Soc Indic Res
DOI 10.1007/s11205-012-0198-2

Calibrating Time-Use Estimates for the British Household Panel Survey

Cristina Borra · Almudena Sevilla · Jonathan Gershuny

Accepted: 1 November 2012

Abstract This paper proposes an innovative statistical matching method to combine the advantages of large national surveys and time diary data. We use data from two UK datasets that share stylised time-use information, crucial for the matching process. In particular, time-diary information of an individual from the Home On-line Study, our donor data set, is imputed to a similar individual from the British Household Panel Survey, our recipient dataset. Propensity score methods are used in conjunction with Mahalanobis matching to increase matching quality.

Keywords Statistical matching · Propensity score · Mahalanobis distance · Childcare time

1 Introduction

This paper suggests an innovative approach to increase the power of stylized time-use data usually found in surveys. Current research distinguishes two methods of measuring time-use: direct, stylized questioning and time-diary methods.[1] Stylised questions are typically incorporated into national surveys. However, this kind of data usually does not cover all

[1] Recently, researchers also consider experience sampling methods whereby respondents record what they are doing at randomly selected moments of time (Juster et al. 2003; Gershuny 2004).

C. Borra (✉)
Department of Economics and Ec. History, Universtity of Seville, Campus Ramón y Cajal, 41018 Sevilla, Spain
e-mail: cborra@us.es

A. Sevilla
School of Business and Management, Queen Mary, University of London, Francis Bancroft Building, Mile End Road, London E1 4NS, UK

J. Gershuny
Deparment of Sociology, Centre for Time-Use Research, University of Oxford, Manor Road Building, Manor Road, Oxford OX1 3UQ, UK

Published online: 18 November 2012

types of daily activities (e.g., Kan and Pudney 2008; Kan and Gershuny 2009). On the contrary, diaries are seen as the most reliable and comprehensive sources of time budgets (Juster et al. 2003). However, keeping a diary is more complex, more expensive, and more time-consuming than obtaining stylised time-use information and therefore diaries are not usually included in most surveys (Bonke 2005; Kan and Gershuny 2009; Schulz and Grunow 2012). As stated by Kan and Pudney (2008), researchers are frequently faced with the dilemma of opting for detailed and presumably more reliable time-use data at the cost of severe constraints on the type of research that can be done or accepting poorer quality, less wide-ranging time-use data, to give them greater research scope.

In this paper we propose an innovative statistical matching method to solve the dilemma by combining the advantages of large national surveys and time diary data. To do this, we use data from two UK datasets: the Home On-line Study (HOL) and the British Household Panel Survey (BHPS). The 1999–2001 HOL is a unique longitudinal data set, which contains both stylized and diary estimates of time devoted to a wide range of activities. It has a smaller sample and covers a shorter period than the BHPS, though. The BHPS gives only stylized estimates of time use devoted to a few activities. Nonetheless, this information can be used to calibrate time-use estimates for BHPS respondents.

The proposed statistical matching procedure uses variables common to both data sets to identify similar individuals in order to generate a new synthetic dataset (Kum and Masterson 2008). In particular, information of an individual from the HOL, our donor data set, is imputed to a similar individual from the BHPS, our recipient dataset. A special feature of our method is that the set of common variables used in the matching process includes the stylised time-use variables present in both surveys.

The practical application of this paper focuses on the important issue of childcare time, although the methods proposed here can be used for other activities such as leisure, personal care, or education time. Unlike housework time and market work time, which are more easily obtained from stylised questions, the BHPS does not contain information on childcare time. Our method provides an alternative way to relying in indirect proxies like working status or number of children for accounting for mothers' time investments in children. This technique can be used to calibrate time devoted to different activities for other longitudinal surveys, such as the PSID and the SOEP, using the auxiliary time use data sets for these countries.

Unlike traditional data imputation techniques which merge evidence across datasets using variables unrelated to the variables of interest, our study contributes to the literature by using a wide range of stylised time use variables (about paid work, childcare responsibilities, and housework) present in both surveys. A previous study by Kan and Gershuny's (2009) also make use of this information. Our study improves on this and former analyses (Sutherland et al. 2002; Bloemen et al. 2010) by using propensity score matching techniques to impute the missing information. These techniques are preferred over OLS regression predictions because, first, they preserve the variability of the data (Peichl and Schaefer 2009); second, they focus the researcher's attention on the direct comparability of the recipient and donor datasets trough direct confrontation of the common support condition (Dehejia and Wahba 2002); and third, given that this method is non-parametric estimates are therefore less sensitive to the choice of functional form in the model (Zhao 2008). With respect to other matching strategies, the use of a propensity score overcomes the dimensionality problem that arises if many covariates are used for matching (Rosenbaum and Rubin 1983). We use recent developments in matching theory to increase matching quality by using Mahalanobis matching on key covariates together with

propensity score matching techniques (Rubin and Thomas 2000; Zhao 2004; Stuart and Rubin 2008).

