statistically matched data sets are used most extensively as inputs to microsimulation models [Cohen(1991)] where the impacts of a policy change are examined across various subgroups of the population where data limitations are often severe. For example, the SOI contains a great deal of detail on taxable income sources and certain expenditures (e.g., health-related costs and home mortgage interest expenses for those taxpayers who itemize) but very little information on transfer payments (e.g., food stamps and AFDC), employment-related fringe benefits (health insurance) and family composition. An ideal data set would combine information from both sources.

From an historical perspective, the first statistically matched data file for use in microsimulation was created for tax policy analysis [Okner (1972)] and combined the SOI with the Survey of Economic Opportunity (SEO). An up-to-date survey of statistical matching methods and a description of how they are implemented in applied work is Cohen (1992).

In this project, we update and extend earlier SOI-CPS matches. We have three goals:

- Construct a statistically matched data set combining the 2004 SOI with the March 2005 CPS that can be a reliable input to tax policy analysis.

- Introduce methods that ease the computational burden of (constrained) statistical matching, are straightforward to implement and have desirable statistical properties.
- Provide documented computer algorithms and source code so the matches can be updated on a regular basis with minimal effort.

Constrained vs. Unconstrained Statistical Matching

Most authors distinguish between two types of statistical matching methods based on how the records in both files are combined. In constrained matching, all of the records in both data files are represented in the final matched data set (File C). In order to achieve this result, there are limits placed on the number of times a particular record in File B can be matched to a record in File A. Records on both files are often "split", or used more than once, in a constrained match and the limits are enforced by making sure that the population weight attached to each record is "used-up" in the match. A necessary condition for performing a constrained match, but one that is difficult to meet in practice, is that both input files have the same weighted population totals.[1]

Unconstrained matching does not require that all of the records in File B be used up the match, but it is almost always the case that each record in the Host file appears in the final matched file. In practice, certain limits are usually imposed on the number of times a Donor (File B) record can be used in an unconstrained match to ensure that the (weighted) distributions of the Z-variables "brought over" in the match are closely aligned with the distributions on the original file.

Experience suggests that, of the two, unconstrained matching is the most popular choice for constructing microsimulation data sets since it makes fewer demands on system resources. On the other hand, there are advantages to be gained in constrained matching and with the advent of more powerful computers, cheaper memory and faster numerical algorithms, this method has become increasingly more common.

A common criticism of statistical matching is that it relies on rather strong assumptions about the relationship between the Y- and Z-variables [Kadane (1978)]. In particular, an implicit assumption in statistical matching is that they are independent (or uncorrelated if normality is assumed) given an observation of X-variables. This Conditional Independence Assumption (CIA) is often violated in practice and has caused researchers to investigate alternative methods of combining data sets [Rubin (1986); Armstrong (1989); Singh et. al. (1993)].

Unconstrained Matching

Unconstrained statistical matching is probably the most popular type of matching being done today, at least with regard to microsimulation applications. Various approaches to

[1] In applied work, it is often the case that the two input data sources are from surveys taken over different time frames so that the weighted population totals are slightly different across the two files. In this case, it is a common procedure to "scale" one of the files (usually the Donor file) so that the weighted population totals agree.

unconstrained statistical matching are comparatively simple to implement, cost effective, easy to replicate and update, and intuitive.

In unconstrained matching, one seeks a record in File B that resembles or is in some sense "close" to a record in File A. This presumes some sort of metric or distance function is introduced. One commonly used metric, the Euclidean distance normalized by the standard deviation, appears frequently in the literature.

Let X^A_{i1}, X^A_{i2}, ... X^A_{in} denote a set of X-variables from record i in File A used to construct the distance function and X^B_{j1}, X^B_{j2}, ..., X^B_{jn} be similar values of the same variables from record j in File B. Then the distance function is:

$$d_{ij} = [\Sigma_k ((\ X^A_{ik} - \ X^B_{jk}\ ) / \sigma_k)^2\ ]^{1/2}$$

where σ_k is the standard deviation of the kth X-variable in File A. It should be emphasized that the choice of an appropriate distance function can have important consequences for the integrity of the matched dataset [Paass(1985)].

Minimum distance matching (or "nearest neighbor" matching) was probably the first large-scale statistical matching method to be performed for use in a microsimulation environment [see Okner (1972)]. It's also very easy to describe: for each record in File A, select the record in File B that is "closest" in a minimum distance sense. Like all unconstrained matches, minimum distance matching suffers from the fact that the marginal distributions of the Z-variables in the matched file could be quite different than on the original file and it is important to check the validity of the results. In a sense, most unconstrained methods represent refinements of this procedure [Armstrong (1989)].

Constrained Matching

Statistically matched data sets constructed by constrained matching have nice properties: means and variances of the X, Y and Z variables in both input files are preserved as a direct consequence of the constraints imposed on the weights occurring in the final matched data set. One drawback of constrained matching, however, is that records may end up being matched with an unacceptably large distance between the X-variables. Constrained matching can also make very large demands on system resources. Barr and Turner (1979, 1980) provide a concise and useful summary of the technical details of constrained matching as well as a discussion of many of the practical problems one encounters in applied work.

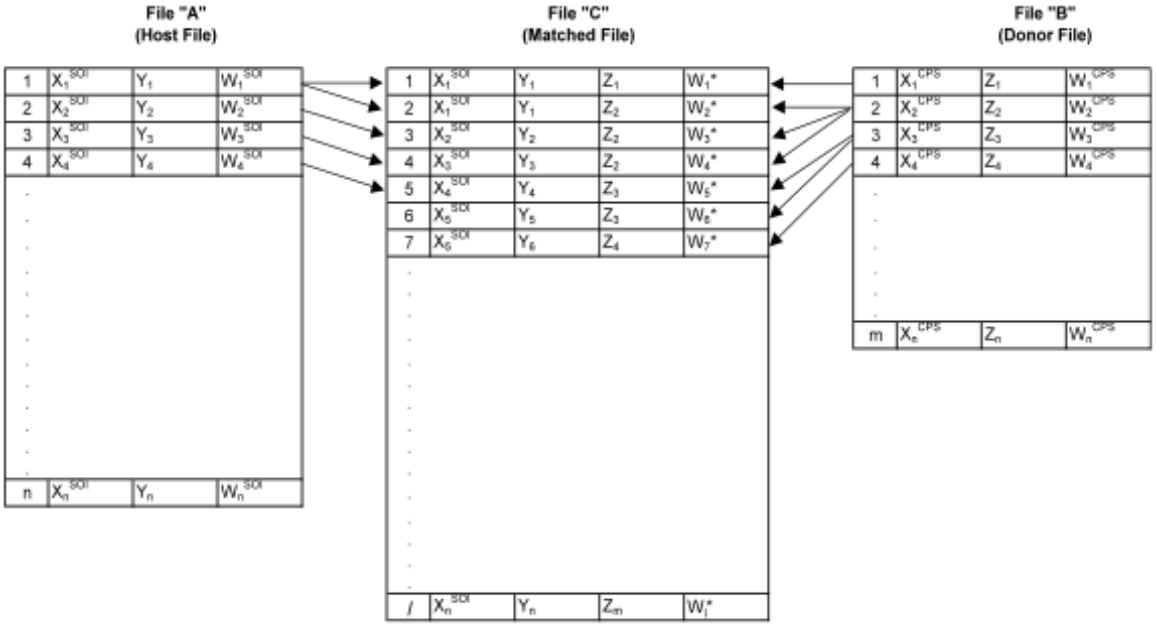
In some implementations of fully constrained matching, one seeks to minimize the overall distance between records in File A and File B that appear in File C. Mathematically, one attempts to minimize:

$$\Sigma_i \Sigma_j (d_{ij} * w_{ij})$$

subject to: $\Sigma_j w_{ij} = a_i$ and $\Sigma_i w_{ij} = b_j$

where a_i and b_j are the original weights on File A and File B and w_{ij} is the weight on File C obtained from matching the ith record on File A with the jth record on File B.[2] This problem has the structure of the transportation (or transshipment) problem in network optimization theory. (Think of the records in File A as the "factories" (sources) and the records in File B as the "warehouses" (sinks). The constraints ensure all of the product (e.g., the weights) gets shipped. Bertsekas (1991) contains a nice discussion of the transportation problem. Figure 1 shows schematically how record are combined in constrained statistical matching.

