Imputation of repeatedly-observed multinomial variables in longitudinal surveys

André Berchtold[1,2] & Joan-Carles Surís[2]

[1] Institute of Sciences Sociales & NCCR LIVES, University of Lausanne, SSP / Géopolis, CH-1015 Lausanne, Switzerland

[2]Research Group on Adolescent Health, Lausanne University Hospital & University of Lausanne, Biopôle 2, Route de la Corniche 10, CH-1010 Lausanne, Switzerland

Andre.Berchtold@unil.ch, Joan-Carles.Suris@chuv.ch

Corresponding author: André Berchtold

Abstract: It is now a standard practice to replace missing data in longitudinal surveys with imputed values, but there is still much uncertainty about the best approach to adopt. Using data from a real survey, we compared different strategies combining multiple imputation and the chained equations method, the two main objectives being 1) to explore the impact of the explanatory variables in the chained regression equations, 2) to study the effect of imputation on causality between successive waves of the survey. Results were very stable from one simulation to another, and no systematic bias did appear. The critical points of the method lied in the proper choice of covariates and in the respect of the temporal relation between variables.

Keywords: Longitudinal survey, missing data, multiple imputation, chained equations, causality.

1 Introduction

This article was motivated by a practical issue encountered during the analysis of the TREE (Transitions from Education to Employment) survey on adolescents. TREE is a longitudinal Swiss study surveying the post-compulsory educational pathways of young people and their entry into the labor market (TREE, 2008). It is based on an initial sample of 6343 youths who participated in the PISA survey in 2000 (Program for International Student Assessment; OECD (1999); Buschor, Gilomen & McCluskey (2003)) and who where then followed up yearly from 2001 (wave 1, T1) to 2007 (wave 7, T7). As in many surveys, there are missing data in the TREE database: Only 63% of the original 2001 sample was still in the study in 2007. Missing data are known to be a problem for all statistical analyses, but this was especially true in our case, since we wanted to track cannabis and tobacco consumption trajectories using statistical tools such as Markov chains. For that purpose, we needed to work on complete trajectories. In practice, however, only 1131 subjects did answer to questions regarding substance use continuously from 2001 to 2007, what was too few for our purpose.

The TREE project being inscribed in a longitudinal framework, data are certainly not independent one wave to another. It was then necessary to take the possible inter-wave relations into account in order to provide estimated values of missing data coherent not only with respect to other variables of the same wave (some of them also including missing data), but also from one wave to another. Moreover, the collection of the TREE data being not finished at the time of this study, at least two additional waves being planned in 2010 and 2014, the mechanism used to estimate missing data on waves 2001 to 2007 should ideally be usable again with the new waves, without having to reestimate the value of missing data in the preceding waves.

Using TREE as an example, we describe in detail an imputation method adapted to the longitudinal context. We chose to rely on multiple imputation, the different completed datasets being generated by the chained equations method. We describe first the base imputation method designed for our analyses, then we discuss six alternative methods differing by the number of covariates and by the importance given to the respect of causality between successive waves. The whole method is evaluated through a series of numerical experiments using simulated missing data from the TREE database. The chained equation method has been already studied in detail in the past (White, Royston & Wood, 2011) and its usefulness is well established, so our main objectives are to compare alternate specifications of the regression equations used during the imputation process, and to explore the influence of the imputation process upon causal models.

2 Missing data and the imputation process

2.1 Missing data

According to Little & Rubin (2002), there are three kinds of missing data: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). We speak of MCAR when the missing data are a random sample of the entire dataset, a very rare situation. MAR means that the probability for a particular data to be missing does not depend on the missing data, but only on other variables of the dataset. For instance, girls could refuse to report their weight more often than boys, the fact of reporting or not reporting the weight being independent from the weight itself. MNAR means that the probability of missing is influenced by the real non-observed value of the data. For instance, the probability of not reporting income could be positively correlated with the income level. Diagnostics exist to assess the missing data mechanism (Hedeker & Gibbons, 1997). When missing data are MCAR, unbiased estimates of the parameters of interest can be obtained by simply removing the missing data. When the general mechanism of missing data is MAR, then it is possible to build statistical models for the relation between several variables, and to use these models to obtain unbiased estimations of the missing data.

In the remaining of this paper, we will work with MCAR simulated missing data. This is certainly not the most frequently assumed missing data mechanism, but since our main goal was to explore particular elements of imputation rather than the whole chained equation process, using a simple missing mechanism suppressed the risk of adding useless noise to the discussion. Moreover, as noted e.g. by Rubright, Nandakumar & Glutting (2014), MCAR data are not always easier to impute than MAR data. Notice also than even the MAR hypothesis can be difficult to justify in a longitudinal framework, some subjects maybe leaving the survey at some point because of the value taken by a variable of interest (Schafer & Graham, 2002), so we implemented the two following mechanisms: First, when several successive values were missing for a subject, only the first of them was imputed, the other ones being kept as missing. Second, we relied on the multiple imputation approach. This approach is known to work well with MCAR and MAR data, but correct results can be achieved even in the MNAR case (Collins, Schafer & Kam, 2001; Demirtas, 2004). Moreover, as noted by Graham (2009), if the MAR assumption is violated, not only the multiple imputation method will be affected, but the majority of missing data treatments. Since we cannot test for the MNAR mechanism, it is then difficult to decide when methods able to handle MNAR data, such as likelihood based methods, should be used in place of multiple imputation.

2.2 Multiple imputation

Many approaches for handling missing data are described in the literature, starting from very simple ones (listwise deletion of all cases with at least one missing data, replacement of the missing data by the mean of observed values) to modern model-based methods (Allison, 2001; Little & Rubin, 2002; Graham, 2009). A distinction must be made between single- and multiple-imputation: In the first case, each missing data is replaced by a single imputed value, the result being a new dataset without missing data called the *completed* dataset. A classical method is to use the value maximizing the log-likelihood of a statistical model, and the well-known Expectation-Maximization (EM) algorithm (McLachlan & Krishnan, 1996) can therefore be used. The objective of this paper is not to review the different options which can be used to deal with missing data, but the interested reader will find much information in e.g. Graham (2012); McKnight et al. (2007) and Van Buuren (2012)

When each missing data is replaced by a single estimated value, the variance of statistical parameters computed from the completed dataset are generally underestimated, what leads to incorrect confidence intervals and p-values. Regardless of the imputation method, the cause of this problem, as noted by Schafer & Olsen (1998), is that replacing missing data by only one particular estimated value does not take into account the inherent variability of missing data. One of the best approaches to overcome this issue is now the multiple imputation (MI) procedure proposed by Rubin (1987). Its principle is to generate not one, but $m > 1$ estimations of each missing value and to constitute then m different sets of completed data called *replications*. The main advantage of this method is to preserve the variability of the missing information. Of course, the algorithm used to create the m replications has to be able to create different imputed values for each missing data. In practice, accurate results can be achieved with m as low as 3 or 5, but we chose here to work in a conservative framework and to use $m = 10$.