This paper is organized as follows. Section 2 describes the data and the methodology used in the analysis. Section 3 presents the results and assesses the quality of the matching. Section 4 performs the robustness check. Section 5 concludes.

2 Data and Methods

HOL is a three-wave household panel data conducted annually in 1999, 2000, and 2001, and containing about 1,000 households drawn from a national random sample (with an over-sample of computer users). As stated by Kan (2008), this study has two main distinctive advantages: first, it contains both stylised estimates and diary-based estimates of time spent on market work and housework; and second, it collected 7-day diaries from respondents, while other studies usually collect only 1- or 2-day diaries. This second feature is especially important, given recent criticisms about time diary studies being able to adequately describe infrequent activities (Gershuny 2012). HOL collected around 2,300 weekly diaries, covering 16,100 diary days.

The BHPS is an annual survey that interviews all members of a random selection of about 5,000 households and 10,000 individuals. From 1994 (Wave 4) onwards, the BHPS asked respondents a series of stylised time-use questions, including usual weekly paid work hours and housework hours, and the distribution of various domestic and childcare tasks within the households.

In both datasets, samples were selected including mothers aged 18–64 with children present in the household. In our main study sample, only waves 9–11 (1999–2001) of the BHPS are used, but we include an analysis for the whole 1994–2006 BHPS dataset in the robustness check section. Our sample is thus composed of 404 observations from HOL and 7,265 observations from BHPS, in the time-restricted case, and 27,538, in the unrestricted case.

Data fusion (also known as statistical matching) provides a means of combining information from different sources into a single data set. In essence, it uses variables common to both data sets to identify similar records that can be linked in order to generate a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets. In particular, the associations between variables never jointly observed are often the main motivation for interest in such a complete, but synthetic, data set (Kum and Masterson 2008). In particular, our aim is to impute values for a variable that is missing in the recipient dataset, the BHPS, from the donor dataset, the HOL. Thus we are concerned with what the literature refers to as unconstrained matching, given that the base recipient file and the supplemental donor file are treated asymmetrically (Ridder and Moffitt 2007).[2] In our study both datasets have a collection of common variables that are labelled X. We want to add one variable, childcare time C, from the HOL to the BHPS. With unconstrained matching, to every unit i in the BHPS we match the unit j in the HOL. It is possible that some unit in the donor HOL is matched to more than one unit in the recipient BHPS, and that some units in the HOL are not matched to any unit in the BHPS. As a consequence, the distribution of X,C in the matched file may differ from that in the original sample (Ridder and Moffitt 2007).

[2] In the alternative constrained matching all the records in both dataset are represented in the matched file. To accomplish this, the units in both samples are replicated to the population size.

We may think of different methods for imputing the missing variable in the recipient dataset. One is the regression method (Rubin 1986). In this approach, the specific variable from the donor dataset C is regressed on the vector of common variables X:

$$C = X\beta + v \tag{1}$$

The estimated coefficients β from the donor dataset are then used to predict the values of C in the recipient dataset (Peichl and Schaefer 2009).[3] In this case, the imputed measures are not values observed on a "similar" individual who participates in the survey, but are simply estimates. The main advantage of this strategy is its simplicity. Connelly and Kimmel (2009) note that one of its disadvantages is that the variance in the imputed variable is lost since it is a predicted value based on estimated coefficients.[4] Also, as long as it is a parametric method, the predicted variable will require the use of exclusion restrictions for identification when used in future analysis in the merged dataset (Angrist and Pischke 2009).

Statistical matching is the other broad option. Matching involves pairing units from different datasets. It uses variables common to both datasets to identify similar records that can be linked in order to generate a new synthetic data set that allows more flexible analysis than would be possible with the two discrete data sets. In particular, the associations between variables never jointly observed are often the main motivation for interest in such a complete, but synthetic, dataset (Kum and Masterson 2008). As stated by Judson and Poppoff (2004), the problem of imputing values for a variable that is missing may be thought to be analogous to constructing a pseudo control group for an experimental design study when a random assignment between treatment and control groups is not possible.