Figure 1 – Constrained Statistical Matching



Paass (1985) and Rodgers (1984) favor constrained matching over unconstrained, nearest-neighbor methods. They believe that the additional cost (in terms of resources) are justified and outweigh the possibility of error introduced by unconstrained matching.

Methodological Approach

The approach we will implement for this project is referred to as fully constrained, predictive mean matching. Constraining the match to utilize all records on both files ensures that the marginal distributions of the Z-variables will be carried over onto the matched file.[3] This is a desirable result. Predictive mean matching is a technique by

[2] Notice that this implies that the sum of the weights on File A must equal the sum of the weights on File B.
[3] This is only the case when the weighted totals are the same on both files. Partitioning of the input files can further distort the marginal distribution of the Z's.

which the correlation between the X-, Y- and Z- variables is exploited in the match by using predicted values to determine the nearest-neighbor match.[4] Predictive mean matching performs well in practice [Armstrong (1989)] and is straightforward to implement.

We employ an extensive partitioning scheme on both input files to create equivalence classes where matches are only allowed within these classes. This procedure has the effect of narrowing the distance between records and allows for a tighter fit across the two data sets.[5] A more detailed description of the procedure is included as an attachment to this document.

Results

In the first step of the match, we construct tax units from the CPS file. We impose the constraint that there should be at least one tax unit in every household. We combine information from the household, family and person records to construct each tax unit. A summary of the March 2005 CPS records is contained in Tables 1A through 1D in Appendix A.

Once person records in each household are combined to form tax units, we make an initial determination of whether that tax unit is legally required to file a federal income tax return based on filing thresholds in place for the Tax Year 2004, the most recent year for which a Public Use File (PUF) is available.[6] In this match, we used a slightly different methodology for choosing tax filers. In prior matches, we adjusted filing thresholds to get the approximate number of returns in broad, taxpayer classes. We modified this approach in the current match to more closely approximate what is done in the Tax Policy Center's (TPC) model. That is, we randomly selected tax units that were initially determined to not be required to file a tax return and reclassified them as a tax filer. We did this across all taxpayer subgroups. Weighted and unweighted counts of CPS tax units are shown in Tables 2A through 2D. Similar counts from the 2004 PUF are shown in Tables 3A and 3B.

In the next step, we partition both files across several dimensions to ensure that matches are not done outside of a particular cell. We construct 18 cells to reflect: (i) dependency status, (ii) marital status, (iii) age, and (iv) the number of dependents. Our partition and the corresponding weighted and unweighted record counts are shown in Table 4.

In our third step, we construct a distance metric by regressing taxable income on the PUF to a collection of independent variables that we believe are related to taxable income.

---

[4] The term was apparently first coined by Little (1986).

[5] A partitioning scheme can also be implicitly "induced" by the careful construction of a distance metric. Imposing an arbitrarily large penalty for matches between certain types of records can have much the same effect as partitioning. In this sense, partitioning is equivalent to assigning an infinite distance to matches between certain records.

[6] A PUF for tax year 2005 became available shortly after this project began, but because of differences in sample designs between the two years, we decided to rely on the 2004 PUF.

These variables include: a indicator if the taxpayer is 65 years of age or older, various income sources (e.g., wages, interest, dividends), the share of total income attributable to capital and labor, respectively, and indicator variables if the taxpayer received either wages or self-employed income. We estimate 18 separate regressions, one for each cell, and construct fitted values on the CPS using the same set of predictor variables. The distance between records on the CPS and the PUF is then calculated as the square root of the (squared) difference between the fitted values. The methodology for matching records between the files is described in Appendix B. The regression coefficients and other measures of goodness-of-fit are shown in Tables C1 through C18 in Appendix C.

In our final step, we evaluate the match by comparing means and standard errors for CPS variables added to the PUF on both the original CPS file and the final Match File. These results are reported in Table 5. The final column in Table 5 is a t-test to examine differences in means between the two files. (The null hypothesis is that the means are equal in the two files.) These results should be interpreted with caution as the values of the standard errors on the Match File are not adjusted for sample design.

Differences Between PwC Match and TPC Match

We are aware of only two differences in the methodology reported here and the methodology described in TPC's documentation.[7]

Partitioning. TPC expanded the number of cells in the partition to include indicators if the tax unit had self-employed income or income from capital. This resulted in a much larger number of cells (49) in the TPC match. One drawback of expanding the number of cells in this way is that certain cells become "unbalanced" and in the process of adjusting the weights on the CPS file to match the weights on the PUF (this is necessary to perform a constrained statistical match) certain CPS variables might not be indicative of their true values.[8] The trade-off between more cells and a potentially large adjustment is more an art than a science. In any event, adding more cells to the PwC match is a simple extension.

Non-filers. TPC relied on a series of probit equations reported by Cilke (1998)[9] to identify potential CPS tax units that were not legally required to file a federal income tax return but did. Cilke related the probability of filing a return (given that the taxpayer was not legally required to do so) as a function of income and numerous demographic variables. These probabilities were then adjusted to approximate control totals for certain taxpayer groups. Incorporating this methodology, while somewhat time-consuming, is straightforward. Because this would have necessitated rerunning the match to include many more demographic variables, we approximated this approach by randomly

[7] *The Urban-Brookings Tax Policy Center Microsimulation Model: Documentation and Methodology for Version 0304*, Rohaly et. al., January 10th, 2005.
[8] As measured by the ratio between the CPS and PUF weighted record counts, several CPS cells in the TPC match were on the order of 2.0 to 2.5 times the PUF value.
[9] *A Profile of Non-Filers*, Office of Tax Analysis Paper 78, U.S. Department of Treasury, Washington, DC.

selecting non-filers in specific taxpayer subgroups and re-classified them as filers so as to match SOI totals in that subgroup.

Description of the SAS Programs

In this project, we perform a constrained statistical match between the 2004 SOI Public Use File (PUF) and March 2005 Current Population Survey (CPS).[10] The PUF plays the role of the Host file while the CPS is the Donor. Before we can perform the match, however, we must make sure both files represent the same observational unit. Because the PUF is a sample of tax returns, an important early step in this process is to construct tax returns, or tax units, from the CPS. Not all tax units constructed this way from the CPS will necessarily be legally required to file and individual income tax return and therefore will not be represented on the SOI. Our algorithms determine which CPS tax units are likely to file an income tax return and these will comprise our Donor file (e.g. the "filers") for matching with the PUF. The remaining CPS tax units are not required to file an income tax return, but they are represented in our final matched data set (e.g., the "non-filers").

We perform the match by running seven (7) SAS programs in sequence. A brief description of these programs and their inputs and outputs follows.

CPS-PREP This program prepares the CPS for processing. Person-level information is combined with family and household information to construct a composite record for each member of a CPS household. A statistical summary of the output files is produced.

Input files: ASEC2005_PUBUSE.PUB (March 2005 CPS File in ASCII format.)

Output files: HOUSHLD (SAS File of CPS household information)
FAMPER (SAS File combining CPS Family and Person level detail.)

CPS-RETS Constructs CPS tax units from the composite files. Processes one household at a time; examines relationships among household members; and determines which members constitute a tax unit. Once a tax unit is constructed, we determine whether or not a tax return is required to be filed. We make this determination by examining the tax filing requirements contained in the IRS Instructions for tax year 2005.

Input files: HOUSHLD
FAMPER

[10] The CPS is conducted in March of every calendar year and questions relating to the annual income of respondents refer to the prior year's income. In contrast, the 2004 PUF reports income received during calendar year 2004.

Output files: CPSRETS (SAS dataset representing the tax filers. The Donor file.)
CPSNONF (SAS dataset representing the non-filers.)

SOI-RETS Creates an extract from the 2004 PUF that will serve has the Host file in the statistical match.

Input files: PUF2004 (SAS Dataset containing the original PUF.)

Output files: SOIRETS (SAS Dataset containing the PUF Extract. The Host file.)