During the statistical analysis of the data, each model is computed separately on the m replications and results are then combined into a single final statistic using the rules defined by Rubin: Let θ be the parameter to be estimated; from each of the m replicated datasets, we obtain an estimation $\hat{\theta}_i$. The MI estimator of θ is then

$$\hat{\theta}_{MI} = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m}$$

The variance of the MI estimator is obtained as a combination of the variance of each $\hat{\theta}_i$ and the variance between the $\hat{\theta}_i$. If $\hat{V}_i$ is the variance of $\hat{\theta}_i$, then

$$\hat{V}_{MI} = \frac{\sum_{i=1}^{m} \hat{V}_i}{m} + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta}_{MI})^2$$

2.3 Chained equations

In the context of multiple imputation, different alternative methods can be used for the estimation of the missing values. A possibility is to use bootstrap (Efron, 1982), but two much more usual methods are now the so-called data augmentation (Tanner & Wong, 1987) and chained equations (Van Buuren, Boshuizen & Knook, 1999) approaches. Similarly to the EM algorithm, these methods rely on an iterative algorithm, but on the contrary of EM, they converge in distribution, so that even after convergence values obtained from a given iteration can differ from the ones of the previous iterations. This feature makes them perfect choices for the generation of as many replications as necessary of the missing data values.

The principle of multivariate imputation by *Chained Equations* (CE) was proposed by Van Buuren, Boshuizen & Knook (1999). It can be seen as a variant of predictive mean matching (Rubin, 1986; Little, 1988), but it is much more general. Compared to alternative approaches such as EM, propensity scores and standard data augmentation, CE seems best suited to longitudinal data in the context of multiple imputation. EM cannot be used here, since it generally always converges to the same solution, except when multiple local optima do exist in the solution space (Berchtold, 2002). Propensity scores do not follow some of the basic hypotheses regarding multiple imputation, leading sometimes to biased results (Allison, 2000). Finally, by relaxing the assumptions made in data augmentation about the distribution of error terms, chained equations can be applied to a wider range of situations (Durrant, 2005).

Chained equations work as follows:

1. Regression models are defined for each variable containing missing data. Continuous and/or categorical dependent variables can be treated simultaneously; for instance, some of the regressions can be linear, while others are binary logistic. Explanatory variables are independently chosen for each regression among the other dependent variables. Covariates which do not appear as dependent variables in a regression can also be used as explanatory variables.

2. Each missing data is first replaced by a random value.

3. Each regression model is estimated and used to provide an imputation of the missing data.

4. The algorithm iterates several times through all the regression models, missing values being each time replaced by the values imputed during the preceding iteration.

5. The imputed values obtained during the last iteration are replaced by the closest values really observed in the dataset in a way similar to hot deck imputation.

Step 5 above could be replaced by other approaches, such as drawings from the normal distribution for instance, but since we considered only categorical variables here, our method presents the advantage of generating only possible values. There are no real guidelines concerning the number of iterations to be used, but 10 is generally considered as a good choice (Van Buuren, Boshuizen & Knook, 1999) and we used this value here. Repeating the whole process m times leads to m different datasets which can then be used for multiple imputation.

3 Simulation study

We concentrate here on a categorical variable representing the average tobacco consumption level during the 30 days preceding the survey. This multinomial variable takes five modalities (never, 1-3 times a month, 1-2 times a week, 3-5 times a week, daily). The number of available values ranges from 5335 in 2001 to 3150 in 2007, but only 1999 complete trajectories are available.

Simulations have been conducted in order to compare the quality of the imputations produced by different forms of chained regression equations. Before imputation, we first looked at the relation existing a) between the successive waves of the variable to be imputed, b) between covariates and the variables to be imputed. Results are reported as Spearman's ρ when comparing the successive observations of tobacco level consumption, as Cramer's V for the relation between tobacco consumption level and nominal covariates, and as η^2 for the relation between tobacco consumption level and continuous covariates.

3.1 The base chained equation imputation method

Many differences exist between cross-sectional and longitudinal data, among them the fact that longitudinal data can be used to assess causality between events. Theoretically, causality can be established if and only if the three following rules are verified (Hill, 1965; Schneider, Carnoy, Kilpatrick, Schmidt & Shavelson, 2007): 1) the cause precedes the effect; 2) there is a strong association between the cause and the effect; 3) all other possible causes have been removed or controlled for. During the imputation process of missing data, it is crucial not to modify the link between possible causes and effects, either in the sense of an attenuation or of an increase of causality. We do not want to produce spurious causalities by reinforcing the association between a possible cause and its expected effect, nor do we want to diminish causality by breaking the temporal link between the cause and the effect. To preserve the possibility to later establish causality, we considered the following rules for the imputation of time dependent variables:

- A regression model is defined for each variable with missing data.

- For each wave, the predictors are composed of i) the lags of the variable to be imputed and ii) a set of covariates, either invariant in time or taken from the previous waves of the survey.
- No imputation is made for the missing data of the first wave, because no previous observations of the same variable are available, so imputation would be based on covariates only, what we consider as too imprecise.
- The observation preceding an imputed value cannot be itself missing. When two or more than two successive observations are missing, only the first one is imputed.

This approach was used to impute missing values for the consumption level of tobacco each year. Even if this variable is ordinal rather than nominal, with five modalities ranging from *never* to *daily*, we chose to use multinomial rather than ordinal regressions, the parallelism assumption required for ordinal regression being rejected here. In addition to the consumption level of tobacco the preceding years, the following variables taken from the 2000 PISA survey, one year before the first wave of the TREE study, were used as time-invariant covariates: Gender, age in months, swiss language area (French / German / Italian), country of birth (Switzerland / other), family wealth (scale), school track (pre-gymnasial / extended requirements / basic requirements / no selection).

Consider for instance the following four subjects for which we simply indicate whether the tobacco consumption level variable is observed (O) or missing (.) on each wave:

Wave	1	2	3	4	5	6	7
Subject A	O	O	O	O	O	O	O
Subject B	O	O	O	O	O	.	O
Subject C	O	O	.	O	.	.	.
Subject D	.	O	O	O	O	O	O

Subject A has all data, so imputation is useless. Subject B has only one missing data on wave 6 and it can be imputed using our method. Subject C has four missing data on waves 3, 5, 6 and 7. The first two can be imputed using our method, but not the last two, because these missing data are themselves preceded by another missing data. Finally, no imputation is possible for the missing data of subject D, because it occured in the first wave. From this example, the rules we fixed for imputation can be seen as very restrictive, but they should ensure a minimal impact on the possible causality between events. Moreover, the importance of these rules will be checked through the use of the six alternative regression shemes defined in the next section.

3.2 Alternative methods

One of our major concerns is the role and the influence of the predictors in the regression models. To evaluate this influence, we defined the six following alternative sets of regression models:

1. Similar to the base method, but without the covariates Gender, Age and Language area.
2. Similar to the base method, but without the covariates Country of birth, Family wealth and School track.
3. Similar to the base method, but with four additional covariates taken from the PISA survey: Level of literacy competence (low / high), family structure (nuclear / other), index of possession of culture items (scale), index of family educational support (scale).
4. Similar to the base method, but the tobacco consumption level of wave t+1 is also used to impute the tobacco consumption level of wave t, except for wave 1.
5. Similar to the base method, but the tobacco consumption level of wave t+1 is also used to impute the tobacco consumption level of wave t, including wave 1.
6. Similar to the base method, but the tobacco consumption level of any wave is explained by the tobacco consumption level of the six other available waves.

