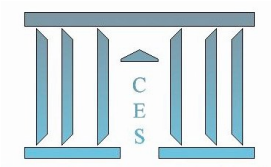
Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data

Anil Alpman, François Gardes, Noel Thiombiano

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers



Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data

Anil ALPMAN, François GARDES, Noël THIOMBIANO

2017.24

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://centredeconomiesorbonne.univ-paris1.fr/
ISSN : 1955-611X

Statistical Matching for Combining Time-Use Surveys with Consumer Expenditure Surveys: An Evaluation on Real Data

Anil Alpman[†], François Gardes[‡], and Noel Thiombiano[§]

May 5, 2017

Abstract

Performing a statistical match to combine two surveys made over the same population by traditional methods is shown to give biased estimates and variance of the imputed values. A method proposed by Rubin (1986) allows imputing an unobserved variable using observations in another dataset by taking into account the partial correlation between the variables that are jointly unobserved for any unit. We use a dataset where households report their expenditures and time-uses to show that fusioning expenditure and time-use surveys by Rubin's procedure allows to recover the true distribution of the missing variables and to yield minimally biased estimates.

Keywords: Data combination, Data Fusion, Missing data, Statistical Matching, Time-use.
JEL: C10, D13.

1 Introduction

Statistical matching allows the constitution of pseudo-panels by combining surveys that were carried out in successive periods or to expand the number of variables by combining two different surveys conducted in the same period (the surveys that are combined must nonetheless sample the same

[*]We would like to thank the the General Directorate for the Promotion of the Rural Economy (DGPER) and the University Ouaga II for providing us the data.
[†]Paris School of Economics and University Paris 1 Pantheon–Sorbonne.
[‡]Paris School of Economics and University Paris 1 Pantheon–Sorbonne.
[§]Université Ouaga II, Cedres.

population). More generally, when Y is a variable that is contained only in the dataset A, Z is a variable available only in the dataset B, and X is a set of variables contained in A and B, statistical matching allows to impute Y in B and Z in A. It is highly difficult to evaluate whether the impuated values are accurate because, in practice, statistical matching is used when Y and Z are not jointly observed for any unit in A or B.

Traditional statistical matching procedures (Cohen [1991] and Rodgers [1984]) impute the missing values under the assumption that Y and Z are conditionally independent given X, which results in biased results whenever this assumption is not verified. On the other hand, the statistical matching procedure proposed by Rubin (1986) allows to take into account the partial correlation between Y and Z given X. Using real and simulated datasets, Alpman (2016) shows that Rubin's procedure, unlike traditional methods, imputes the missing values accurately enough so that empirical analyses performed with imputed values yield unbiased estimates.

The traditional methods have been used in Jenkins and O'Leary (1996) and Bonke (1992) to combine time-use surveys with consumer expenditure surveys. In more recent applications, Gardes (2014) and Gardes and Starzec (2017) have also used the traditional procedures to combine time-use surveys with consumer expenditure surveys whereas Alpman and Gardes (2016) have used the procedure of Rubin (2016). As time and expenditures were not jointly observed for any unit, these papers were not able to assess the accuracy of the imputed variables (although Alpman and Gardes [2016] show that the distribution of the imputed expenditures is almost identical to the distribution of the observed expenditures). In this paper, we compare the performances of three different statistical matching procedures using a dataset where rural households in Burkina Faso were surveyed regarding their expenditures and their use of time. Our analysis reveals that statistical matching, especially the procedure proposed by Rubin (2016), can be effectively used for combining time-use surveys with surveys informing consumers' expenditures. Therefore, our findings indicate that the results obtained in Alpman (2016) remain valid even for surveys where the relations between the variables are quite complex.

In Section 2 we present the different statistical matching procedures that we use. The dataset in presented in Section 3. We discuss the results in Section 4.