One of the advantages of these matching methods over regression is that the variation in the imputed variable that occurs in the donor dataset is simulated as closely as possible, given that a unique donor amount can be found for each recipient record (Connelly and Kimmel 2009). Another benefit is that it restricts inferences to samples for which there is overlap in covariate distributions across data sets (the common support region), thereby avoiding unwarranted model extrapolations (Dehejia and Wahba 2002). Finally one more advantage of matching over regression analysis is that it is non-parametric: matching does not impose functional form restrictions such as linearity and homogeneous effects on the distribution of X, C, both assumptions being usually unjustified either by economic theory or by the data (Zhao 2008). Moreover matching does not require exclusion restrictions for the identification of the imputed variable when used in the combined dataset.

Exact matching is only possible for a reduced number of common variables. However, when dealing with multiple covariates it becomes very difficult to find matches with close or exact values of all covariates. Different methods have been devised to summarize the information in the covariates in just one scalar. The two most popular are the propensity score and the Mahalanobis metric (Stuart and Rubin 2008).

To implement propensity score methods, both datasets are combined and a dummy variable is constructed taking value one if the observation belongs to the recipient file BHPS and value 0 if the observation belongs to the donor file HOL. The propensity score is formally defined as the probability of belonging to the recipient database conditional on the common observed covariates:

[3] Kan and Gershuny (2009) and Connelly and Kimmel (2009) include examples of this method.

[4] See the Rubustness Check section for recent developments in this methodology which overcome this drawback.

$$p(X_i) = \Pr(i \in BHPS | X = x). \tag{2}$$

Rosenbaum and Rubin (1983) show that at each value of the propensity score, the distribution of the covariates X defining the propensity score is the same in the recipient and donor groups. In other words, conditioning on covariates and conditioning on the propensity score will both make the distribution of the covariates in the recipient group the same as the distribution of the covariates in the donor group.

The Mahalanobis metric is a measure of dissimilarity between observations. The Mahalanobis metric measures the distance between units i from the recipient dataset BHPS and j from the donor dataset HOL weighting each coordinate of X in inverse proportion to the variance of that coordinate (Zhao 2004):

$$d_M = (X_i - X_j) D^{-1} (X_i - X_j)' \tag{3}$$

Where D is the variance–covariance matrix of X.

Gu and Rosenbaum (1993) and Rubin and Thomas (2000) compare the performance of matching methods based on Mahalanobis metric matching and propensity score matching, and they find that the two distance measures perform similarly when there are a relatively small number of covariates, but that propensity score matching works better than Mahalanobis metric matching with large numbers of covariates (greater than 5). Nonetheless Zhao (2004) reports that when the sample size is too small, propensity score matching does not perform well compared with Mahalanobis matching, which is relatively robust to specification error.

One possible solution to these complexities considered in the literature is combining both distance measures in the matching algorithm (Lechner 2002; Zhao 2004; Stuart and Rubin 2008). This can be especially illuminating for cases where there are some key covariates on which particularly close matches are desired (Rubin and Thomas 2000). Mahalanobis matching on the key covariates can be combined with propensity score matching to improve matching quality (Stuart and Rubin 2008).

This is the approach followed in this paper. Previous studies on the determinants of mothers' childcare time have underlined the importance of market work and the age of the child in deciding mothers' time spent with their children (Zick and Bryant 1996; Baydar et al. 1999; Bittman 2004; Joesch and Spiess 2006). Thus, in performing the matching we would like to obtain very high quality pairings with respect to these two variables that are believed to be particularly predictive of the variable of interest childcare time. In order to do so we follow the subsequent matching protocol (Rosenbaum and Rubin 1985; Rubin and Thomas 2000; Lechner 2002): We first specify and estimate a binomial probit model of the probability of belonging to the BHPS sample, that is, we obtain the propensity score. Second, we restrict the BHPS sample to observations whose estimated propensity score lies within the ranges of estimated propensity scores of the HOL sample, that is, we impose the common support condition. Third, all HOL subjects within intervals surrounding each BHPS subject's estimated propensity score are identified as potential matches. And finally, Mahalanobis metric matching is applied on the propensity score and the key covariates considered: usual working hours and age of youngest child. Rubin and Thomas (2000) underline that Mahalanobis metric matching on key covariates within relatively coarse propensity score calipers is an effective method for ensuring matching quality while allowing researchers to use information about the relative prognostic value of different covariates. After the matching process, we check the quality of the matching. Rosenbaum and Rubin (1983) suggest that we compare the mean covariate values in the groups, i.e. that each of the observable covariates within the treatment group has the same average value within the matched control group. Before matching we expect differences, yet after

matching the variables should be balanced in both groups and significant differences should not persist. Equality of the first moments does not imply equality of the entire distribution of covariates, but for binary variables—and most of the covariates used here are of this type—there is no need to compare higher order moments.