Methods 1-3 were defined to investigate the importance of the choice of covariates upon the results, while methods 4-6 were defined to evaluate the importance of keeping a strict temporal relation between waves. For the purpose of comparability with the base method, the rule defined above specifying that when two or more than two successive observations are missing only the first one is imputed was also used.

3.3 Simulation procedure

The first point to be examined when speaking about imputation is the quality of the imputation as can be measured by the difference between values obtained from the imputation and the corresponding truly observed values. The difficulty of course lies in the fact that in real situations, where imputation would typically be used, we do not know the true values! It is then very important to investigate this aspect of the imputation process by the mean of simulated data. The following procedure was adopted:

1. The TREE observations for which the level of tobacco consumption was continuously available from wave 1 to wave 7 were selected (n=1999).

2. Fifty datasets were created with 10% of missing data randomly generated on each wave.

3. The missing data of each of the 50 datasets were imputed using the base method and the six alternative methods described before.

4. The distribution of the tobacco consumption level was computed for each wave using i) only the 10% of imputed data; ii) the whole variable including the 10% of imputed data and the 90% of observed data.

5. Each distribution was compared to the corresponding fully observed distribution. Comparisons were made for both the overall difference between observed and imputed distributions, and for the relative difference defined as the percentage of over or under estimation of the imputed distribution compared to the observed one.

6. To study the impact on causality of the different imputation methods, we used the consumption level of tobacco during waves 1 to 6 (either the six waves or only one of them) to explain the same consumption level on wave 7 through a multinomial regression model. The difference of fit between the model computed on the whole observed dataset and the models computed on data including imputed values was then investigated through the analysis of the adjusted Nagelkerke's pseudo R^2.

The whole process was also repeated using 50 datasets with 20% of missing data on each wave. Results are reported as boxplots showing the distribution of the difference between really observed and imputed data. It may be important to recall that since we are working within the general framework of multiple imputation, the result for each dataset is an average using Rubin's rule of the results computed separately on each of the m=10 replications.

As with most surveys, sampling weights are provided for the TREE data in order to work with a truly representative sample of the population. Nevertheless, since our goal here is not to produce results valid for the Swiss adolescent population, but to investigate the properties of an imputation approach, we chose not to use weights. In that way, inherent issues related to the computation of weights and to their adjustment to accomodate for longitudinal frameworks and for missing data cannot impact our results. On the other hand, it must be remembered throughout the lecture of the Results section that proportions related to tobacco consumption are just given here as examples and that they do not directly apply to the Swiss adolescent population.

All computations were performed using Stata version 10 (StataCorp., 2007) and the ICE implementation of CE developed for Stata 9 and later by Royston (2004), but since the release of version 12, ICE has become a standard feature of the multiple imputation procedure implemented

in Stata. Similar results can also be achieved for instance in R with the MICE library (Van Buuren & Groothuis-Oudshoorn, 2011).

4 Results

We summarize hereafter the main findings of our study. More complete results can be found in Berchtold & Surís (2012).

4.1 Association between variables

Table 1 summarizes the association between the seven waves of the tobacco consumption level variable. The consumption level of any wave is highly positively and significantly correlated with the consumption level of any other wave. As expected, the association decreases as the lapse of time between two waves increases. On the other hand, the correlation tends to be larger between the last waves, as compared to the first ones.

TABLE 1 ABOUT HERE.

Similarly, Table 2 gives the relation between each wave of tobacco consumption level and the different covariates used during the imputation process. The level of significance is very variable, ranging from highly significant to non-significant associations. All covariates are significantly associated at the 95% level with tobacco consumption in at least one wave. Family structure is associated with all seven waves and two other covariates (Literacy competence, Family wealth) are associated with six waves. Even when significant, the level of association stays always low.

TABLE 2 ABOUT HERE.

4.2 Base method

Figures 1 to 4 report the main findings for the simulations conducted with the base method on the tobacco consumption level. Figures 1 and 3 report results computed on the 10% or 20% of simulated missing data only, while Figures 2 and 4 report results computed on the n=1999 available data points, including a large majority of non-missing data.

The distribution of the 50 differences for any wave and any category of consumption is always approximatively Gaussian, as indicated by a Kolmogorov-Smirnov test of normality at the 95% level (results not shown). The mean is always very close to zero, indicating that the imputation is not biaised toward a systematic over- or under-estimation of some of the categories.

As could have been expected, the magnitude of the error is related to the magnitude of the proportion to be estimated. In absolute value, the error is generally larger for the "Never" modality, which is the most frequent one as indicated by the observed data, but the relation is not strictly observed and the magnitude of the error is sometimes similar for quite different proportions. On the other hand, when looking at the relative rather than the absolute difference between observed and imputed values, the error becomes logically larger for rare modalities. A few number of imputations are outliers, the most extreme case being a relative over-estimation of more than ten times encountered for the "3-5 times a week" modality at T4 with 10% of missings. Fortunately, considering the whole set of simulations, such situations are very unusual.

In practice only a small percentage of data are missings. When considering the distribution not only of the imputed data, but of all available data (Figures 2 and 4), the importance of outliers decreases even more. More generally, these two figures show that even in term of relative difference and with 20% of missing data, the impact of outlier imputations stays always smaller than ± 0.2 for the most extreme cases, more than 50% of the simulations presenting errors smaller than ± 0.05.

Finally, in terms of absolute difference and considering the 10% or 20% of imputed data, the variability represented by the full range of the boxplots (Figures 1 and 3) is smaller with 20% rather than 10% of missing data. This is an indication of the quality of the imputation process. Even with a doubled rate of missing data, the quality of the imputation is only slightly reduced: if a larger number of missing data implies an increase of the variability between imputed values, this effect is more than counterbalanced by the doubling of the number of data points, resulting in a decrease of the standard deviation. When considering the boxplots computed from the whole set of 1999 observations (Figures 2 and 4), the distributions with 10% of missing data are less variable than the distributions with 20% of missing data, because the sample size is equal in both cases, the only difference being the increase of the variability in the 20% situation as compared to the 10% one.

FIGURES 1 TO 4 ABOUT HERE.

4.3 Alternative methods

Figure 5 summarizes results from the simulations comparing the base method (method 0) with the six alternative methods (methods 1 to 6 as described before). Boxplots for the absolute difference between observed and imputed data are provided for waves T2, T4 and T7 and for datasets with 10% of missing values. Results with 20% of missing values are not provided here, since they do not really differ from the ones with 10% of missing values, but more complete results can be found in Berchtold & Surís (2012). Results are shown for the "Never" category, which is the most frequent

one, and the "3-5 times/week" category which is one of the two least frequent categories.

FIGURE 5 ABOUT HERE.

Overall, no method does systematically achieve the best results and the level of over- or under-estimation is comparable for all methods. The range of the distribution computed over the 50 simulations is not related to the particular method used. Depending on the category, the wave, and the percentage of missing data, each method can provide a larger or smaller range of results than alternatives. Moreover, even if outliers remain rare, they can also appear with any method.

This lack of difference between the seven methods can be due to at least two causes: First, the covariates used during the imputation process are insufficiently predictive of the level of tobacco consumption, so adding or removing covariates has only a marginal effect. Second, the high correlation observed among successive values of the tobacco consumption level implies that without any other covariate, the results achieved by all the methods are similar. Using posterior waves during the imputation process does not have a visible impact on the results in terms of distribution.

