Transfer of variables between different data-sets, or Taking ‘previous research’ seriously

Bojan Todosijević
Transfer of Variables between Different Data Sets, or Taking "Previous Research" Seriously

Bojan Todosijevic

Bulletin de Méthodologie Sociologique 2012 113: 20
DOI: 10.1177/0759106311426992

The online version of this article can be found at:
http://bms.sagepub.com/content/113/1/20

Published by:
Association Internationale de Methodologie Sociologique

RC33

and

http://www.sagepublications.com

Additional services and information for Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique can be found at:

Email Alerts: http://bms.sagepub.com/cgi/alerts

Subscriptions: http://bms.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://bms.sagepub.com/content/113/1/20.refs.html

>> Version of Record - Jan 5, 2012

What is This?
Transfer of Variables between Different Data Sets, or Taking “Previous Research” Seriously

Bojan Todosijević

COMPSOC, University of Twente

With or without missing data, the goal of a statistical procedure should be to make valid and efficient inferences about a population of interest – not to estimate, predict, or recover missing observations nor to obtain the same results that we would have seen with complete data. (Schafer and Graham, 2002: 149)

Résumé

Transfer de variables entre différentes bases de données, ou prendre au sérieux “des recherches précédentes” : Tenant compte de deux enquêtes méthodologiquement similaires, une question non posée dans une enquête pourrait être considérée comme un cas particulier du problème des données manquantes. Ainsi, le transfert des données entre des bases de données (« appariement statistique » ou « fusion de données ») pourrait être obtenu en appliquant les procédures bayésiennes d’imputation multiple des valeurs manquantes. Afin de s’attaquer au problème de l’indépendance conditionnelle créé par cette approche, un ensemble de données simulées pourrait servir comme « troisième ensemble de données » qui transmet l’information sur la relation entre les variables qui ne sont pas ordinairement observées. Ce document présente un modèle de transfert de données entre des bases de données différentes, basé sur l’approche de l’imputation multiple (IM). Les résultats montrent que l’appariement statistique fondé sur les principes MI peut être un outil de recherche utile. Ce travail est fondé sur une prise en compte sérieuse « des recherches précédentes ».

Corresponding Author:
Bojan Todosijević, 27 Nova 52, 24214 Tavankut, Serbia
Email: bojan.todosijevic@gmail.com
Abstract
Given two methodologically similar surveys, a question not asked in one survey could be seen as a special case of the missing data problem. Hence, the transfer of data across data sets (“statistical matching” or “data fusion”) could be achieved applying the procedures for Bayesian multiple imputation of missing values. To tackle the problem of conditional independence, which this approach creates, a simulated data set could serve as the “third data set” that conveys information about the relationship between variables not commonly observed. This paper presents a model for transferring data between different data sets based on multiple imputation (MI) approach. The results show that statistical matching based on MI principles can be a useful research tool. The entire enterprise is interpreted in the sense of taking the “previous research” into account seriously.

Mots clés
Appariement statistique, Imputation multiple, Etude électorale néerlandaise

Keywords
Statistical Matching, Multiple Imputation, Dutch Election Study

Introduction to the problem
It is an often-encountered problem that variables of interest are scattered throughout different data sets. For instance, a survey analyst might be interested in combining survey and census, or other official statistical data. In a series of comparable surveys it may happen that some questions of interest are not asked in all surveys. Further examples include when, within the same survey, some questions are not asked in all sample strata or clusters, or when random sections of the sample are asked different questions for experimental purposes (Gelman et al., 1998; Aluja-Banet and Thiò, 2001). Although collecting more data or abandoning the problem might seem as the obvious alternatives, they are not necessarily the only possibilities. The question is, thus, what would be an efficient way to use data available in different datasets?

A number of different ways have been proposed for dealing with the problem of transferring data between different sources, such as data aggregation, hot deck imputation, or record linkage approach (Rässler, 2002). Given two methodologically similar surveys, a question not asked in one survey could be seen as a special case of missing data problem (Gelman et al., 1998). Hence, the procedures for imputation of missing values could be applied not only when some respondents failed to provide valid answers, but also when some questions were not posed to all respondents (Rubin, 1987; Schafer, 1997; Schafer and Graham, 2002).

This paper presents a model for transferring data between different datasets based on multiple imputation (MI) approach. It builds on theoretical arguments developed by Rubin (1986), Gelman et al., (1998), and Rässler (2002, 2003). An early proposal of this approach could be found in Franklin (1989). A precedent, with theoretical justification and criticisms, can be found in Gelman et al. (1998), and further elaboration in Rässler.
(2002, 2003). Although data fusion seems to be an accepted practice in various areas of applied statistics, and there are already instances of its applications to substantive problems (He et al., 2007), it is still far from the mainstream statistical toolbox. The goal of this paper is to demonstrate the applicability and usefulness of this method to studying substantive problems in social sciences.

The present paper elaborates the theoretical arguments for the feasibility of the MI approach to statistical matching, and demonstrates its application to a substantive research problem. The paper makes two further points. First, a simple method for dealing with the problem of conditional independence, based on the use of simulated data, is proposed. Second, the general approach is interpreted in the sense of “taking the previous research seriously”, or, in other words, as a method of incorporating previous research directly into the research design.