3 Results

Consistent with our matching strategy we first estimate the propensity score. We run a probit regression of the binary indicator taking value 1 for observations in the BHPS sample (and 0 for the HOL sample) over the set of common variables. In particular we consider demographic and personal characteristics of the mother (a quadratic in age, highest educational level in three categories, civil status), household characteristics (number of children in different age brackets), reported time-use behaviour (average weekly working hours, a quadratic in usual housework hours, and, for those living in couples, whether childcare is a shared responsibility with their partners). We also include year indicators and a control for computer use to take account of over sampling in the HOL survey.

Table 1 displays the results from the probit model of the likelihood of belonging to the BHPS sample. Many of the variables considered significantly explain sample membership; interestingly, all the variables computed with stylised time-use information.

We impose the common support assumption next. It implies that, for each individual in the recipient database, there is another individual from the donor database who can be used as a matched comparison observation. This can be tested graphically. Figure 1 shows the propensity score histogram in both data sets. As can be observed, given the high degree of overlap between the two distributions, for the large majority of the treated individuals there is a similar control group individual, in such a way that the common support assumption is satisfied.[5]

Mahalanobis matching within balanced propensity score intervals is then performed. Within each block, we pair each recipient unit with that donor for which the Mahalanobis distance metric is lower. We consider three variables in the computation of this metric: the propensity score, the average weekly working hours, and the age of the youngest child.

In order to analyze the quality of the matching Rosenbaum and Rubin (1983) suggest that we check whether significant differences between the average values of the variables for both groups exist after matching. We follow the test of stratification suggested by Dehejia and Wahba (2002). We divide the observations into blocks based on the estimated propensity score and we check whether within each block significant differences in the distribution of each of the explanatory variables persist. Table 6 in the Appendix displays the mean values of the variables used in the analysis for both datasets before matching and the significance of the *t* test of equality of means after matching in each of the blocks considered.[6] Overall, the figures in Table 6 confirm that both samples, though initially somewhat different, look quite similar after matching, with no significant differences in the majority of measured background characteristics remaining between the two groups.[7] We therefore conclude that in general the distribution of the common variables is balanced after matching.

As a further assessment of the virtues of this method, the distribution of the imputed variable is compared to the distribution of the original variable and the distribution of a simple OLS regression estimation of it. Table 2 presents descriptive statistics of the three

[5] 120 BHPS observations had to be discarded because they had no HOL close matches available.

[6] We used the procedure developed by Becker and Ichino (2002) for STATA. We obtained 9 blocks.

[7] In fact, none of the differences are significant at the 1 % level.

Table 1 Propensity score coefficient estimates

Variables	Label	Coef.	SE
Demographic and personal variables			
Age	Age	−0.1811	0.068***
agesq	Age squared	0.0013	0.001
Married		0.1185	0.158
Education levels			
Degree_further	=1 degree	0.3286	0.165**
Alev	=1 A-level	−0.7139	0.178***
Olev	=1 O-level	−0.6156	0.154***
Less	=1 less than O-level	Ref	
Household characteristics			
nch02	Number children 0–2 years	−0.3853	0.146***
nch34	Number children 3–4 years	−0.4186	0.133***
nch511	Number children 5–11 years	0.0039	0.072
Time use and childcare behaviour			
howlng	Housework time	0.0365	0.011***
howlngsq	Housework time squared	−0.0003	0.000*
tothrs	Working hours	0.0115	0.004***
joint_childcare	=1 childcare joint responsability	−0.3454	0.209*
tot_joint	interaction tothrs*joint_childcare	−0.0159	0.007**
Design variables			
wave10	=1 year 2000	0.7029	0.131***
wave11	=1 year 2001	0.6282	0.129***
cd8use	=1 computer use	−0.5440	0.132***
cons		7.5264	1.290***

This table shows the probit regression of treatment status on available covariates. The samples include all mothers 18–64, where mother is defined as having a child under the age of 15 in the house

* Significant at 10 %; ** significant at 5 %; *** significant at 1 %

variables. The distribution of the Childcare variable estimated by the matching method resembles the distribution of the original HOL variable better than the OLS prediction. In particular, the variation in the original variable is closely simulated, as measured by the estimated standard deviation. In addition, the matching method is utterly superior to the simple OLS regression in that no negative childcare hours are predicted and in that the estimated proportion of mothers reporting zero hours of childcare is closer to the original variable, as evidenced by the almost negligible proportion mothers with OLS-estimated zero (or negative) childcare hours. Also, our matching procedure preserves better the joint distribution of Childcare and the two key covariates, total working hours and age of the youngest child, as evidenced by the estimated correlations between the variables.