4.4 Causality

Longitudinal data can be used to assess causality between successive events. It is then crucial to ensure that the imputation process remains neutral regarding the relationship between waves. More specifically, we checked whether the use of one or several posterior waves for the estimation of wave t does impact the temporal relation between waves. A multinomial regression model was computed for the level of tobacco consumption on wave 7 using the level of tobacco consumption on waves 2 to 6 as explanatory factors. Wave 1 was not used here, since most of the alternative imputation methods do not impute missing data on this particular wave. The quality of fit of each model is assessed by Nagelkerke's pseudo-R^2. For the purpose of comparison, we also computed the same models on the 1999 original data without missing values. The model using explanatory waves 2 to 6 achieved a pseudo-R^2 of 0.4935, while the models using only one explanatory wave achieved 0.4294 and 0.3663 for the explanatory wave 6 and 5 respectively. Figure 6 summarizes the results.

Results show a very clear difference between methods 0-3 based on a strict respect of the temporality, and methods 4-6 breaking this relationship. This difference tends to be larger when 20% of the data are missing. The average results of methods 4-6 are always better than the results of methods 0-3, method 6 showing generally the highest values. In many cases, methods 4-6 even obtain a pseudo-R^2 better than the one computed on the original data without missing values. On the other hand, such a situation is never observed when using methods 0 to 3. These results clearly

demonstrate that not respecting the temporality between waves during the imputation process has a direct impact on statistical models trying to identify a possible causal mechanism between waves.

FIGURE 6 ABOUT HERE.

5 Conclusion

The imputation method described in this paper leads to accurate results in a longitudinal context. Longer sequences of non-interrupted data can be obtained at a relatively small computational cost, implying the possibility of analysing more complete datasets, and to put more easily into evidence causal effects. The use of the chained equation approach presents the advantage of seamlessly combining variables of different kinds (for instance continuous and categorical) into a same imputation procedure. This approach is perfectly suited for the needs of multiple imputation, since convergence in probability lets obtain different reliable imputed values for a same missing observation.

The success of the whole imputation process is directly related to both the selection of a thoretically sound method regarding the missing data, and the use of enough information to apply this method reliably. Since the chained equation methods relies on regression models to explain the missing values, covariates must be correlated with the dependent variable, as for any traditional regression. We conjecture that the use of covariates much more associated to the level of tobacco consumption would have helped improve the quality of imputation, but such covariates were either not available in the dataset, or they themselves contained too many missing data to be useful in the context of a simulation study.

Accurate imputations at the individual level are very difficult to achieve, but fortunately the distributional level is generally sufficient, since statistical studies are rarely concerned by each case taken separately. It follows that the distributional properties of imputations are of great interest. In this paper, and even if this was already known since it is a consequence of the method itself, we can see that imputations obtained through the chained equation method are generally Gaussian-distributed around the true value. There is no systematic bias towards an over- or under-estimation of some particular categories of the variable, so the method can be considered as reliable.

Another crucial point lies in the relation between the imputed data themselves and the statistical use of these data. When the study design is purely cross-sectional, any method aiming at improving the quality of the imputations could be used. In particular, we have seen in Section 4.3 that the use of posterior waves during the imputation process does not have a visible impact on the resulting distributions. On the other hand, the situation is completely different when the research design is longitudinal and when causality is of interest. As clearly stated by Hill (1965), one of the

crucial points to be respected when speaking about causality is the temporality: the cause always precedes the consequence. It follows that any intervention related to the temporal order can have consequences, and that the temporal relationship between waves should be strictly respected during the imputation process. Even if it could be tempting to improve the accuracy of point imputations by using not only past, but also future information, this approach can potentially modify the possible causal link between waves, leading to the impossibility to later demonstrate a causality, or inversely to lead to a spurious causality. This point is very important as we have shown that the consequences of the non-respect of the temporality are not always visible. For instance, results obtained by methods 4 to 6 do not appear as really different from the ones of methods 0 to 3, the problem appearing only when the same data are used in a longitudinal context (Figure 6).

The correct handling of missing information is a key point in the proper use of longitudinal data like the TREE survey, the final goal being to obtain more useful statistical results. In that sense, the results presented here should prove interesting to any user of longitudinal datasets including missing information. Even if imputation cannot provide exact replacement values for individual subjects, this methodology is able to provide larger data files to be analyzed at the aggregated level without systematic bias. It can in particular be recommended to anyone needing to analyze long uninterrupted sequences of longitudinal data with the purpose of identifying specific types of trajectories.

Acknowledgments

This research was supported by the Swiss Tobacco Prevention Fund and by the Swiss National Centre of Competence in Research LIVES, which is financed by the Swiss National Science Foundation.

References

Allison PD (2000) Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods Research*, 28:301-309.

Allison PD (2001) *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences 136. Thousand Oaks: Sage.

Berchtold A (2002) Optimisation of Mixture Models: Comparison of Different Strategies. *Computational Statistics*, 19 (3), 385-406.

Berchtold A, Surís JC (2012) Multiple imputation in a longitudinal context: A simulation study using the TREE data. *Cahiers Recherche et Méthodes*, 1. University of Lausanne, Switzerland. Available online at http://www.unil.ch/webdav/site/consultation-statistique/shared/Cahiers_CREM/CREM_1.pdf

Buschor E, Gilomen H, McCluskey H (2003) PISA 2000 - Synthèse et recommandations. Neuchâtel: OFS/CDIP. Available online at http://www.pisa.admin.ch/bfs/pisa/fr/index/hidden_folder/publications.Document.26277.pdf

Collins LM, Schafer JL, Kam CM (2001) A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods* 6: 330-351.

Demirtas H (2004) Modeling Incomplete Longitudinal Data. *Journal of Modern Applied Statistical Methods* 3: 305-321.

Durrant GB (2005) A Semi-Parametric Multiple Imputation Data Augmentation Procedure. *Proceedings of the Survey Research Methods Section, ASA*. Available at http://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSM2005-000425.pdf

Efron B (1982) *The Jacknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: The Society for Industrial and Applied Mathematics.

Graham JW (2009) Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* 60: 549-576.

Graham JW (2012) Missing Data Analysis: Analysis and Design. New York: Springer.

Hedeker D, Gibbons RD (1997) Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies. *Psychological Methods* 2: 64-78.

Hill AB (1965) The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* 58:295-300.

Little RJA (1988) Missing-Data Adjustments in Large Survey. *Journal of Business & Economic Statistics*, 6(3): 287-296.

Little RJA, Rubin DB (2002) *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley & Sons.

McKnight PE, McKnight KM, Sidani S, Figueredo AJ (2007) *Missing Data: A Gentle Introduction.* New York: Guilford.

McLachlan GJ, Krishnan T (1996) *The EM Algorithm and Extensions.* New York: John Wiley & Sons.

OECD (1999) Measuring Student Knowledge and Skills: A New Framework for Assessment. Paris: OECD Publications Service. Available online at http://www.pisa.oecd.org/dataoecd/45/32/33693997.pdf

Royston PB (2004) Multiple imputation of missing values. *The Stata Journal*, 4:227-241.

Rubin DB (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics*, 4(1): 87-94.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Rubright JD, Nandakumar R, Glutting JJ (2014) A Simulation Study of Missing Data with Multiple Missings X's. *Practical Assessment, Research & Evaluation*, 19(10): 1-8.