PHASE-I Prepares both Host and Donor files for the match. Partitions each file along similar dimensions; constructs independent variables to be used in the predictive mean calculations; performs the predictive mean regressions within each partition; and merges the fitted values back to the original Host and Donor files. Our dependent variable in the regressions is taxable income (TINCX) on the PUF.

Input Files: SOIRETS
CPSRETS

Output Files: BETA (SAS Dataset with regression coefficients.)
CPSFILE (SAS Dataset extract for CPS.)
SOIFILE (SAS Dataset extract for PUF.)

PHASE-2 Performs the match. Creates working extracts from SOIRETS and CPSRETS containing record identifiers and fitted values; sorts both files by the fitted values (YHAT) within each partition; scales the sample weights on the Donor file to ensure that weighted population counts are identical across Host and Donor files for each partition; and performs the match using the predictive mean matching algorithm.

Input files: SOIFILE
CPSFILE

Output files: MATCH (SAS Dataset containing the results of the match: A record ID from the PUF, a record ID from the CPS and a newly-constructed, final match weight.

ADDCPSVARS Creates a preliminary version of the Production file by linking the Donor records from the CPS with the matched file.

Input files: CPSRETS
SOIRETS
MATCH

Output files: PROD2004_V1 (SAS Dataset with preliminary version of the Production File.)

ADDNONFILERS Adds non-filer records to the Production File. Maps CPS variables into their PUF counterpart to ensure a consistent record layout; gives each record a unique sequence number; and creates updated version of the Production File.

Input files: CPSNONF
PROD2004_V1

Output files: PROD2004_V1 (Updates w/ non-filer records.)

APPENDIX A – Results

Table 1A - March 2005 Current Population Survey, Household Record Summary

Variable	N	Mean	Sum	Std Dev	Minimum	Maximum
HID	76,447	50,166	3,835,040,341	28,669	3	98,664
NUMPER	76,447	3	210,648	2	1	16
NUMFAM	76,447	1	87,149	0	1	10
HTYPE	76,447	3	233,173	2	1	9
TENURE	76,447	1	100,386	0	1	3
HHINC	76,447	61,838	4,727,356,327	64,573	-28,454	1,125,395
STATE	76,447	54	4,119,659	26	11	95
REGION	76,447	3	199,105	1	1	4
HHENGVAL	76,447	9	711,687	71	0	1,999
HHDVAL	76,447	147	11,257,894	710	0	9,999

Table 1B - March 2005 Current Population Survey, Family Record Summary

Variable	N	Mean	Sum	Std Dev	Minimum	Maximum
HID	87,149	50,594	4,409,182,966	28,622	3	98,664
FID	87,149	1	100,315	0	1	10
FKIND	87,149	2	158,734	1	1	3
FTYPE	87,149	2	145,634	1	1	5
FSIZE	87,149	2	216,938	1	1	16
HEADIDX	87,149	1	110,614	1	1	16
WIFEIDX	87,149	1	71,484	1	0	15
HUSBIDX	87,149	1	60,662	1	0	16
SPOUIDX	87,149	1	87,286	1	0	16
FWEIGHT	87,149	1,491	129,967,630	978	30	16,263
FMVSL	87,149	71	6,188,682	218	0	3,592

Table 1C - March 2005 Current Population Survey, Person Record Summary

Variable	N	Mean	Sum	Std Dev	Minimum	Maximum
HHD	210,648	53,703	11,313,447,349	28,971	3	98,664
HID	210,648	1	228,427	0	1	10
PID	210,648	41	8,907,846	1	41	56
RELCODE	210,648	4	874,552	3	1	14
AGE	210,648	34	7,104,897	23	0	85
MARITAL	210,648	4	900,056	3	1	7
SEX	210,648	2	319,094	0	1	2
PSTAT	210,648	2	317,079	1	1	3
FAMNUM	210,648	1	189,417	0	0	6
FAMTYP	210,648	1	278,591	1	1	5
FAMREL1	210,648	2	393,713	1	0	4
FAMREL2	210,648	2	458,201	1	0	5
FAMREL3	210,648	3	537,812	2	1	8
FAMREL4	210,648	4	774,741	3	1	11
FAMREL5	210,648	11	2,415,391	17	1	51
PWEIGHT	210,648	1,382	291,166,196	933	21	18,081
SCHOOL	210,648	0	35,605	0	0	2
WAS	210,648	17,597	3,706,855,466	36,268	0	748,263
SELF	210,648	1,085	228,461,355	10,377	-19,998	507,982
FARM	210,648	133	28,013,062	3,441	-19,998	433,850
LCOMP	210,648	81	17,121,490	877	0	99,999
SOCSEC	210,648	1,188	250,213,096	3,737	0	50,000
PENSIONS	210,648	699	147,207,622	4,525	0	179,600
INTST	210,648	483	101,812,604	3,457	0	55,524
DIV	210,648	268	56,404,087	2,182	0	35,416
RENTS	210,648	194	40,788,116	3,014	-9,999	76,259
ALIMONY	210,648	19	4,029,705	285	0	64,152
HDLC	210,648	30	6,251,106	17	0	46
WORK	210,648	40	8,512,877	861	0	85,000
PUBA	210,648	22	4,662,339	353	0	75,000
DISI	210,648	70	14,676,756	1,358	0	107,454
HICO	210,648	1	203,825	1	0	2
HIPL	210,648	1	151,582	1	0	2
HIEM	210,648	0	81,751	1	0	2
HIFP	210,648	1	112,570	1	0	3
PECO	210,648	1	158,043	1	0	2
PEIN	210,648	0	70,772	1	0	2
HEAL	210,648	2	444,650	1	1	5
HCOV	210,648	1	254,528	1	0	2
VETH	210,648	79	16,658,007	1,317	0	99,999
VNA	210,648	35	7,362,519	813	0	57,893
SSIN	210,648	92	19,428,340	855	0	25,000
CHSU	210,648	109	23,061,980	1,051	0	96,280
SPOINTER	210,648	1	134,057	1	0	15
MCARE	210,648	2	397,784	0	1	2
MCAID	210,648	2	397,002	0	1	2
CHAMP	210,648	2	413,280	0	1	2
PENATVTY	210,648	85	17,802,175	81	57	555

Table 1D - March 2005 Current Population Survey, Composite Record Summary

Variable	N	Mean	Sum	Std Dev	Minimum	Maximum
HHD	210,648	53,703	11,313,447,349	28,971	3	98,664
FID	210,648	1	278,427	0	1	10
FKIND	210,648	1	278,423	1	1	3
FTYPE	210,648	1	278,591	1	1	5
FSIZE	210,648	3	700,357	2	1	16
HEADIDX	210,648	1	244,236	1	1	16
WIFEIDX	210,648	1	233,728	1	0	15
HUSBIDX	210,648	1	199,473	1	0	16
SPOUIDX	210,648	1	286,651	1	0	16
FWEIGHT	210,648	1,372	288,956,752	897	30	16,263
FMVSL	210,648	136	26,629,905	300	0	3,592
PID	210,648	42	8,907,846	1	41	56
RELCODE	210,648	4	874,552	3	1	14
AGE	210,648	34	7,104,897	22	0	85
MARITAL	210,648	4	900,056	3	1	7
SEX	210,648	2	319,094	0	1	2
PSTAT	210,648	2	317,079	1	1	3
FAMNUM	210,648	1	189,417	0	0	6
FAMTYP	210,648	1	278,591	1	1	5
FAMREL1	210,648	2	393,713	1	0	4
FAMREL2	210,648	2	458,201	1	0	5
FAMREL3	210,648	3	537,812	2	1	8
FAMREL4	210,648	4	774,741	3	1	11
FAMREL5	210,648	11	2,415,391	17	1	51
PWEIGHT	210,648	1,382	291,166,196	933	21	18,081
SCHOOL	210,648	0	35,605	0	0	2
WAS	210,648	17,597	3,706,855,466	36,268	0	748,263
SELF	210,648	1,085	228,461,355	10,377	-19,998	507,982
FARM	210,648	173	36,013,062	3,441	-19,998	477,850
LCOMP	210,648	81	17,121,490	877	0	99,999
SOCSEC	210,648	1,188	250,213,096	3,737	0	50,000
PENSIONS	210,648	699	147,207,622	4,525	0	179,600
INTST	210,648	483	101,812,604	3,457	0	55,524
DIVS	210,648	268	56,404,087	2,182	0	35,416
RENTS	210,648	194	40,788,116	3,014	-9,999	76,259
ALIMONY	210,648	19	4,029,705	785	0	64,152
EDUC	210,648	30	6,251,106	17	0	46
WORK	210,648	40	8,512,877	861	0	85,000
PUBA	210,648	22	4,662,339	357	0	75,000
DISI	210,648	70	14,676,758	1,358	0	107,454
HICO	210,648	1	203,825	1	0	2
HIPL	210,648	1	151,582	1	0	2
HIEM	210,648	0	81,751	1	0	2
HIEP	210,648	1	112,570	1	0	3
PECO	210,648	1	158,043	1	0	2
PEIN	210,648	0	70,772	1	0	2
HEAL	210,648	2	444,650	1	1	5
HCOV	210,648	1	254,528	1	0	2
VETB	210,648	79	16,658,007	1,317	0	99,999
FINA	210,648	35	7,362,519	813	0	57,893
SSIN	210,648	92	19,428,340	855	0	75,000
CHSU	210,648	109	23,061,980	1,051	0	76,280
SPOINTER	210,648	1	134,057	1	0	15
MCARE	210,648	2	397,784	0	1	2
MCAID	210,648	2	397,002	0	1	2
CHAMP	210,648	2	413,280	0	1	2
PENATVTY	210,648	85	17,802,175	81	57	555