4 Robustness Check

In order to test the sensitivity of our results with respect to the sample selection criteria being used, we perform the matching process using the unrestricted BHPS sample

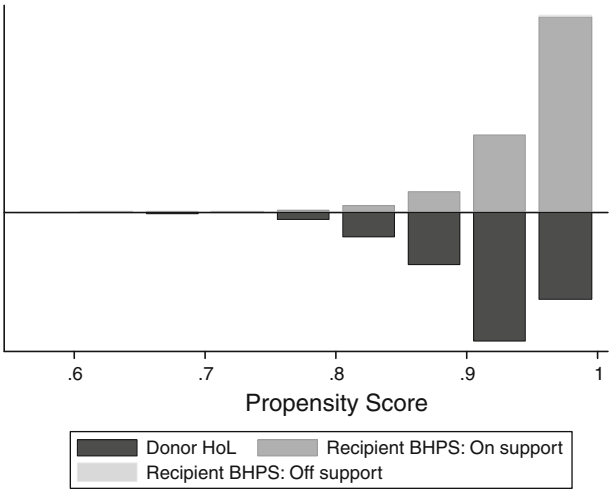

Fig. 1 Distribution of the estimated propensity score in both datasets

Table 2 Descriptive statistics of original and estimated variables (1999–2001 sample)

Variable	No. obs	Mean	SD	Min	Max	Proportion zero childcare	Correlation with working hours	Correlation with age of youngest child
Original childcare	404	15.18	19.61	0.00	128.50	0.141	−0.28	−0.50
Matching childcare	6,668	18.00	19.63	0.00	97.00	0.090	−0.29	−0.57
OLS childcare	6,788	17.44	10.79	−3.86	51.52	0.002	−0.35	−0.90

The samples include all mothers 18–64, where mother is defined as having a child under the age of 15 in the house. All estimates are adjusted to take into account sample weights

(1994–2006) together with the HOL sample. In this case, the calibration involves assuming that time-use patterns among different social groups have not changed significantly over the years, as well supported by past research (Gershuny 2000).

In the estimation of the propensity score, the year dummies were obviously not included. Additional interaction variables were used as explanatory variables in order to reach convergence in Dehjia and Wahba's (2002) proposed algorithm. Nonetheless, results were quite similar to those reported in Table 1 for the restricted sample.[8]

Table 3 reports the distribution of the estimated matched variable, together with the distribution of the original variable and an OLS regression estimation of it. Again the

[8] The additional variables were interaction terms between total working hours and age, civil status, education categories, number of children, childcare responsibility, and computer use. Results are available upon request.

Table 3 Descriptive statistics of original and estimated variables (1994–2006 sample)

Variable	No. obs	Mean	SD	Min	Max	Proportion zero childcare	Correlation with working hours	Correlation with age of youngest child
Original childcare	404	15.18	19.61	0.00	128.50	0.141	−0.28	−0.50
Matching childcare	25,206	17.97	19.69	0.00	128.50	0.077	−0.24	−0.52
OLS childcare	25,206	17.58	11.06	−7.69	57.40	0.012	−0.40	−0.90

The samples include all mothers 18–64, where mother is defined as having a child under the age of 15 in the house. All estimates are adjusted to take into account sample weights

matched estimates resemble the original variable more closely with regard to variability of the data, proportion of zero outcomes, and correlation with key covariates.

It has been shown that asymptotically all matching algorithms should yield the same results, because with growing sample size they all become closer to comparing only exact matches (Smith 2000). However, in small samples the choice of the matching procedure can be important. So additionally we test the sensitivity of our results to the choice of matching algorithm. In our case it could be interesting to compare results obtained using only propensity score matching or Mahalanobis matching. In doing so, we try to assess the degree of accuracy of the proposed methodology.