Schneider B, Carnoy M, Kilpatrick J, Schmidt WH, Shavelson RJ (2007) *Estimating Causal Effects Using Experimental and Observational Designs.* (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Schafer JL, Graham JW (2002) Missing Data: Our View of the State of the Art. *Psychological Methods* 7: 147-177.

Schafer JL, Olsen MK (1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research* 33: 545-571.

StataCorp. (2007) Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* 82:528-550.

TREE (ed.) (2008) TREE Project Documentation 2000-2008. Berne/Basel: TREE. Available online at http://tree.unibas.ch/en/the-project/description/

Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* 18: 681-694.

Van Buuren S, Groothuis-Oudshoorn K (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45(3): 1-67.

Van Buuren S (2012) *Flexible Imputation of Missing Data.* Boca Raton: Chapman & Hall/CRC.

White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377:399.

Table 1: Association between the seven waves of the tobacco consumption level variable. Only observations for which the level of tobacco consumption is continuously available from T1 to T7 are used (n=1999). We provide for each comparison the value of the Spearman correlation. All correlations are significant at the 99.9% level.

				Wave			
Wave	T1	T2	T3	T4	T5	T6	T7
T1	1						
T2	0.724	1					
T3	0.678	0.793	1				
T4	0.619	0.729	0.791	1			
T5	0.595	0.683	0.726	0.801	1		
T6	0.566	0.659	0.700	0.789	0.829	1	
T7	0.553	0.621	0.667	0.728	0.774	0.818	1

Table 2: Relation between the seven waves of the tobacco consumption level variable and covariates used during the imputation process. Only observations for which the level of tobacco consumption is continuously available from T1 to T7 are used (n=1999). For categorical covariates, we provide the value of the Cramer's V association coefficient and its associate p-value (in italic). For continuous covariates, we provide the η^2 association measure and the p-value of the corresponding one-way ANOVA.

	Wave						
Covariates	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Categorical:							
Gender	0.0812	0.0854	0.0735	0.0675	0.0640	0.0929	0.0492
	0.010	*0.006*	*0.029*	*0.058*	*0.085*	*0.002*	*0.304*
Language area	0.0715	0.0475	0.0643	0.0344	0.0665	0.0399	0.0435
	0.009	*0.341*	*0.036*	*0.786*	*0.024*	*0.608*	*0.477*
Country of birth	0.0510	0.0553	0.0602	0.0547	0.0829	0.0703	0.0718
	0.270	*0.191*	*0.125*	*0.201*	*0.008*	*0.043*	*0.036*
School track	0.0756	0.0638	0.0861	0.0624	0.0572	0.0466	0.0471
	0.001	*0.018*	*0.006*	*0.025*	*0.075*	*0.370*	*0.348*
Literacy competence	0.0651	0.0752	0.0929	0.0893	0.0984	0.1137	0.0983
	0.076	*0.024*	*0.002*	*0.003*	*0.001*	*<0.001*	*0.001*
Family structure	0.1076	0.1155	0.1197	0.0995	0.1088	0.1137	0.0767
	<0.001	*<0.001*	*<0.001*	*0.001*	*<0.001*	*<0.001*	*0.020*
Continuous:							
Age	0.0069	0.0013	0.0034	0.0028	0.0021	0.0021	0.0024
	0.008	*0.619*	*0.142*	*0.227*	*0.392*	*0.375*	*0.316*
Family wealth	0.0068	0.0059	0.0084	0.0068	0.0042	0.0034	0.0076
	0.009	*0.016*	*0.003*	*0.007*	*0.062*	*0.0166*	*0.005*
Cultural possessions	0.0030	0.0030	0.0045	0.0075	0.0045	0.0040	0.0075
	0.179	*0.231*	*0.052*	*0.004*	*0.065*	*0.094*	*0.004*
Educational support	0.0065	0.0041	0.0059	0.0024	0.0041	0.0036	0.0000
	0.009	*0.080*	*0.022*	*0.259*	*0.074*	*0.144*	*0.970*

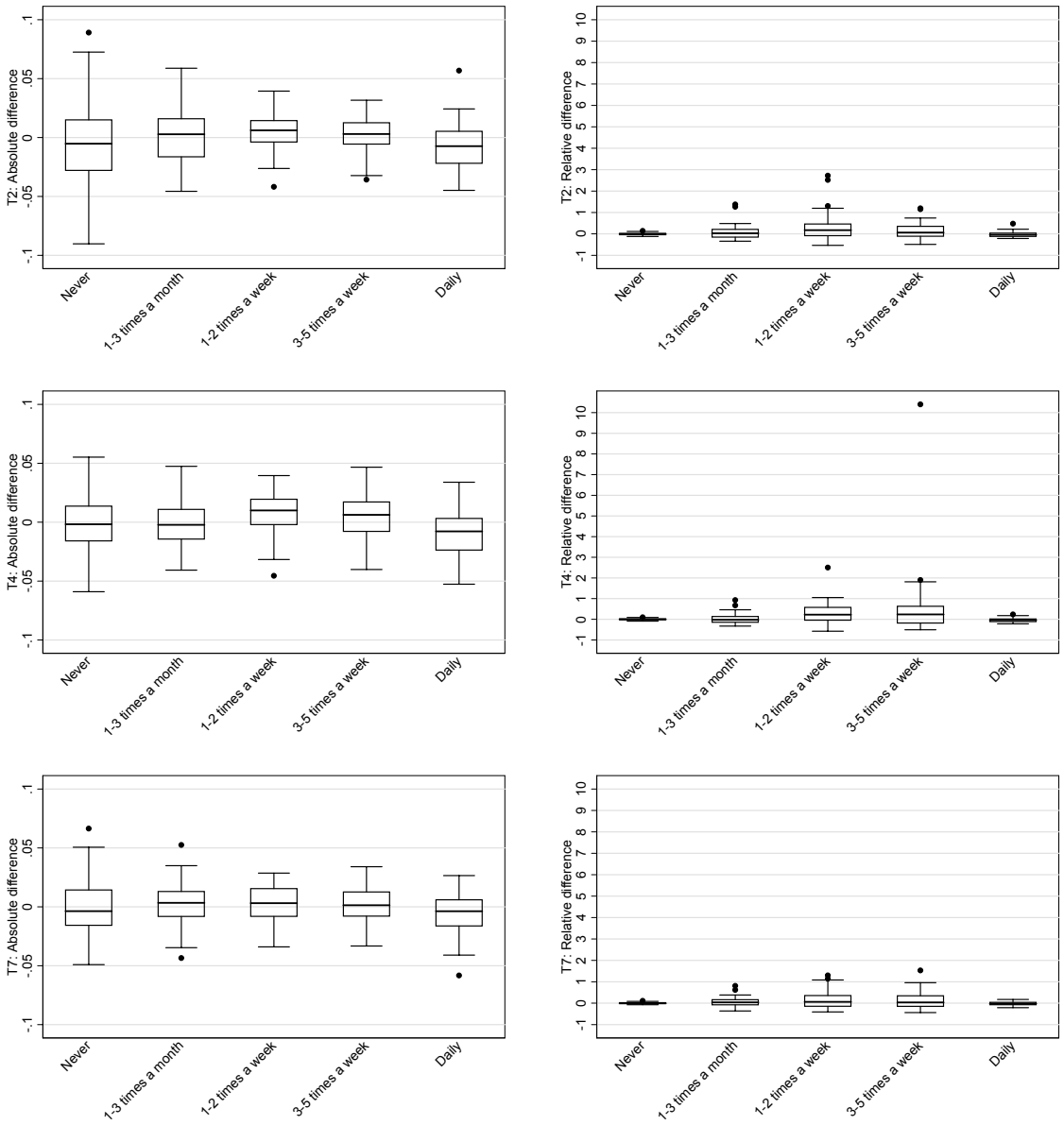

Figure 1: Simulated tobacco data with 10% of missing data on each wave. Boxplots for the difference between the proportions computed on the imputed and observed data. Only the 10% of imputed data points are taken into account. The left column shows the absolute difference between the proportions computed on the imputed and observed values for waves T2, T4 and T7, while the right column provides their relative difference.