Table 2A. - CPS Tax Units, by Type of Taxpayer: Filers (Unweighted)

	CPS Tax Units						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	33,382	1,616	13,276	34,155	568	10,051	76,048
Aged	4,781	47	3,118	387	101	434	8,868
Total	37,163	1,663	15,394	34,542	669	10,485	84,916
Dependent Filers							
Non-Aged	5,464	0	52	0	0	0	5,516
Aged	51	0	4	0	0	0	55
Total	5,515	0	56	0	0	0	5,571
All Taxpayers							
Non-Aged	37,846	1,616	13,328	34,155	568	10,051	81,564
Aged	4,832	47	3,122	387	101	434	8,923
Total	37,678	1,663	15,450	34,542	669	10,485	90,487

Source: *March 2005 Current Population Survey*

Table 2B. - CPS Tax Units, by Type of Taxpayer: Filers (Weighted)

	CPS Tax Units						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	43,517,339	2,057,078	20,103,231	38,581,130	1,000,055	13,047,179	108,276,608
Aged	8,150,009	62,082	5,451,326	538,144	164,423	650,785	15,016,769
Total	51,667,338	2,119,160	25,525,063	39,119,274	1,164,478	13,698,064	123,293,377
Dependent Filers							
Non-Aged	6,384,705	0	69,037	0	0	0	6,453,743
Aged	72,345	0	7,752	0	0	0	80,097
Total	6,456,550	0	76,789	0	0	0	6,533,339
All Taxpayers							
Non-Aged	49,901,514	2,057,078	20,142,774	38,581,130	1,000,055	13,047,179	114,729,851
Aged	8,222,354	62,082	5,459,078	538,144	164,423	650,785	15,096,866
Total	58,123,868	2,119,160	25,601,852	39,119,274	1,164,478	13,698,064	129,826,716

Source: *March 2005 Current Population Survey*

Table 2C. - CPS Tax Units, by Type of Taxpayer: Non-Filers (Unweighted)

	CPS Tax Units						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	1,884	368	1,631	687	67	243	4,875
Aged	3,046	39	3,403	183	43	167	6,881
Total	4,930	407	5,034	865	110	410	11,756
Dependent Filers							
Non-Aged	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Aged	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Total	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
All Taxpayers							
Non-Aged	1,884	368	1,631	687	67	243	4,875
Aged	3,046	39	3,403	183	43	167	6,881
Total	4,930	407	5,034	865	110	410	11,756

Source: *March 2005 Current Population Survey*

Table 2D. - CPS Tax Units, by Type of Taxpayer: Non-Filers (Weighted)

	CPS Tax Units						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	3,128,700	573,500	2,709,705	955,010	107,325	340,145	7,758,885
Aged	5,329,423	53,339	5,932,322	244,002	69,595	248,415	11,877,096
Total	8,458,123	576,839	8,641,577	1,199,012	171,920	588,560	19,635,981
Dependent Filers							
Non-Aged	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Aged	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Total	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
All Taxpayers							
Non-Aged	3,128,700	573,500	2,709,705	955,010	107,325	340,145	7,758,885
Aged	5,329,423	53,339	5,932,322	244,002	69,595	248,415	11,877,096
Total	8,458,123	576,839	8,641,577	1,199,012	171,920	588,560	19,635,981

Source: *March 2005 Current Population Survey*

Table 3A. - SOI Tax Units, by Type of Taxpayer: Filers (Unweighted)

	SOI Tax Returns						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	24,653	1,259	31,572	51,446	759	8,752	118,441
Aged	7,246	58	18,564	1,469	146	290	27,773
Total	31,899	1,317	50,136	52,915	905	9,042	146,214
Dependent Filers							
Non-Aged	3,793	0	1	0	0	0	3,794
Aged	39	0	0	0	0	0	39
Total	3,832	0	1	0	0	0	3,833
All Taxpayers							
Non-Aged	28,446	1,259	31,573	51,446	759	8,752	122,235
Aged	7,285	58	18,564	1,469	146	290	27,812
Total	35,731	1,317	50,137	52,915	905	9,042	150,047

Source: *2004 Statistics of Income, Public Use File*

Table 3B. - SOI Tax Units, by Type of Taxpayer: Filers (Weighted)

	SOI Tax Returns						
	Filing Status						
	Single		Joint		Head of Household		
Dependency And Aged Status	With Dependents	Without Dependents	With Dependents	Without Dependents	With Dependents	Without Dependents	Total
Non-Dependent Filers							
Non-Aged	39,315,476	1,256,915	16,831,004	27,593,754	1,101,780	18,283,112	104,381,543
Aged	6,712,729	40,883	9,119,152	817,893	196,546	331,510	17,218,713
Total	46,028,205	1,297,798	25,950,156	28,411,647	1,297,876	18,614,622	121,600,255
Dependent Filers							
Non-Aged	9,563,918	0	3,346	0	0	0	9,567,264
Aged	58,377	0	0	0	0	0	58,377
Total	9,622,295	0	3,346	0	0	0	9,625,641
All Taxpayers							
Non-Aged	48,879,394	1,256,915	16,834,350	27,593,754	1,101,780	18,283,112	114,948,806
Aged	6,771,106	40,883	9,119,152	817,893	196,546	331,510	17,277,090
Total	55,650,500	1,297,798	25,953,503	28,411,647	1,297,876	18,614,622	132,225,896