2 Matching Procedures

2.1 The Simple Regression Method

In the simple regression method, Y is regressed on $(1, X)$ using the dataset A, and Z is regressed on $(1, X)$ in the dataset B. Using these regression coefficients, Y and Z can be predicted in B and A, respectively, since X is observed in both datasets. The simple regression procedure is easy to implement but it is subject to several shortcomings. As noted in Frazis and Stewart (2011), this procedure results "in a loss of the variation that is not accounted for by the covariates", that is, the

variance of the imputed values of Y in B is smaller than the variance of Y in A, and the variance of the imputed values of Z in A is smaller than the variance of Z in B.

A greater flaw of the simple regression method is the assumption that Y and Z are independent given X. As a result, "the same Y values, for example, would be imputed to units with different Z and identical X when in fact the imputed Y values should differ with Z for a given X if the partial correlation between Y and Z given X is different than zero" (Alpman, 2016). For instance, if Y is weight, Z is height, and X is age, income and gender, the simple regression method imputes the same weight, regardless the height, as soon as individuals share the same X. Therefore, when the partial correlation between Y and Z given X is different than zero, the simple regression method yields biased results (the bias increases as the partial correlation gets further from zero).

2.2 Statistical Matching Using Cells

In this method, the observations in A and, on the other hand, the observations in B are grouped according to some identical characteristics (e.g., age, education level, geographical location, and family type). The different groups are referred to as *cells*. The mean values of Y computed for each cell in A are imputed to the corresponding cells in B. Thus, the same Y is imputed for all the observations in a given cell. This procedure shares the shortcomings of the simple regression method.

This method has been commonly used to generate pseudo-panel following the seminal work of Deaton (1985): the individuals are replaced by cohorts that are tracked over the years. Browning et al. (1985) group households according to the birth date of the family head (5 years cohorts) and according to a profession characterisitic (blue collars and the others), Blundell et al. (1998) define the cells according to 10 years cohorts and two levels of education, and Propper et al. (2001) chose to group households according to 7 years cohorts and 10 regions. Note that individuals can be grouped into cells according to other non-parametric procedures using, for instance, neuronal Kohonen maps as in Gardes et al. (1996). This method minimizes the measurement errors and it allows to generate long series of cross-sections (e.g., Boelaert et al. [2017] construct a pseudo-panel over 42 years) under an ergodicity assumption (i.e., all the variables have a similar distribution over the period). The main difficulty that arises when using cells is the tradeoff between the number of cells and their homogeneity: a Monte Carlo simulation (Verbeek and Nijman, 1992) suggests that a cell size of minimum 100 individuals diminishes appropriety the measurement errors correlated with the cell size.

2.3 The Procedure of Rubin (1986)

The statistical matching procedure proposed by Rubin (1986) allows to take into account the partial correlation between Y and Z given the common covariates X. As a result, Y is predicted as a function of Z and X, and Z is predicted as a function of Y and X (see Rubin [1986] and Moriarity

and Scheuren [2003] for greater details). Once the missing values are predicted, Rubin (1986) matches the units for which the values are missing with those for which the values are observed by minimizing the distance between the predicted values conditional on a set of covariates. When the matches are identified, the unobserved value is replaced by the observed value of the match. If the partial correlation value is chosen accurately, Rubin's procedure yield unbiased results and the distributions of the imputed values are closer to the true distributions (see Alpman, 2016).

The critical issue in Rubin's procedure is the choice of the partial correlation value between Y and Z given the common covariates X. In this paper, we are able to compute the partial correlation value because all the variables are jointly observed. However, as statistical matching is required when the variables of interest are not jointly observed, the partial correlation value will be unobserved. A possible solution to guess accurately the unknown partial correlation value would be to compute the average values of the variables over some characteristics such as the geographical location and the income. As a result, a representative agent's average amount of time and income allocated to an activity would be jointly observed. This approach allows us to approximate reasonably well (i.e., up to 85%) the partial correlation value between leisure time and leisure expenditures (given the sex, age, marital status, education level, family size, number of children, and total spendings).