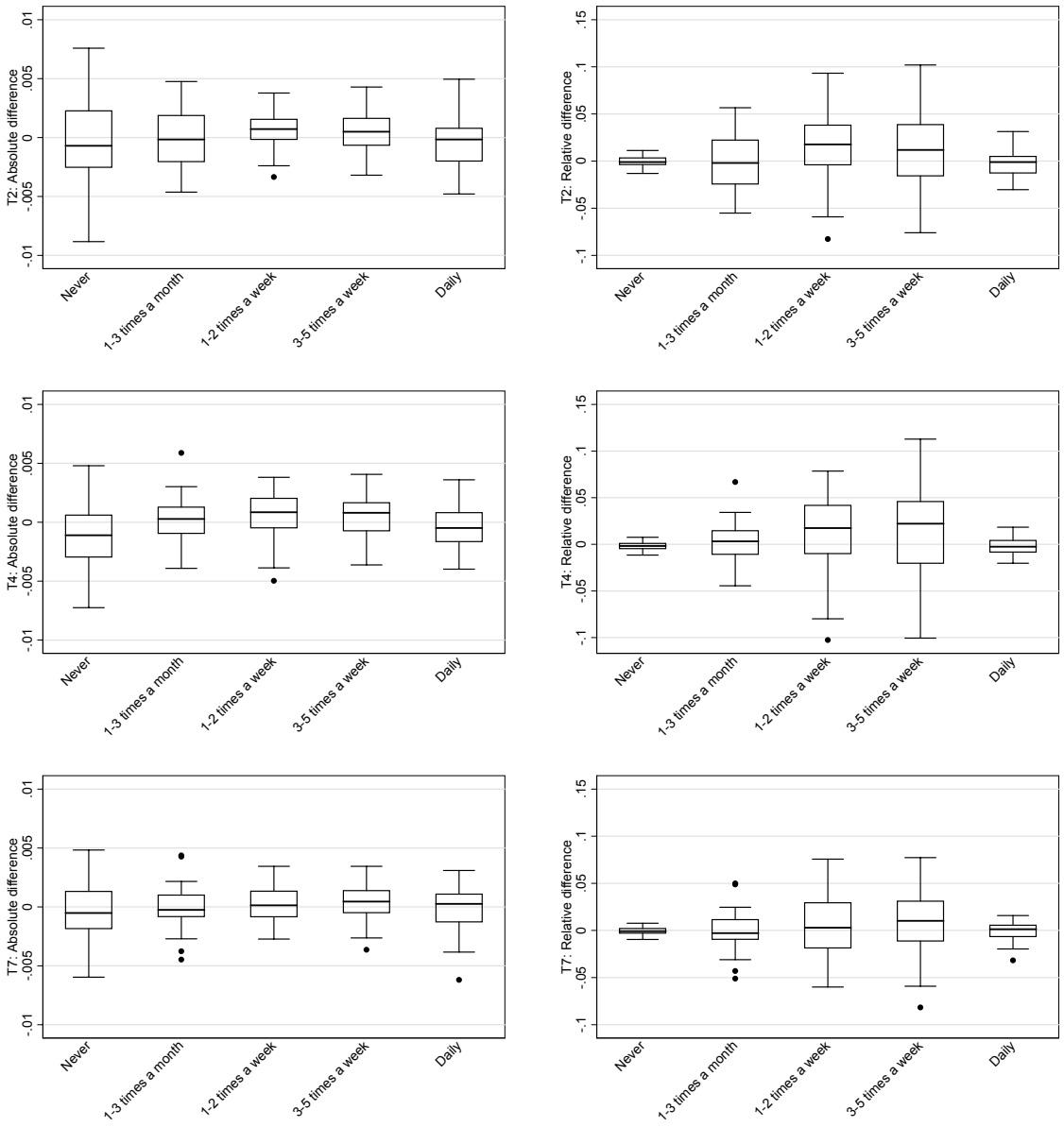

Figure 2: Simulated tobacco data with 10% of missing data on each wave. Boxplots for the difference between the proportions computed on the imputed and observed data. All the n=1999 data points, both imputed or not imputed, are taken into account. The left column shows the absolute difference between the proportions computed on the imputed and observed values for waves T2, T4 and T7, while the right column provides their relative difference.