Table 4 presents the results for propensity score radius matching and Mahalanobis matching and compares them to our previous Mahalanobis-within-propensity-score-callipers results. Mahalanobis matching uses only the information from our key covariates: market work time and age of the youngest child. The standard deviation figure shows that the variability in the data is even greater than in our original HOL sample. We may interpret that this fact reflects the lack of precision of the estimates compared to our preferred methodology. However, there are no observations off the common support in this case; in this matching procedure the closest unit, in terms of Mahalanobis distance, from the donor dataset is always used as matching unit. The joint distribution between the estimated Childcare variable and our key covariates is not improved however, given that the values of the estimated correlations are further from the original correlations in the HOL dataset (see figures in Table 2). In the radius matching case, given that we use all the observations in the donor dataset within a calliper of the propensity score to estimate the matching unit, we find a clear increase in precision which is reflected by the reduction in the standard deviation figure. This feature comes at a price, however. The proportion of mothers with zero estimated childcare hours drops dramatically, the number of observation off the common support increases significantly,[9] and the correlations with the key covariates clearly depart from previous figures. From our point of view, the method used in the main analysis combines the benefits of both propensity score and Mahalanobis matching. On one hand it obtains really good matches with respect to key covariates; on the other, it prevents matching individuals with very different characteristics with respect to other features.

[9] Instead of the 120 cases in our preferred method, radius matching leaves 360 observations out of the analysis.

Table 4 Descriptive statistics of estimated variables

Method	No. obs	Mean	SD	Min	Max	Proportion zero childcare	Correlation with working hours	Correlation with age of youngest child
Mahalanobis-within-callipers matching	6,688	18.00	19.63	0.00	97.00	0.090	−0.29	−0.57
Mahalanobis matching	6,788	20.52	23.54	0.00	97.00	0.080	−0.33	−0.59
Radius matching	6,328	17.73	13.28	0.00	85.00	0.029	−0.04	−0.06

Different matching algorithms. 1999–2001 sample

The samples include all mothers 18–64, where mother is defined as having a child under the age of 15 in the house. All estimates are adjusted to take into account sample weights

Table 5 Descriptive statistics of estimated variables

Method	No. obs	Mean	SD	Min	Max	Proportion zero childcare	Correlation with working hours	Correlation with age of youngest child
Mahalanobis-within-callipers matching	6,668	18.00	19.63	0.00	97.00	0.090	−0.29	−0.57
Variance-adjusted OLS	6,788	16.51	17.08	−37.44	85.02	0.16	−0.18	−0.53

Recently developed methods (1999–2001 sample)

The samples include all mothers 18–64, where mother is defined as having a child under the age of 15 in the house. All estimates are adjusted to take into account sample weights

Recent developments in regression imputation methods add random residuals to the regression estimates, therefore overcoming the problem with the lack of variability in regression estimates –obviously at the cost of loosing its simplicity (van Buuren et al. 1999; Räessler 2002). So as a final assessment of the virtues of our proposed methodology, we compare the distribution of the variables estimated by our Mahalanobis-within-callipers matching procedure and the alternative variance-adjusted OLS method (van Buuren et al. 1999).[10] The results are shown in Table 5. Both methods give comparable results, with respect to the marginal distribution of Childcare, in terms of variability of the data and proportion of zero outcomes (as long as predicted negative outcomes are considered zero in the regression scheme). However, as expected, our Mahalanobis-within-calipers method outperforms variance-adjusted OLS with respect to the joint distribution of Childcare and our key covariates. In particular, the correlation obtained for Childcare and market work is significantly lower than that found in the original donor sample (see Table 2) when variance-adjusted regression is used. That is logical, provided that our proposed methodology

[10] We used the Stata program uvis to compute this last estimation (Royston 2004).

is intended to precisely obtain very good matches with respect to especially decisive covariates. In addition, as previously stated, it is important to bear in mind that, compared to any regression routine, matching methods are non-parametric and thus less sensitive to the choice of functional form in the model (Zhao 2008) and free from requirements of exclusion restrictions for identification, if the predicted variable is used in future analysis in the recipient dataset (Angrist and Pischke 2009). Also, they focus the researcher's attention on the common support condition, thus ensuring that only comparable individuals are compared (Dehejia and Wahba 2002).

5 Conclusions

This paper proposes an innovative statistical matching method to combine the advantages of large national surveys and time diary data. In particular we use stylized time use data, usually found in large national datasets, to match similar individuals from the national survey and the diary data. The proposed method involves performing Mahalanobis matching on special key variables within relatively coarse intervals of the estimated propensity score.