Source: *2004 Statistics of Income, Public Use File*

Table 4A - SOI-CPS Match Blocking Partition for Filers

Dependency Status	Filing Status	Aged Status	Presence of Dependents	SOI Records	CPS Records	SOI Records (Weighted)	CPS Records (Weighted)	Factor	Cell
Non-Dependent Filer	Single Returns	Non-Aged Return	No Dependents	74,653	77,382	89,315,476	43,517,339	0.9034	1
Non-Dependent Filer	Single Returns	Non-Aged Return	1 Dependent	857	1,009	1,246,369	1,262,779	1.2349	2
Non-Dependent Filer	Single Returns	Non-Aged Return	2 Dependents	306	313	536,270	386,766	1.3866	3
Non-Dependent Filer	Single Returns	Non-Aged Return	3 Dependents	101	294	174,277	418,083	0.4169	4
Non-Dependent Filer	Single Returns	Aged Return	n.a.	7,304	4,828	6,753,613	8,212,091	0.8224	5
Non-Dependent Filer	Joint Returns	Non-Aged Return	No Dependents	31,572	17,276	16,831,004	20,073,737	0.8385	6
Non-Dependent Filer	Joint Returns	Non-Aged Return	1 Dependent	14,850	8,629	9,716,956	10,627,334	0.9143	7
Non-Dependent Filer	Joint Returns	Non-Aged Return	2 Dependents	21,133	9,707	11,785,947	11,318,575	0.9971	8
Non-Dependent Filer	Joint Returns	Non-Aged Return	3 Dependents	13,786	4,111	4,755,531	4,698,052	1.0122	9
Non-Dependent Filer	Joint Returns	Non-Aged Return	4 Dependents	986	1,184	1,330,409	1,342,692	0.9909	10
Non-Dependent Filer	Joint Returns	Non-Aged Return	5+ Dependents	391	574	504,910	594,488	0.8493	11
Non-Dependent Filer	Joint Returns	Aged Return	n.a.	20,033	3,505	9,937,045	5,989,470	1.6591	12
Non-Dependent Filer	Head of Household	Non-Aged Return	No Dependents	759	568	1,101,780	1,000,055	1.1012	13
Non-Dependent Filer	Head of Household	Non-Aged Return	1 Dependent	4,757	5,174	9,581,713	6,915,773	1.3856	14
Non-Dependent Filer	Head of Household	Non-Aged Return	2 Dependents	2,967	3,018	6,578,037	3,774,994	1.7425	15
Non-Dependent Filer	Head of Household	Non-Aged Return	3 Dependents	1,028	1,859	2,123,362	2,357,063	0.9009	16
Non-Dependent Filer	Head of Household	Aged Return	n.a.	436	535	538,056	815,208	0.6478	17
Dependent Filer	n.a.	n.a.	n.a.	3,833	5,571	9,625,641	6,533,339	1.4733	18
				150,017	90,187	137,225,896	129,826,715	1.0185	

Table 5A - All Observations: Summary Statistics for the Matched Data File

		Matched File (SOI-Host)				CPS-Donor				
Item	Variable	n	N (Millions)	m_x	s	n	N (Millions)	m_y	s	$Diff=m_x-m_y$
1	Age of Tax Unit Head	240,515	132.2	42.18	17.02	90,487	129.8	42.44	16.69	-0.02
2	Age of Tax Unit Spouse	240,515	132.2	18.49	24.63	90,487	129.8	18.20	23.72	0.12
3	Age of Dependent #1	240,515	132.2	6.15	9.31	90,487	129.8	5.64	9.07	14.46
4	Age of Dependent #2	240,515	132.2	7.42	5.79	90,487	129.8	7.21	5.51	9.85
5	Age of Dependent #3	240,515	132.2	0.70	3.58	90,487	129.8	0.71	3.60	-0.29
6	Age of Dependent #4	240,515	132.2	0.19	2.05	90,487	129.8	0.20	2.10	-0.94
7	Age of Dependent #5	240,515	132.2	0.07	1.16	90,187	129.8	0.07	1.17	0.69
8	Age of Youngest Child	240,515	132.2	3.15	6.93	90,487	129.8	2.91	6.72	9.13
9	Age of Oldest Child	240,515	132.2	4.46	8.76	90,487	129.8	4.18	8.56	8.56
10	HI: Covered (HEAD)	240,515	132.2	1.27	0.45	90,487	129.8	1.27	0.44	2.25
11	HI: Employer-Provided (HEAD)	240,515	132.2	0.60	0.61	90,487	129.8	0.61	0.60	-4.34
12	HI: Employer Pays (HEAD)	240,515	132.2	0.85	0.98	90,487	129.8	0.87	0.98	-4.45
13	HI: Covered (SPOUSE)	240,515	132.2	0.46	0.62	90,487	129.8	0.47	0.62	-3.03
14	HI: Employer-Provided (SPOUSE)	240,515	132.2	0.22	0.46	90,487	129.8	0.22	0.46	-0.33
15	HI: Employer Pays (SPOUSE)	240,515	132.2	0.32	0.73	90,487	129.8	0.33	0.74	-4.37
16	Pension: Offered (HEAD)	240,515	132.2	1.24	0.72	90,487	129.8	1.23	0.71	2.90
17	Pension: Included (HEAD)	240,515	132.2	0.53	0.67	90,487	129.8	0.54	0.67	-1.42
18	Pension: Offered (SPOUSE)	240,515	132.2	0.12	0.70	90,187	129.8	0.11	0.71	6.55
19	Pension: Included (SPOUSE)	240,515	132.2	0.20	0.46	90,487	129.8	0.22	0.47	-7.97
20	Health Status (HEAD)	240,515	132.2	2.20	1.05	90,487	129.8	2.20	1.05	-0.16
21	Health Status (SPOUSE)	240,515	132.2	0.87	1.24	90,487	129.8	0.87	1.23	-0.57
22	Supplemental Security Income	240,515	132.2	91.57	871.44	90,487	129.8	93.38	880.96	-0.53
23	Public Assistance (TANF)	240,515	132.2	28.62	411.33	90,487	129.8	28.19	419.62	0.27
24	Workmen's Compensation	240,515	132.2	66.72	1,079.10	90,487	129.8	70.17	1,120.41	-0.80
25	Veterans Benefits	240,515	132.2	128.71	1,662.63	90,487	129.8	128.24	1,661.81	0.07
26	Child Support	240,515	132.2	130.93	1,527.84	90,487	129.8	190.20	1,374.90	7.36
27	Disability Income	240,515	132.2	102.36	1,698.10	90,487	129.8	106.56	1,740.74	-0.67
28	Social Security Income	240,515	132.2	3,091.98	5,826.55	90,487	129.8	1,809.77	5,234.35	13.42
29	Home Ownership (TENURE)	240,515	132.2	0.51	0.50	90,187	129.8	0.51	0.50	3.71
30	Wage Share (Lesser Earner)	240,515	132.2	0.07	0.14	90,487	129.8	0.07	0.15	-10.77
31	Energy Assistance	240,515	132.2	4.27	48.98	90,487	129.8	4.04	46.99	1.22
32	Food Stamps	240,515	132.2	88.55	545.26	90,487	129.8	85.45	549.02	1.45
33	School Lunches	240,515	132.2	53.00	186.21	90,487	129.8	50.81	187.93	2.99
34	Medicare (Head)	240,515	132.2	1.87	0.34	90,487	129.8	1.88	0.33	-6.86
35	Medicaid (Head)	240,515	132.2	1.93	0.26	90,487	129.8	1.93	0.25	-3.84
36	Champus (Head)	240,515	132.2	1.97	0.18	90,487	129.8	1.97	0.18	-0.61
37	Country of Origin (Head)	240,515	132.2	89.82	87.10	90,487	129.8	90.28	87.57	-1.33
38	Medicare (Spouse)	240,515	132.2	0.74	0.94	90,487	129.8	0.78	0.96	-9.57
39	Medicaid (Spouse)	240,515	132.2	0.78	0.97	90,487	129.8	0.80	0.97	-4.57
40	Champus (Spouse)	240,515	132.2	0.78	0.96	90,187	129.8	0.79	0.97	1.67
41	Country of Origin (Spouse)	240,515	132.2	37.52	73.24	90,487	129.8	38.60	74.46	-3.74