3 Data

3.1 Presentation of Burkina Faso

Burkina Faso is a country located in West Africa with a population of 16.9 millions inhabitants where 46% of the individuals have less than 15 years old and the average life expectancy is 56 years. Burkina Faso is ranked 181^{th} out of 187 countries regarding its human development index (United Nation Program for Development evaluation in 2014). The per capita GDP is 720 dollars, mainly concentrated in the service sector (52%); industry and agriculture constitute only 26 and 22% of the GDP, respectively. The economic growth in Burkina Faso is 5% by year since 2000, the unemployment rate is 3%, and 83% of the population is below the poverty line according to the UNPD multidimensional index. 40.1% of the population is under the poverty line defined by the statistical office (INSD, 2015), with 92% of the poors in rural areas.

3.2 The Agricultural Family survey, Descriptive Statistics, and Estimation Procedure

The RGA (Recensement Général de l'Agriculture, 2008) surveys 6,941 households in 71 villages located in 45 provinces. This survey contains information on family characteristics (income from agriculture or other activities, age, number and age of children, education level, accessibility to social services, income, financial situation, equipment etc.), on households' expenditures (over 40

Table 1: Descriptive Statistics.

	Monetary budget share	Time share	Full income budget share	Monetary expenditures (euros)	Full expenditures (money+ time value*) (euros)	Ratio of monetary exp. to time value	Ratio of monetary exp. to time value	Ratio of monetary exp. to time value	Ratio of monetary exp. to time value
	All households						Singles	Two adults, no child	Two adults with children
Food	0.745	0.459	0.716	932 723	989 567	16.41	52.26	20.36	15.30
House	0.079	0.274	0.099	97 755	136 803	2.60	2.32	1.54	1.96
Leisure	0.176	0.267	0.185	219 792	255 586	6.14	9.73	2.99	4.01
TOTAL	1	1	1	1 252 291	1 381 955	9.58	18.14	7.89	7.94

Note: time is valued by the shadow price of time estimated at the individual level.

goods and services), and on the time use over 14 activities: unproductive activities, agriculture in rainy periods, gardening, culture of trees, cattle breeding, fishing, gathering, wood harvesting for selling on a market, wood harvesting for family needs, search for water, market work, other household activities, personal activities, and other activities.

Time uses are recorded for all adults in the family, while expenditures are at the family level, including the children (the numbers of adults and children are in average 5.36 and 5.72). We assume therefore that only adults contribute to the household production. Time allocated to activities such as gardening and cattle breeding are used for personal consumption and to the sale of the products or services on the market. We have no precise information regarding the share of time allocated to personal consumption and to the sale of the products on the market; based on the statistics in Burkina Faso, we assume that 70% of this time corresponds to personal consumption and 30% to the production that is sold on the market.

In this paper, the monetary expenditures and the time-use have been grouped into three activities: food, household activities, and other. The full prices (i.e., the shadow prices of these activities) are computed using the monetary expenditures and the value of time allocated to the activity as explained below. Expenditures are recorded weekly for food and quarterly for other expenditures, while time uses correspond to one week. All have been transformed into yearly values. As family size can be very large (with an average of 5.35 adults per household), time uses corresponding to all adults in the household may be performed in fact by a small part of these households (say two or three). The descriptive statistics in Table 1 indicate indeed that couples with two adults have a significantly greater ratio of monetary expenditures to time-uses than singles. In order to correct this potential bias, time-uses have been multiplied by the ratio of the OECD equivalence scale (one for the first adult and 0.7 for the others) over the number of adults (which perhaps still overstates the true number of adults corresponding to recorded time uses).

We estimate full income and full price elasticities using an Almost Ideal Demand System (AIDS).

We proxy the full income by the full epxenditure; the latter writes $\sum_i (p_i x_i + \omega t_i)$ where $p_i x_i$ is the monetary expenditure on activity i, ω is the shadow price of time, and t_i is the amount of time allocated to activity i. The full price corresponds to the value of time and market goods required to produce an additional unit of activity i. Therefore, the dependent variables are the full budget shares of the activities: $(p_i x_i + \omega t_i)/\sum_i (p_i x_i + \omega t_i)$ The shadow price of time, the full income elasticities, and the price elasticities are estimated using the procedure proposed by Gardes (2014).