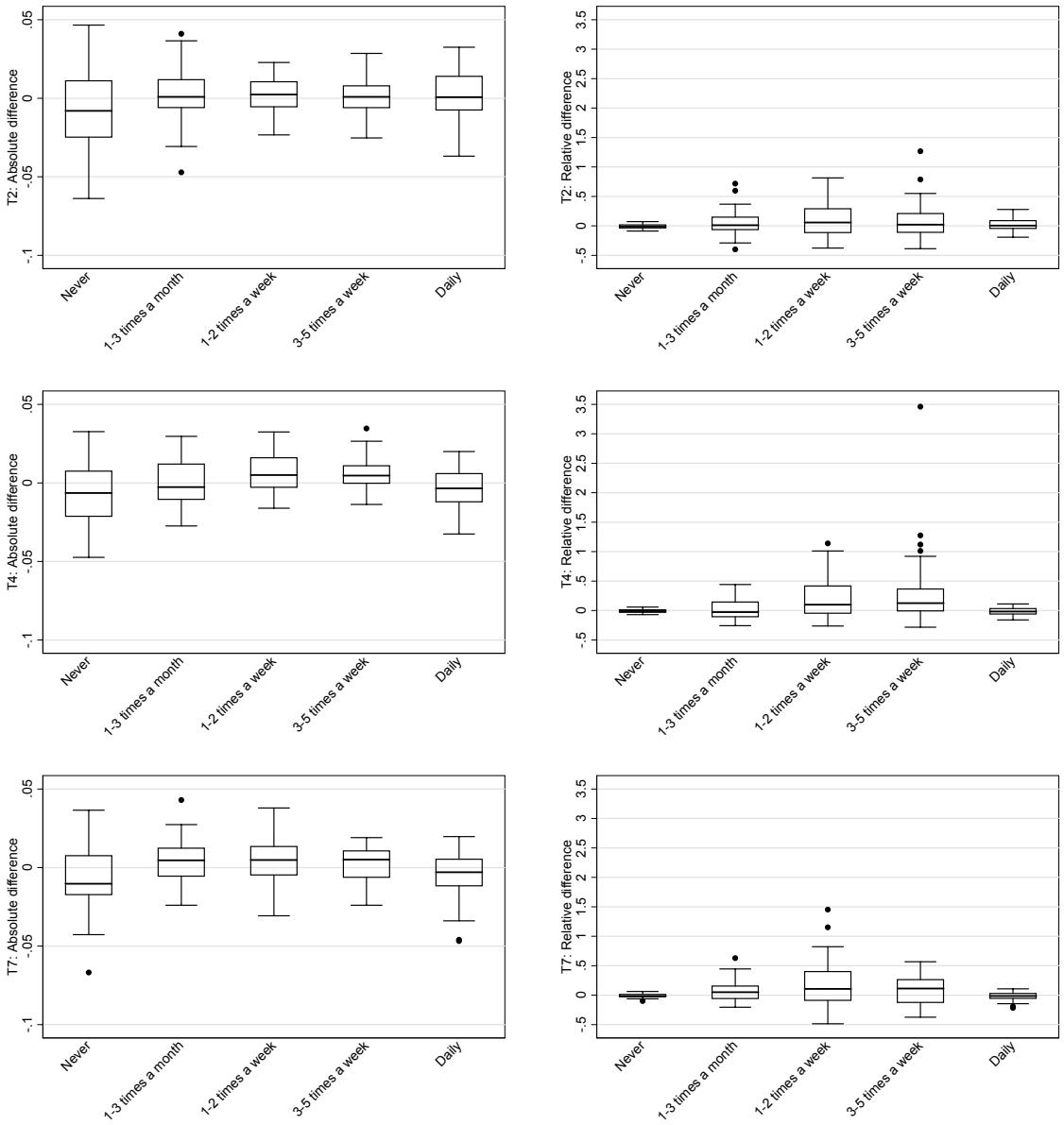

Figure 3: Simulated tobacco data with 20% of missing data on each wave. Boxplots for the difference between the proportions computed on the imputed and observed data. Only the 20% of imputed data points are taken into account. The left column shows the absolute difference between the proportions computed on the imputed and observed values for waves T2, T4 and T7, while the right column provides their relative difference.

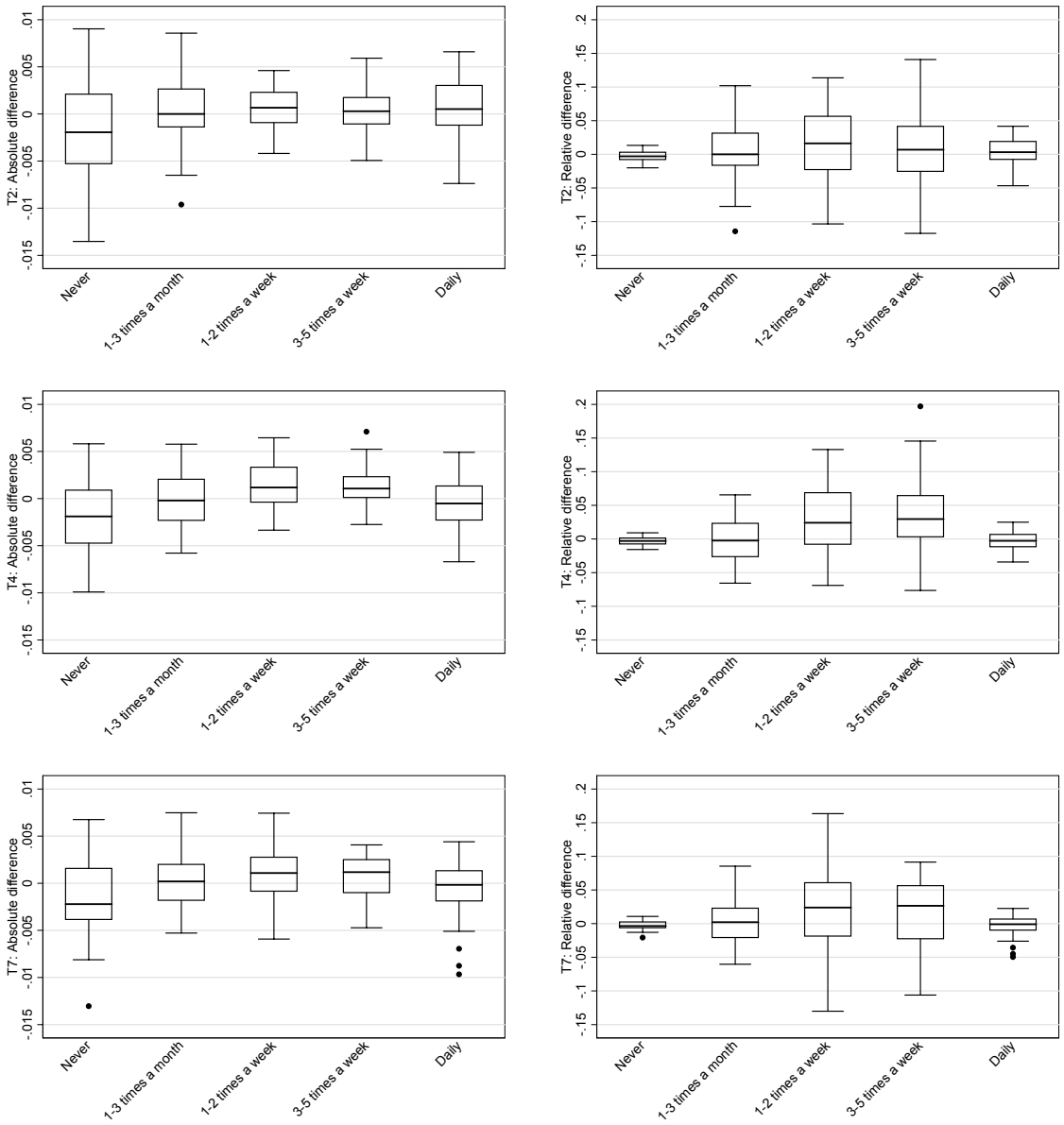

Figure 4: Simulated tobacco data with 20% of missing data on each wave. Boxplots for the difference between the proportions computed on the imputed and observed data. All the n=1999 data points, both imputed or not imputed, are taken into account. The left column shows the absolute difference between the proportions computed on the imputed and observed values for waves T2, T4 and T7, while the right column provides their relative difference.