Table 5B - Non-Zero Observations: Summary Statistics for the Matched Data File

		Matched File (SOI Host)				CPS Donor				
Item	Variable	n	N (Millions)	m_x	s	n	N (Millions)	m_y	s	$(m_x-m_y)/s$
1	Age of Tax Unit Head	240,515	132.2	42.18	17.02	90,487	129.8	42.44	16.69	-0.02
2	Age of Tax Unit Spouse	139,639	57.7	46.40	15.07	38,768	53.8	44.74	13.99	0.38
3	Age of Dependent #1	111,130	58.5	13.90	9.39	44,529	53.2	13.75	9.11	0.01
4	Age of Dependent #2	66,942	31.2	10.26	7.84	24,344	28.4	10.08	7.69	0.03
5	Age of Dependent #3	24,197	10.1	9.17	9.46	8,732	10.1	9.06	9.50	0.09
6	Age of Dependent #4	4,252	1.8	9.10	10.95	1,515	1.8	9.02	11.04	0.09
7	Age of Dependent #5	1,251	0.8	11.25	15.12	761	0.8	10.90	14.62	0.02
8	Age of Youngest Child	86,280	40.5	10.28	9.12	30,899	37.1	10.18	9.18	1.62
9	Age of Oldest Child	97,792	44.1	13.39	10.52	33,645	40.4	13.40	10.56	-0.18
10	HI: Covered (HEAD)	240,504	132.2	1.27	0.45	90,480	129.8	1.27	0.44	2.25
11	HI: Employer-Provided (HEAD)	140,887	71.1	1.12	0.32	47,877	71.1	1.12	0.32	1.55
12	HI: Employer Pays (HEAD)	120,958	61.4	1.83	0.50	41,848	61.6	1.83	0.50	1.67
13	HI: Covered (SPOUSE)	138,131	57.0	1.17	0.37	38,154	53.1	1.16	0.37	2.26
14	HI: Employer Provided (SPOUSE	71,863	36.0	1.11	0.32	18,906	36.0	1.09	0.29	7.79
15	HI: Employer Pays (SPOUSE)	64,032	33.8	1.85	0.51	17,017	33.3	1.85	0.50	0.92
16	Pension: Offered (HEAD)	204,316	110.2	1.49	0.50	76,987	108.4	1.48	0.50	5.76
17	Pension: Included (HEAD)	115,495	56.2	1.24	0.43	40,187	56.6	1.23	0.42	4.25
18	Pension: Offered (SPOUSE)	107,089	39.1	1.41	0.49	30,605	40.8	1.40	0.49	2.30
19	Pension: Included (SPOUSE)	68,363	23.1	1.15	0.36	18,461	24.4	1.15	0.36	0.75
20	Health Status (HEAD)	240,515	132.2	2.20	1.05	90,487	129.8	2.20	1.05	-0.16
21	Health Status (SPOUSE)	139,691	52.7	2.18	1.00	38,802	52.9	2.14	0.99	6.66
22	Supplemental Security Income	3,044	2.1	5,702.93	3,910.90	1,489	2.1	5,717.16	3,919.96	-0.11
23	Public Assistance (TANF)	1,717	1.4	2,761.38	2,962.35	995	1.2	2,968.85	3,132.85	-1.70
24	Workmans Compensation	3,004	1.4	6,532.18	8,471.35	1,020	1.4	6,643.59	8,670.64	-0.34
25	Veterans Benefits	3,893	1.7	9,902.66	10,765.76	1,111	1.7	9,966.41	10,796.71	0.17
26	Child Support	8,234	6.1	5,005.47	5,166.93	4,326	5.0	4,918.97	5,062.75	0.90
27	Disability Income	1,401	1.0	13,608.34	14,127.32	665	1.0	13,764.75	14,254.86	-0.75
28	Social Security Income	38,377	20.3	13,621.08	8,000.30	11,454	18.6	13,660.75	7,371.75	17.00
29	Home Ownership (TENURE)	155,656	66.9	1.00	0.00	46,670	66.6	1.00	0.00	-
30	Wage Share (Lesser Earner)	79,458	28.1	0.31	0.14	22,695	30.0	0.31	0.14	-1.72
31	Energy Assistance	3,419	1.8	318.21	381.17	1,417	1.7	315.91	373.39	0.25
32	Food Stamps	6,907	5.6	2,096.57	1,682.04	3,906	5.1	2,158.67	1,771.70	-1.78
33	School Lunches	47,557	22.4	317.45	351.23	18,260	20.8	317.46	368.81	-1.58
34	Medicare (Head)	240,515	132.2	1.87	0.34	90,487	129.8	1.88	0.33	-6.86
35	Medicaid (Head)	240,515	132.2	1.93	0.26	90,487	129.8	1.93	0.25	-1.84
36	Champus (Head)	240,515	132.2	1.97	0.18	90,487	129.8	1.97	0.18	0.61
37	Country of Origin (Head)	240,515	132.2	89.82	87.10	90,487	129.8	90.28	87.57	-1.35
38	Medicare (Spouse)	139,691	57.7	1.87	0.34	38,802	53.9	1.91	0.28	-29.15
39	Medicaid (Spouse)	139,691	57.7	1.95	0.21	38,802	53.9	1.96	0.21	-0.42
40	Champus (Spouse)	139,691	57.7	1.95	0.22	38,802	53.9	1.95	0.21	-2.75
41	Country of Origin (Spouse)	139,691	57.7	94.08	90.16	38,802	53.9	94.81	91.05	-1.40

APPENDIX B – Brief Description of Predictive Mean Matching[11]

- Partitioning: Attempt to keep (unweighted) cell sizes to a minimum of 30 and a maximum of 500. Check the weighted cell counts for each of the partitions and check for unbalanced cells.

- Estimation: Fit a (weighted) linear model for two Y- and Z-variables as a function of the X-variables common to both files.

Host File: A(X,Y)	Donor File: B(X,Z)
$Y_1 = F_1(X) = XB_1 + e_1$	$Z_1 = F_1(X) = XB_1 + e_1$
$Y_2 = F_2(X) = XB_2 + e_2$	$Z_2 = F_2(X) = XB_2 + e_2$

- Predicted Values: Calculate fitted values of all four variables on each file.

Host File: A(X,Y)	Donor File: B(X,Z)
$Y^*_1, Y^*_2, Z^*_1, Z^*_2$	$Y^*_1, Y^*_2, Z^*_1, Z^*_2$

- Align Partitions: Scale the weights for each cell in the Donor file so that they are equal to the weights in Host file. Next, sort cells on the predicted value of one of the Z-variables so records with the closest values of the match variables are ordered correctly.
- Perform Match: Match each record in Host File to the closest Donor record, splitting records if necessary to ensure all the "weight" on the Donor Records are used up

[11] This material is from "Statistical Matching Meeting", March 24th, 2003, *Urban-Brookings Tax Policy Center*.

APPENDIX C – Predictive Mean Matching Results

Table C1 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 1

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-605,9770	3106.7391	-0.2000
AGEDF	-	-	-
WAS	1.0420	0.0130	80.0100
INTST	7.6250	0.0593	44.2500
DIV	1.3916	0.0396	35.1700
BIL	0.6655	0.0447	14.8800
UI	0.2860	0.2039	1.4000
SCHC	1.1166	0.0085	130.9500
PENSIONS	0.1723	0.0480	3.5900
SSINC	0.3275	0.5324	0.6200
LCAGIX	0.5053	0.6366	0.7900
ALIMONY	0.6079	0.5093	1.1900
WAGESIR	-11,470.0000	5285.7867	-2.1700
CAPSIR	2,474.1002	6057.8901	0.4100
WAGEFLAG	1,235.7076	5085.2832	0.2400
SETFFLAG	-1,518.9551	2630.6414	-0.5800

Root MSE	4,480,880	R-Square	0.5928
Dependent Mean	33,141	Adj R-Sq	0.5926
Coeff Var	90,238		

Table C2 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 2

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-4,739.5197	11,745.0001	-0.4000
AGEDF	-	-	-
WAS	0.8722	0.0384	22.7200
INTST	0.7003	0.0440	15.9300
DIV	1.3071	0.1546	8.4500
BIL	0.8656	0.1418	6.1100
UI	0.3319	0.9231	0.3600
SCHC	0.7500	0.0443	16.9300
PENSIONS	0.2012	0.3299	0.6100
SSINC	-0.3438	5.0823	-0.0700
LCAGIX	-0.3879	1.5713	-0.2500
ALIMONY	0.7975	7.5317	0.1100
WAGESIR	-19,297.0000	16,617.0000	-1.1600
CAPSIR	94,442.0000	31,448.0000	3.0000
WAGEFLAG	15,980.0000	15,051.0000	1.0600
SETFFLAG	-5,822.0630	8,645.9656	-0.6700

Root MSE	7,299.792	R-Square	0.5374
Dependent Mean	15,434	Adj R-Sq	0.5297
Coeff Var	14.901		