4 Results

Table 2 reports the estimates obtained with real values (A), the simple regression method (B), the statistical matching procedure using cells (C), and with Rubin's procedure (D). Regarding the full income elasticities, the three matching procedures yield biases less than 9%. The results show that the income elasticities estimated with procedures (B) and (C) give slightly less biased results than the estimates in (D): when the absolute value of the biases of the income elasticities are averaged, the bias with traditional methods is 4.00% whereas it is 6.26% with Rubin's procedure.

On the other hand, the results show that traditional procedures yield highly biased uncompensated and compensated full price elasticities for food (roughly 21% and 55%, respectively) while Rubin's procedure produce biases below 4%. Regarding the leisure activity, Rubin's procedure produce once more less biased results (below 3%) whereas the biases with the traditional methods are 9%. For the housing activity, the price elasticities are less biased with traditional methods even though, with Rubin's procedure, the biases do not exceed 8.33%.

When we compute the mean of the absolute value of all the biases, we obtain 5.01% for Rubin's procedure against 11.82% for method (B) and 11.99% for method (C). Therefore, Rubin's procedure allows to obtain overall less biased results. In addition, the results obtained with Rubin's procedure are quite consistent, that is, the biases do not exceed 8.36%; with traditional methods, the biases can reach 56.79%. In addition, Figure 1 shows that Rubin's procedure allow to recover the distribution of the missing variables much more accurately than the traditional methods.

Finally, note that time and market goods gets increasingly substitutable as the economy develops (because more market alternatives are available for the use of time in a particular activity). In Burkina Faso, the partial correlation value between time and expenditures is less than 0.013 in absolute value. As this value goes up, the results obtained with traditional methods will get increasingly biased since they assume that time and expenditures are conditionally independent.

5 Conclusion

In this paper, we use a dataset informing simultaneously the use of time and the allocation of income to assess the performances of various statistical matching procedures in the context where the use of time and the allocation of income are informed by different datasests. Combining these

two dimensions allows to conduct empirical research based on the household production theory. Among the various procedures that we have evaluated, the one proposed by Rubin (1986) gave the most accurate results: among all the estimated elasticities, the biases of the estimated coefficients never exceeded 8.5%.

The results obtained with the traditional statistical matching procedures were often close to those obtained with the method proposed by Rubin (1986), which can be explained by the low partial correlation value between expenditures and time use in Burkina Faso. As this partial correlation value gets further away from zero, one should expect that the results obtained with the traditional methods gets more biased. Nevertheless, despite the similar results obtained with the different statistical matching procedures, the highest biases with the traditional methods were above 50%.

Traditional statistical matching procedures operate often with fitted values: thus, the distribution of the imputed values differ from the distribution of the real values and the variance of a variable reduces during the imputation process. On the other hand, the operation of Rubin's procedure allows to recover accurately the distribution (and therefore the variance) of a variable. Whenever the partial correlation value between the variables that are jointly unobserved is chosen accurarelty, the procedure of Rubin (1986) allows to combine time quite accurately an expenditure survey with a survey informing the use of time. Regarding the calibration of the partial correlation value, we show that aggregating individual observations according to some specific characteristics (in order to jointly observe expenditures and time-uses) such as geographical location and income allows to approximate reasonably well the unknown partial correlation value.