The method is tested using two UK datasets, the BHPS and the HOL, that share stylized time use information. We focus on the important issue of childcare time, estimating time devoted to that activity for all mothers in the BHPS sample. In fact matching estimates are utterly superior to simple OLS estimates with respect to variability of the data, proportion of zero outcomes, and joint distribution of estimated values and key covariates. They are comparable to variance-adjusted regression methods with respect to the marginal distribution of the estimated variable, while better preserving the original joint distribution of the estimated variable and the especially considered key covariates. In addition, being a non-parametric method, it avoids relying on specific functional form assumptions, whereas its focus on the region of common support, avoids unwarranted model extrapolations. The proposed method also combines the benefits of propensity score and Mahalanobis matching, emphasizing the influence of key covariates while preventing matching individuals with very dissimilar characteristics with respect to other variables.

The methods proposed here can be expanded to calibrate diary time devoted to different activities such as leisure, personal care, or education. It also opens new avenues for imputing time use information for other longitudinal surveys, such as the PSID and the SOEP, using the auxiliary time use data sets for these countries.

Acknowledgments This paper has benefited from comments provided by Man Yee Kan, David Berrigan, and Oriel Sullivan. Any remaining errors are our own. This paper was partly prepared while Dr. Borra was an Academic Visitor in the Centre for Time Use Research at the University of Oxford (summer 2009, autumn 2010). The authors would like to express their thanks for the financial support provided by the Spanish Ministry of Education and Science (Project ECO2008-01297 and Movility Grant "José Castillejo" Convocatoria 2010), by the Andalusian Government (Convocatoria IAC 2009 2), and by the Economic and Social Research Council (Grant Number RES-060-25-0037).

Appendix

See Table 6.

Table 6 Matching quality

Variable	Unmatched sample				Matched sample					
	Recipient	Donor	% bias	*t* test	*t* test.4th block	*t* test 5th block	*t* test 6th block	*t* test 7th block	*t* test 8th block	*t* test 9th block
Personal and demographic variables										
Age	35.43	38.34	−41	−7.58***	1.33	−0.92	0.19	1.21	0.21	−0.76
agesq	1309.4	1516.1	−39.4	−7.54***	1.26	−0.81	0.17	1.16	0.03	−0.73
Married	0.79	0.85	−14.9	−2.70**	−0.76	0.39	0.51	−0.24	0.30	−0.02
Education levels										
degree_further	0.37	0.26	22.5	4.13***	1.36	0.05	1.10	−0.77	−0.40	−0.23
Alev	0.12	0.19	−18.9	−3.93***	−0.29	0.74	0.74	0.00	−0.15	0.57
Olev	0.26	0.35	−21.3	−4.25***	−0.62	−0.37	0.17	0.14	1.49	−1.23
Household characteristics										
nch02	0.24	0.19	11.5	2.08**	−1.37	−0.52	0.77	0.33	−0.08	1.56
nch34	0.22	0.22	1	0.20	−1.81*	−0.15	0.65	−1.05	1.90*	−0.84
nch511	0.81	0.78	3.7	0.68	−1.00	−0.01	−0.46	1.45	0.38	−1.48
age_youngest	6.04	7.87	−39.5	−7.47***	−2.08**	−0.37	−0.40	−0.15	−1.05	0.74
Time-use and childcare behaviour										
howlng	19.31	17.29	16.8	3.11**	−1.76*	−0.53	0.35	1.03	0.49	−1.57
howlngsq	526.83	433.13	12.3	2.20*	−0.86	−0.74	0.56	0.80	0.39	−1.02
tothrs	18.28	19.87	−9.1	−1.75	1.56	0.51	−0.57	−0.13	−0.04	−0.09
joint_childcare	0.22	0.33	−26.4	−5.36***	1.08	2.28**	−0.42	0.26	−1.96*	−1.07
tot_joint	5.27	9.14	−26.7	−5.69***	1.01	−1.52	0.03	−0.03	−2.09**	0.02
Design variables										
wave9	0.33	0.46	−27.8	−5.45***	0.97	0.01	−0.17	−1.18	1.62	1.66
wave10	0.34	0.26	18.7	3.45***	−0.88	0.73	1.10	1.04	0.39	−0.26
wave11	0.33	0.28	10.7	2.00*	1.82*	0.63	1.19	0.22	−1.94*	−1.31
cd8use	0.64	0.76	−27.4	−4.97***	−1.56	0.62	1.18	−0.01	−1.07	1.62

This table shows the difference in means before the matching and the *t* test for whether any significant differences remain in the different blocks after the matching. The stata program pscore (Becker and Ichino 2002) used to perform the calculations

* Significant at 10 %; ** significant at 5 %; *** significant at 1 %

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricists companion*. Princeton: Princeton University Press.