Table C3 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 3

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	3,664.0733	23,163.0000	0.1600
AGEDIF	-	-	-
WAS	0.9465	0.0267	35.4900
INTST	3.0685	0.6888	4.4500
DIV	1.5740	0.4026	3.9100
BIL	0.6866	0.2380	2.8800
TIL	1.1025	0.7074	1.5600
SCHT	0.7620	0.0667	11.4300
PENSIONS	0.1007	0.3705	0.2700
SSINC	0.1059	9.3848	0.0100
LCAGIX	-1,646?	4.5737	-0.3600
ALIMONY	0.0000	.	.
WAGESHR	-26,659.0000	33,592.0000	-0.7900
CAPSHR	81,090.0000	56,084.0000	1.4500
WAGEFLAG	11,794.0000	26,170.0000	0.4500
SELFFLAG	-13,579.0000	18,570.0000	-0.7300
Root MSE	?,189,601	R-Square	0.8373
Dependent Mean	9,706	Adj R-Sq	0.8301
Coeff Var	??,559		

Table C4 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 4

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-9,338.2424	23,809.0000	-0.3900
AGEDIF	-	-	-
WAS	0.9884	0.0965	10.2500
INTST	7.8727	3.2631	2.4100
DIV	0.8674	0.4490	1.9300
BIL	0.8791	0.0936	9.4000
TIL	1.4019	1.3117	1.0700
SCHC	0.6296	0.1493	4.2200
PENSIONS	-0.3401	0.7056	-0.4800
SSINC	0.6597	7.3506	0.0900
LCAGIX	0.9247	8.0208	0.1200
ALIMONY	-	-	-
WAGESHR	-7,438.7502	41,603.0000	-0.1800
CAPSHR	164,788.0000	153,680.0000	1.0700
WAGEFLAG	473.6707	47,229.0000	0.0100
SELFFLAG	-1,244.5746	21,378.0000	-0.0600
Root MSE	1,764,903	R-Square	0.8262
Dependent Mean	12,954	Adj R-Sq	0.8002
Coeff Var	13,624		

Table C5 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 5

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	7,535.1008	17,314.0000	0.4400
AGEDIF	-5,192.8043	16,527.0000	-0.3100
WAS	1.2506	0.0505	24.7800
INTST	1.9441	0.0888	21.9000
DIV	1.3694	0.0462	29.6700
BIL	-3.1953	0.1277	-25.0100
FIL	-1.4428	0.3038	-4.7500
SCHE	0.7715	0.0209	36.9300
PENSIONS	0.1995	0.0550	3.6300
SSINC	0.7982	0.2817	2.8300
LCAGIX	0.2987	3.0799	0.1000
ALIMONY	0.6031	0.8298	0.7300
WAGESHR	-21,457.0000	17,619.0000	-1.7000
CAPSHR	-16,395.0000	7,571.9561	-2.4300
WAGEFLAG	-2,817.5122	7,584.7389	-0.3700
SELFFLAG	28,086.0000	7,493.4520	3.7500

Root MSE	4,974.785	R-Square	0.4624
Dependent Mean	24.725	Adj R-Sq	0.4613
Coeff Var	20.120		

Table C6 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 6

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	3,993.7633	8,206.9431	0.4900
AGEDIF	-	-	-
WAS	1.0151	0.0097	105.1000
INTST	3.1423	0.0500	63.8900
DIV	1.3060	0.0276	47.2600
BIL	0.6675	0.0493	13.5300
FIL	0.1290	0.2228	0.5800
SCHE	0.8231	0.0103	80.1800
PENSIONS	0.0400	0.0402	1.0000
SSINC	0.0226	0.6733	0.0300
LCAGIX	-0.0105	1.2950	-0.0100
ALIMONY	0.3519	4.7901	0.0700
WAGESHR	-43,553.0000	9,590.6083	-4.5400
CAPSHR	36,021.0000	18,506.0000	1.9500
WAGEFLAG	19,155.0000	10,238.0000	1.8700
SELFFLAG	-6,269.5321	4,949.3760	-1.2700
Root MSE	8,239.848	R-Square	0.5061
Dependent Mean	63,529	Adj R-Sq	0.5059
Coeff Var	13.178		

Table C7 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 7

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	3,620.3422	9,021.7106	0.4000
AGEDF	-	-	-
WAS	0.9715	0.0101	96.7100
INTST	4.1015	0.0394	103.9900
DIV	1.1205	0.0208	53.9600
BIL	0.7193	0.0517	13.9100
TIL	-1.7539	0.2995	-5.8600
SCHT	0.6542	0.0125	52.4800
PENSIONS	0.0622	0.0657	0.9500
SSINC	-0.1333	0.9604	-0.1400
LCAGIX	0.1678	1.1771	0.1400
ALIMONY	0.6256	2.8552	0.2200
WAGESHR	-43,008.1000	9,947.5303	-4.3200
CAPSHR	40,035.0000	23,616.0000	1.7000
WAGEFLAG	19,177.0000	11,208.0000	1.7100
SELFFLAG	-8,057.8595	4,352.5725	-1.8500

Root MSE	5,144,069	R-Square	0.7095
Dependent Mean	58,962	Adj R-Sq	0.7092
Coeff Var	8.724		

Table C8 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 8

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-6,211.4790	9,767.1813	-0.6400
AGEDIF	-	-	-
WAS	1.0425	0.0111	93.8000
INTST	2.8492	0.0547	52.0700
DIV	2.0687	0.0608	34.0100
BIL	0.9653	0.0528	18.2800
TIL	0.1416	0.3271	0.4300
SCHT	0.6404	0.0118	54.0900
PENSIONS	0.0101	0.0805	0.1000
SSINC	0.5706	1.9680	0.2900
LCAGIX	0.4830	1.3122	0.3700
ALIMONY	0.5531	2.8027	0.2000
WAGESHR	-42,098.1000	10,778.0000	-3.9100
CAPSHR	100,766.0000	29,501.0000	3.4200
WAGEFLAG	16,501.0000	12,209.0000	1.3500
SELFFLAG	-7,943.6762	4,542.7785	-1.7500
Root MSE	6,072,872	R-Square	0.5066
Dependent Mean	65,535	Adj R-Sq	0.5062
Coeff Var	9.266		

Table C9 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 9

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-23,070.0000	19,781.0000	-1.1700
AGEDIF	-	-	-
WAS	1.1416	0.0152	75.1300
INTST	3.5133	0.0728	48.2900
DIV	1.3055	0.0257	50.8500
BIL	1.2457	0.0934	13.3300
TIL	0.5764	0.2808	2.0500
SCHE	0.7626	0.0156	49.0500
PENSIONS	0.3448	0.1618	2.1300
SSINC	0.7356	3.8589	0.1900
LCAGIX	0.9902	2.8185	0.3500
ALIMONY	0.2122	17.4908	0.0100
WAGESHR	-36,617.0000	22,201.0000	-1.6500
CAPSHR	417,711.0000	68,810.0000	6.0700
WAGEFLAG	15,197.0000	24,981.0000	0.6100
SELFFLAG	-11,680.0000	10,057.0000	-1.1600

Root MSE	8,676,544	R-Square	0.5562
Dependent Mean	76,811	Adj R-Sq	0.5558
Coeff Var	11.296		

Table C10 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 10

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-3,825.5583	3,366.4348	-1.1400
AGEDIF	-	-	-
WAS	0.8713	0.0149	58.6000
INTST	0.9176	0.2237	4.1000
DIV	0.7283	0.1178	6.1800
BIL	0.4248	0.0509	8.3500
TIL	0.1306	0.1385	0.9400
SCHF	0.6584	0.0215	30.5600
PENSIONS	0.1161	0.0500	2.3200
SSINC	0.0631	0.7835	0.0800
LCAGIX	1.1344	0.4972	2.2800
ALIMONY	-	-	-
WAGESHR	-24,456.1000	4,457.4540	-5.4900
CAPSHR	8,630.5609	9,626.5717	0.9000
WAGEFLAG	3,849.8362	4,218.3158	0.9100
SELFFLAG	-2,660.9279	1,713.5880	-1.5500

Root MSE	777.151	R-Square	0.8776
Dependent Mean	30.308	Adj R-Sq	0.8760
Coeff Var	7.399		