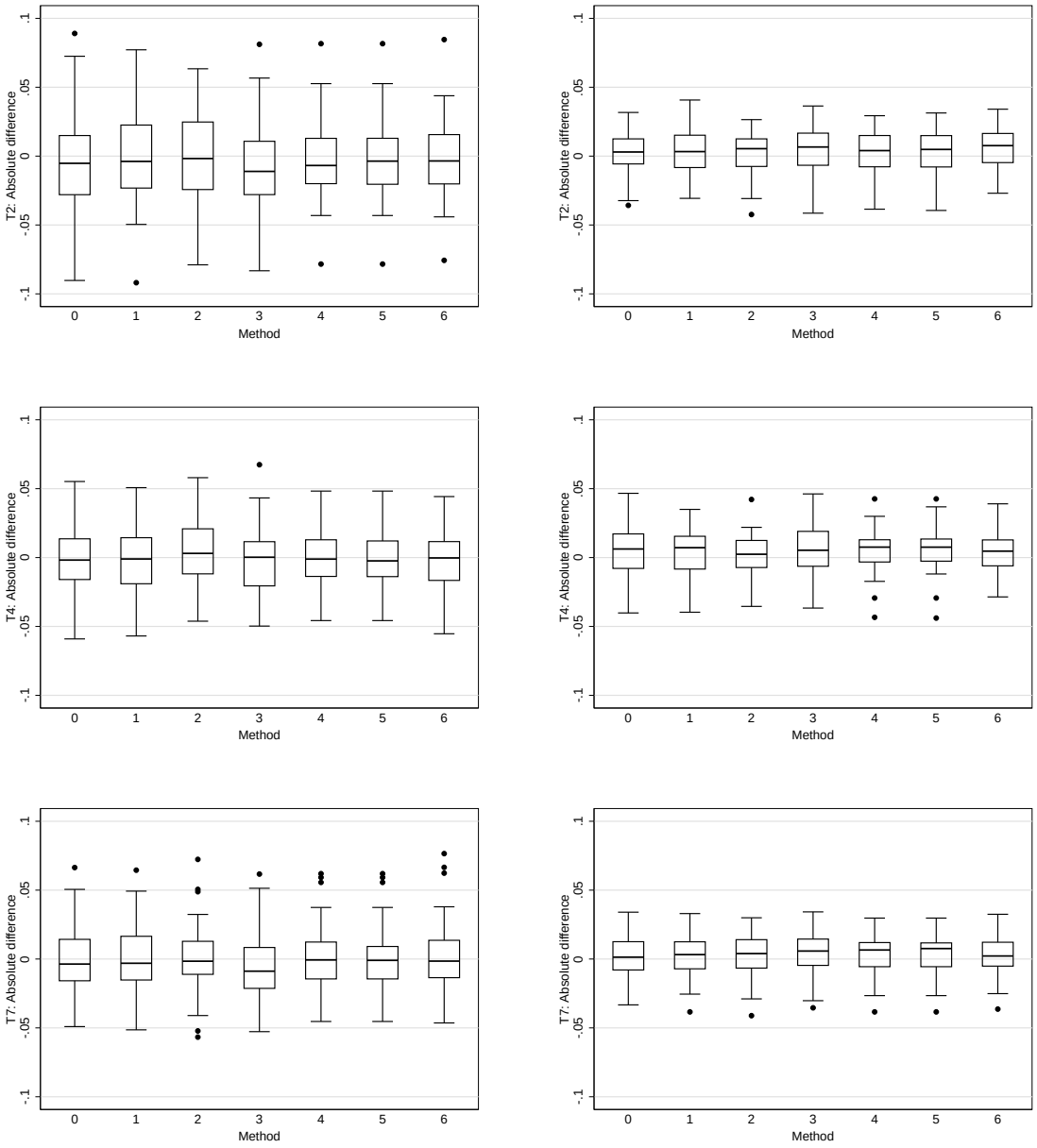

Figure 5: Simulated tobacco data with 10% of missing data on each wave. Boxplots for the absolute difference between the proportions computed on the imputed and observed data of the "Never" (left column) and "3-5 times/week" categories in waves T2, T4 and T7. Only the 10% of imputed data points are taken into account.

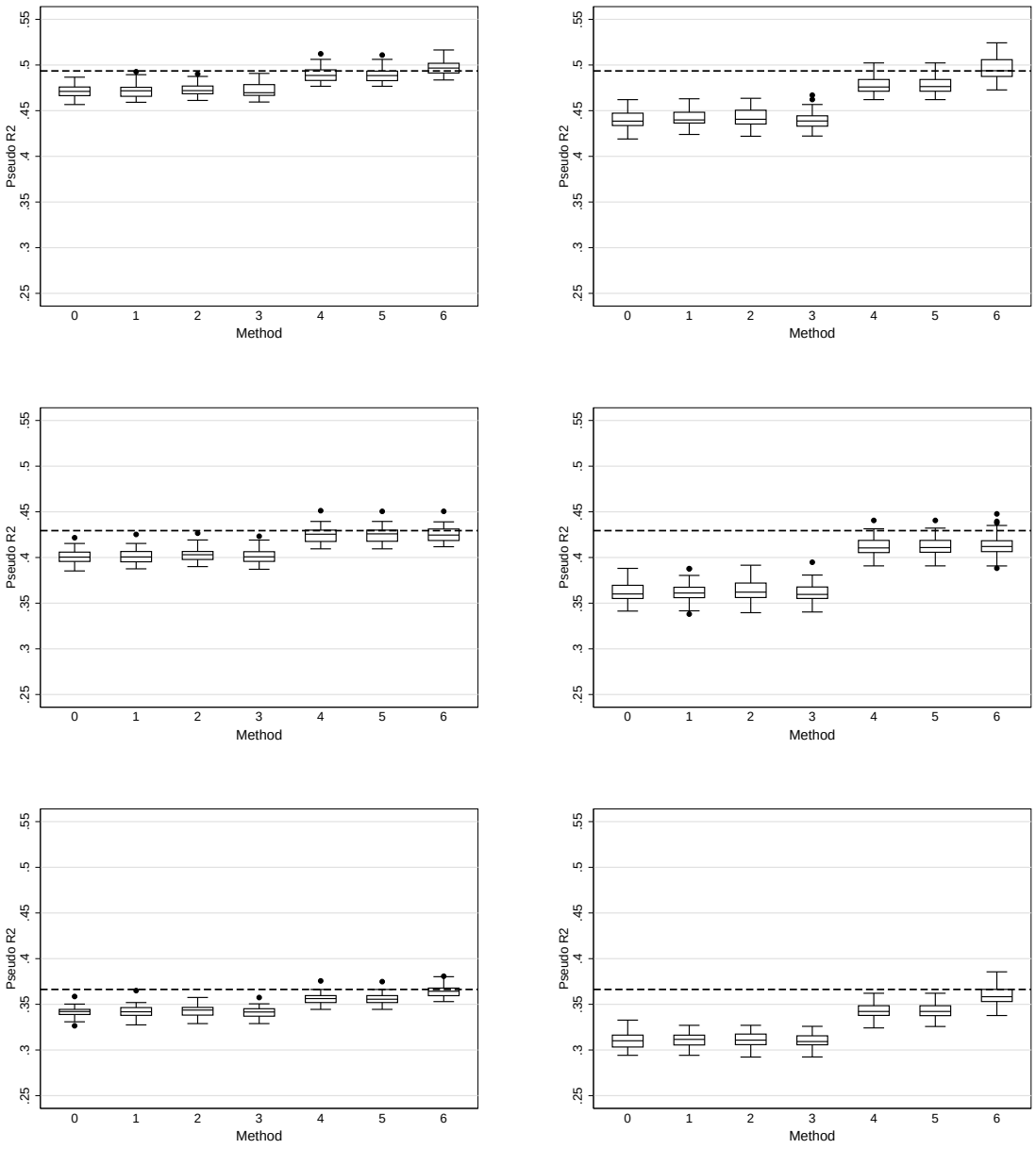

Figure 6: Nagelkerke's pseudo-R^2 for the regressions explaining the tobacco consumption level at T7 with data from T2 to T6 (up), T6 only (middle) and T5 only (down). Figures on the left report results with 10% of missing data and figures on the right report results with 20% of missing data. The dashed line represents the value of Nagelkerke's pseudo-R^2 when computed on the original dataset without missing data (0.4935 for the model using explanatory waves 2 to 6, 0.4294 for the model using only explanatory wave 6, and 0.3663 for the model using only explanatory wave 5).