Baydar, N., Greek, A., & Gritz, M. R. (1999). Young mothers' time spent at work and time spent caring for children. *Journal of Family and Economic Issues, 20*, 61–84.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal, 2*(4), 358–377.

Bittman, M. (2004). Parenting and employment what time-use surveys show. In Michael. Bittman & Nancy. Folbre (Eds.), *Family time: The social organization of care* (pp. 152–170). London, New York: Routledge.

Bloemen, H., Pasqua, S., & Stancanelli, E. (2010). An empirical analysis of the time allocation of Italian couples: Are they responsive? *Review of Economics of the Household, 8*(3), 345–369.

Bonke, J. (2005). Paid work and unpaid work. Diary information versus questionnaire information. *Social Indicators Research, 70*, 349–368.

Connelly, R., & Kimmel, J. (2009). Spousal influences on parents' non-market time choices. *Review of Economics of the Household, 7*(4), 361–394.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*(1), 151–161.

Gershuny, J. (2000). *Changing times: Work and leisure in postindustrial society*. Oxford: Oxford University Press.

Gershuny, J. (2004). Costs and benefits of time sampling methodologies. *Social Indicators Research, 67*, 247–252.

Gershuny, J. (2012). Too many zeros: A method for estimating long-term time-use from short diaries. *Annales d'Ėconomie et de Statistique, 105*(106), 247–270.

Gu, X., & Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405–420.

Joesch, Jutta. M., & Spiess, K. (2006). European mothers' time spent looking after children—Differences and similarities across nine countries. *Electronic International Journal of Time Use Research, 3*(1), 1–27.

Judson, D. H., & Poppoff, C. L. (2004). Selected general methods. In J. S. Siegel & D. Swanson (Eds.), *The methods and materials of demography* (pp. 667–732). San Diego, CA: Elsevier.

Juster, F. T., Ono, H., & Stafford, F. P. (2003). An assessment of alternative measures of time use. *Sociological Methodology, 33*, 19–54.

Kan, M. Y. (2008). Measuring housework participation: The gap between "stylised" questionnaire estimates and diary-based estimates. *Social Indicators Research, 86*(3), 381–400.

Kan, M. J., & Gershuny, J. (2009). Calibrating stylised time estimates using UK diary data. *Social Indicators Research, 93*, 239–243.

Kan, M. Y., & Pudney, S. (2008). Measurement error in stylized and diary data on time use. *Sociological Methodology, 38*, 101–132.

Kum, H. & Masterson, T. (2008). Statistical matching using propensity scores: Theory and application to the levy institute measure of economic wellbeing. Economics working paper archive wp_535, Levy Economics Institute. http://www.levyinstitute.org/pubs/wp_535.pdf. Accessed December 15, 2010.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics, 84*, 205–220.

Peichl, A., & Schaefer, T. (2009). FiFoSiM: An integrated tax benefit microsimulation and CGE model for Germany. *International Journal of Microsimulation, 2*(1), 1–15.

Räessler, S. (2002). *Statistical matching: A frequentist theory, practical applications and alternative Bayesian approaches*. New York: Springer.

Ridder, G., & Moffitt, R. (2007). The econometrics of data combination. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6, pp. 5469–5547). Amsterdam: Elsevier.

Rosenbaum, P., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33–38.

Royston, P. (2004). Multiple imputation of missing values. *Stata Journal, 4*, 227–241.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics, 4*, 87–94.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95*, 573–585.
Schulz, F. & Grunow, D. (2012). Comparing diary and survey estimates on time use. *European Sociological Review, 28*(5), 622–632.
Smith, J. (2000). A critical survey of empirical methods for evaluating active labor market policies. *Swiss Journal of Economics and Statistics, 136*(3), 247–268.
Stuart, E. A., & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155–176). Thousand Oaks, CA: Sage Publications.
Sutherland, H., Taylor, R., & Gomulka, J. (2002). Combining household income and expenditure data in policy simulations. *Review of Income and Wealth, 48*(4), 517–536.
van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*, 681–694.
Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics, 86*(1), 91–107.
Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters, 98*(3), 309–319.
Zick, C. D., & Bryant, W. K. (1996). A new look at parents' time spent in child care: Primary and secondary time use. *Social Science Research, 25*, 1–21.