Table C11 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 11

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	1,036.0901	4,884.1707	0.2100
AGEDF	-	-	-
WAS	1.0105	0.0142	71.0100
INTST	-0.2514	0.6003	-0.4200
DIV	-0.3329	0.9496	-0.3500
BIL	0.4710	0.0777	6.0600
TIL	0.2183	0.1529	1.4300
SCHT	0.3816	0.0495	7.7200
PENSIONS	0.0582	0.0781	0.7400
SSINC	-	-	-
LCAGIX	0.7527	0.9474	0.7900
ALIMONY	-	-	-
WAGESHR	-44,119.0000	7,789.5134	-5.6600
CAPSHR	13,717.0000	26,690.0000	0.5100
WAGEFLAG	6,440.3139	6,577.9777	0.9800
SELFFLAG	-4,306.9159	2,920.5306	-1.4700

Root MSE	707,242	R-Square	0.9355
Dependent Mean	23,476	Adj R-Sq	0.9335
Coeff Var	3,013		

Table C12 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 12

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-315.7710	9,091.9007	-0.0300
AGEDIF	-5,578.0165	4,784.4138	-1.1700
WAS	0.9710	0.0171	56.7100
INTST	2.0768	0.0462	44.9400
DIV	1.1760	0.0171	68.6400
BIL	0.9037	0.0565	15.9900
TIL	0.6117	0.2605	2.3500
SCHT	0.9388	0.0109	86.3400
PENSIONS	0.2616	0.0362	7.3200
SSINC	1.1032	0.2256	4.8900
LCAGIX	-0.7961	2.5264	-0.3200
ALIMONY	0.9802	5.2138	0.1900
WAGESHR	-16,904.0000	10,578.0000	-1.6000
CAPSHR	77,013.0000	11,699.0000	7.3100
WAGEFLAG	1,858.6017	6,250.6939	0.3000
SELFFLAG	3,135.9937	5,521.1161	0.5700
Root MSE	6,747.114	R-Square	0.5323
Dependent Mean	50.709	Adj R-Sq	0.5320
Coeff Var	13.306		

Table C13 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 13

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	947.5454	15,417.0000	0.0600
AGEDIF	-	-	-
WAS	0.9149	0.0707	12.9400
INTST	6.4673	0.6035	10.7200
DIV	0.8588	0.4627	1.8600
BIL	0.9515	0.1408	6.7600
UI	-0.8217	1.3402	-0.6100
SCHC	0.8091	0.0674	12.0000
PENSIONS	0.3493	0.2663	1.3100
SSINC	0.1984	3.1481	0.0600
LCAGIX	0.2269	2.7887	0.0800
ALIMONY	0.4535	1.3600	0.3300
WAGESHR	-16,504.0000	19,747.0000	-0.8400
CAPSHR	7,492.1404	53,198.0000	0.1400
WAGEFLAG	5,566.7458	21,088.0000	0.2600
SELFFLAG	-4,415.9173	10,539.0000	-0.4200
Root MSE	3,407,650	R-Square	0.5157
Dependent Mean	27,490	Adj R-Sq	0.5066
Coeff Var	12.396		

Table C14 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 14

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-2,440.5893	4,627.0073	-0.5300
AGEDF	-	-	-
WAS	0.9724	0.0203	47.8800
INTST	1.4439	0.1226	11.7800
DIV	2.1542	0.1033	20.8500
BIL	0.9135	0.0578	15.8100
TIL	-0.3782	0.7947	-0.4800
SCHF	0.8032	0.0215	37.4300
PENSIONS	0.2601	0.1201	2.1600
SSINC	0.2796	0.9375	0.3000
LCAGIX	0.4149	0.6605	0.6300
ALIMONY	0.3112	0.4332	0.7200
WAGESHR	-14,266.1000	6,884.5196	-2.0700
CAPSHR	6,849.7170	15,929.0000	0.4300
WAGEFLAG	3,568.7496	6,519.0671	0.5500
SELFFLAG	-5,106.6689	3,319.9375	-1.5400

Root MSE	7,507.842	R-Square	0.5449
Dependent Mean	15,877	Adj R-Sq	0.5435
Coeff Var	15,795		

Table C15 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 15

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-4,705.1589	11,935.0000	-0.3900
AGEDF	-	-	-
WAS	1.0582	0.0430	24.6000
INTST	6.7206	0.4700	14.3000
DIV	1.4783	0.2527	5.8500
BIL	1.1058	0.2315	4.7800
UI	0.2613	1.7167	0.1500
SCHC	0.5535	0.0274	20.2100
PENSIONS	-0.0500	0.2555	-0.2000
SSINC	0.9406	2.6717	0.3500
LCAGIX	0.5548	1.6928	0.3300
ALIMONY	0.7163	0.8881	0.8100
WAGESHR	-12,551.0000	15,473.0000	-0.8100
CAPSHR	5,486.4132	56,246.0000	0.1000
WAGEFLAG	549.2929	14,376.0000	0.0400
SELFFLAG	-6,051.0244	8,234.7659	-0.7300
Root MSE	4,718.768	R-Square	0.6051
Dependent Mean	10,419	Adj R-Sq	0.6032
Coeff Var	45.285		

Table C16 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 16

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-6,667.1345	7,201.7565	-0.9300
AGEDF	-	-	-
WAS	0.8683	0.0254	34.1600
INTST	5.0354	0.5904	8.5300
DIV	1.0902	0.1267	8.6100
BIL	0.9462	0.1308	7.2300
TIL	1.3925	0.8304	1.6800
SCHT	0.6148	0.0285	21.5500
PENSIONS	0.2016	0.1550	1.3000
SSINC	-0.3684	1.6989	-0.2200
LCAGIX	0.2084	0.7973	0.2600
ALIMONY	0.4998	0.4580	1.0900
WAGESHR	-17,043.0000	10,173.0000	-1.6800
CAPSHR	33,792.0000	19,907.0000	1.7000
WAGEFLAG	7,834.7661	9,686.7969	0.8100
SELFFLAG	-4,898.8154	5,103.9830	-0.9600

Root MSE	1,493,371	R-Square	0.6889
Dependent Mean	10,466	Adj R-Sq	0.6846
Coeff Var	14.269		

Table C17 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 17

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	t-VALUE
Intercept	-9,437.6825	23,468.0000	-0.4000
AGEDIF	10,127.0000	22,305.0000	0.4500
WAS	0.9735	0.0751	12.9600
INTST	1.3478	0.3807	3.5400
DIV	1.8776	0.4095	4.5900
BIL	0.6615	0.2626	2.5200
TIL	-2.6816	1.5935	-1.6800
SCHF	0.7523	0.0591	12.7300
PENSIONS	0.1912	0.0792	2.4100
SSINC	0.6096	0.3052	2.0000
LCAGIX	0.4533	2.3656	0.1900
ALIMONY	0.5788	0.6993	0.8300
WAGESHR	-13,880.0000	12,113.0000	-1.1500
CAPSHR	-7,957.4853	16,377.0000	-0.4900
WAGEFLAG	-1,447.6736	8,336.0908	-0.1700
SELFFLAG	974.8701	9,103.3114	0.1000
Root MSE	1,697,810	R-Square	0.5764
Dependent Mean	17,299	Adj R-Sq	0.5613
Coeff Var	9,815		

Table C18 - Regression Results From the Predictive Mean Matching Algorithm

CELLID = 18

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR		t-VALUE
Intercept	1,275.8103	1,978.8046		0.6400
AGEDIF	-5,263.4722	8,666.9063		-0.6100
WAS	0.8016	0.0258		31.1000
INTST	2.3257	0.1441		16.1400
DRIV	1.3017	0.0592		22.0000
BIL	0.7178	0.3980		1.8000
FIL	0.0818	4.9132		0.0200
SCHE	1.0043	0.0202		49.8100
PENSIONS	0.4995	0.4466		1.1200
SSINC	0.4144	1.0131		0.4100
LCAGIX	0.4989	1.6378		0.3000
ALIMONY	.	.		.
WAGESHR	-2,997.6241	3,101.8965		-0.9700
CAPSHR	-1,335.6215	2,210.9834		-0.6000
WAGEFLAG	-807.0593	2,477.8700		-0.3300
SELFFLAG	-1,265.1518	2,093.0755		-0.6000
Root MSE	881.494	R-Square		0.5750
Dependent Mean	2,364	Adj R-Sq		0.5734
Coeff Var	37.290			