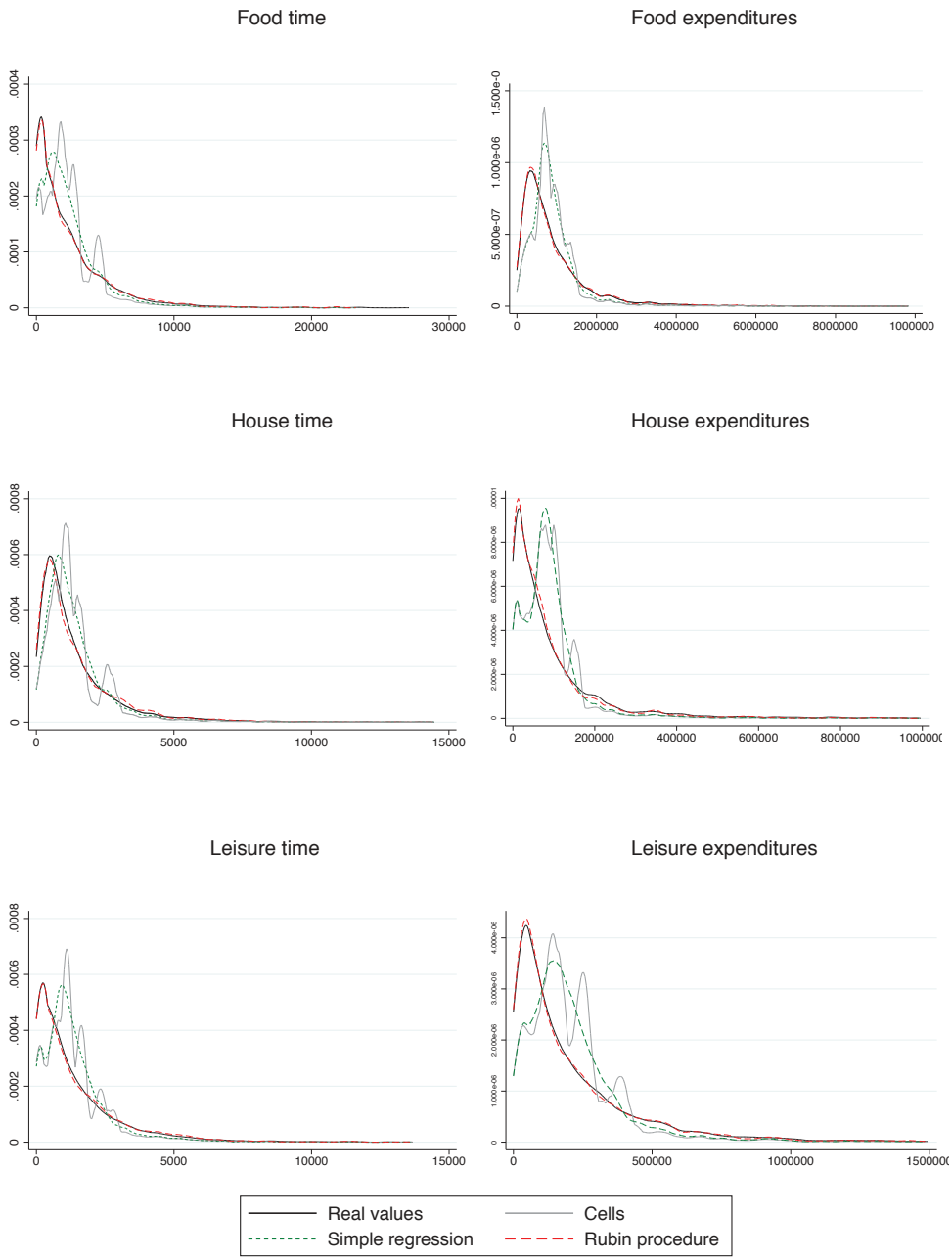


Figure 1: Real vs Imputed Values.

Table 2: Full income and full price elasticities.

	Full income elasticity				Uncompensated full price elasticity				Compensated full price elasticity			
	(A)	(B)	(C)	(D)	(A)	(B)	(C)	(D)	(A)	(B)	(C)	(D)
Food	1.072***	1.051***	1.044***	1.110***	−1.165***	−1.419***	−1.400***	−1.188***	−0.442***	−0.693***	−0.680***	−0.456***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.012)	(0.013)	(0.012)	(0.015)	(0.013)	(0.014)	(0.013)	(0.016)
Housing	0.766***	0.793***	0.809***	0.702***	−1.164***	−1.160***	−1.155***	−1.072***	−1.069***	−1.068***	−1.061***	−0.980***
	(0.012)	(0.012)	(0.012)	(0.015)	(0.013)	(0.011)	(0.010)	(0.013)	(0.014)	(0.011)	(0.010)	(0.013)
Leisure	0.903***	0.942***	0.957***	0.841***	−1.224***	−1.324***	−1.329***	−1.193***	−1.043***	−1.141***	−1.143***	−1.018***
	(0.011)	(0.010)	(0.010)	(0.013)	(0.011)	(0.012)	(0.012)	(0.015)	(0.011)	(0.013)	(0.013)	(0.016)
					Difference with respect to the real estimates							
Food		−1.96%	−2.61%	3.54%		21.80%	20.17%	1.97%		56.79%	53.85%	3.17%
Housing		3.52%	5.61%	−8.36%		−0.34%	−0.77%	−7.90%		−0.09%	−0.75%	−8.33%
Leisure		4.32%	5.98%	−6.87%		8.17%	8.58%	−2.53%		9.40%	9.59%	−2.40%

Notes: standard errors are in brackets. $*, **, ***$ indicate significance different from zero respectively at 90%, 95% and 99% confidence. (A) denotes the results obtained with real values; (B) denotes the results obtained with values predicted by simple regression; (C) denotes the results obtained with values predicted by regression using cells; and (D) denotes the results obtained by Rubin's procedure.

References

Alpman, A. (2016). Implementing Rubin's alternative multiple-imputation method for statistical matching in Stata. *Stata Journal*, 16, 717–739.

Alpman, A. and F. Gardes (2016). Welfare Analysis of the Allocation of Time During the Great Recession. CES Working Papers, 2015.12

Blundell, R., Duncan, A., and C. Meghir (1998). Estimating Labor Supply Responses Using Tax Reforms. *Econometrica*, 66, 827–861.

Boelaert, J., Gardes, F., and S. Langlois (2017). Convergence des consommations entre classes socioéconomiques et contraintes non monétaires au Canada. Forthcoming in *L'Actualité Economique*.

Bonke, J. (1992). Distribution of Economic Ressources: Implications of Including Household Production. *Review of Income and Wealth*, 38, 281–293.

Browning, M., Deaton, A., and M. Irish (1985). A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle. *Econometrica*, 53, 503–544.

Cohen, M. L. (1991). Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling. In *Best Practices in Quantitative Methods*, ed. C. Citro and E. Hanushek. National Academy Press.

Deaton, A. (1986). Panel Data from a Time Series of Cross Sections. *Journal of Econometrics*, 30, 109–126.

Frazis, H. and J. Stewart (2011). How Does Household Production Affect Measured Income Inequality? *Journal of Population Economics*, 24, 3–22.

Gardes, F. (2014). Full price elasticities and the value of time: A Tribute to the Beckerian model of the allocation of time. CES Working Papers, 2014.14.

Gardes, F. and C. Starzec (2017). Individual prices and Households'size: a Restatement of Equivalence Scales Using Time and Monetary Expenditures Combined. Forthcoming in *The Review of Income and Wealth.*

Gardes, F ., Gaubert, P., and Rousset, P. (1996). Cellulage de Données d'Enquête de Consommation par une Méthode Neuronale. Prépublication du SAMOS n.69.

Gardes, F ., Langlois, S., and D. Richaudeau (1996). Cross-Section vs Time-Series Income Elasticities: an Estimation on a Pseudo-Panel of Canadian Surveys. *Economic Letters*, 51, 169–175.

Jenkins, S. and N. O'Leary (1996). Household Income Plus Household Production: The Distribution of Extended Income in the U.K. *Review of Income and Wealth,* 42, 401–419.

Moriarity, C. and F. Scheuren (2003). A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 21, 65–73.

Propper, C., Rees, H., and Green K. (2001). The Demand for Private Medical Insurance in the UK: A Cohort Analysis. *The Economic Journal,* 111, C180–C200

Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91–102.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 4, 87–94.

Verbeek M. and T. Nijman (1992). Can cohort data be treated as genuine panel data? *Empirical Economics*, 17, 9–23.