The paper is structured as follows. The first part of the paper demonstrates the application of the MI procedure. After outlining its main characteristics, by comparing the imputed and actual respondents’ responses, follows the illustration of the simulated “third data source” approach to the conditional independence problem. The second part of the paper presents a substantive research example focused on the moderating influence of group identification onto the relationship between authoritarianism and ethnocentric attitudes. The exercises are based on the integrated analysis of two Dutch studies: Social and Cultural Developments in The Netherlands (SOCON 2000) and the Dutch Election Study (NKO 2002). Finally, the MI approach is discussed in the broader framework as a method for the advanced use of previous research.

**MI approach - Illustration and assessment**

Bayesian multiple imputation is a method for reaching valid inferences based on incomplete data (Rubin, 1987; Schafer, 1997). The key feature of the MI approach, which differentiates it from deterministic methods that impute single values, is that each missing response is replaced by several plausible values (multiply imputed values), which enables incorporating the uncertainty introduced by the imputed values. Multiply imputed data sets are then analyzed using the standard methods, and the estimates are integrated using the simple Rubin’s rules (1987).

One important condition for applying MI is that missing data are missing at random (MAR), or that nonresponses are ignorable (Rubin, 1987; Schafer and Graham, 2002). Missing data are MAR if probability of missingness does not depend on missing data:

\[
P(R|Y_{\text{complete}}) = P(R|Y_{\text{observed}})
\]

where \(R\) is distribution of missingness (Rubin, 1987; Schafer and Graham, 2002). Normally, when data are missing because a question was not included in a survey, the MAR assumption applies (Gelman et al. 1998).

The MI approach to data fusion is concerned with the situation when variables of interest (\(Y\) and \(Z\)) are not both observed within a given data set. If there is a set of variables that appear in both data files (variables \(X\)), they can provide ground for the imputation of the variable \(Y\) to the second data set. In the example that follows, I imputed a variable that actually exist in the second data (variable “\(Y \text{ true}\)”), in order to compare
the imputed with the actual respondents’ scores. It should be also noted that variable Y is imputed \( m \) times, and therefore \( m \) versions of the second data files are created and subsequently analyzed.

To assess the feasibility of the approach, we proceeded with the following three steps:

1. Two data sets were selected: SOCON 2000 (Social and Cultural Developments in the Netherlands\(^1\)), and the Dutch Parliamentary Election Study (NKO 2002\(^2\)). Both surveys are based on random samples of Dutch population. They were conducted close in time, and in methodologically similar manner (face to face interviews). What is particularly important, they contain a number of equivalent variables that could serve to construct the imputation model. Both data sets are freely accessible through the DANS archive.\(^3\)

2. Left-Right self-placement is chosen as the variable to be imputed; that is, transferred from the SOCON to the NKO data file. Both data sets contain this variable (in fact, the NKO data contain three L-R variables, collected at each of the three survey waves). The reason for imputing the variable that exists in the second data set is to be able to compare the imputed variable with the actual answers given by the respondents.

3. Finally, various test and analyses were performed to assess the viability of the MI approach to solving the data-transfer problem. This included, for instance, examining the degree of association between the imputed L-R variable and the actual respondents’ scores. The major test is the comparison of the associations of the transferred variable and the original NKO variables with a number of relevant variables existing in the second data set but not used in the imputation model. This exercise is intended to provide an insight into whether one would reach similar conclusions if one used the original or imputed target variable.

**Imputation procedure and software**

Among the variety of recently developed imputation techniques, the one based on chained equations received significant theoretical and empirical attention. This approach was developed by Van Buuren and Oudshoorn (2000) and termed MICE (from Multivariate Imputation by Chained Equations). The imputations are drawn from the multivariate distribution, so that all available information from other variables is taken into account. The multivariate distribution is estimated from the incomplete data in a Gibbs sampling process (Van Buuren and Oudshoorn, 1999).

In addition to the strong theoretical ground, this approach to MI fared favorably in a number of comparative assessments of imputation procedures (Acock, 2005; Horton and Lipsitz, 2001; He et al., 2007; Schafer 2003), and emerged as the state of the art approach to MI (Shafer and Graham, 2002). According to Rässler (2003) MICE performed especially well in exploiting the information from auxiliary data sets, which is particularly relevant for the present purpose. In the present study, I used the Ice program, which is a Stata module implementing MICE procedure, written by Patrick Royston (2005).\(^4\)

The following imputation procedure is designed so that the second data file is “enriched” with the imputed variable. This is different from the standard MI procedure,
which would do the imputation symmetrically, and analyze the merged dataset. The illustrative imputation procedure consists of three main steps.

1. First, a set of variables that appear in both SOCON and NKO studies is selected and adjusted. Twelve such variables are selected, and include mostly the usual background variables, but also several attitudinal measures (post-materialism index, political interest). These variables serve to predict/impute the target variable. The more variables appear in both data sets the better; since, in principle, it should be possible to construct a better predictive equation.

2. Before imputing the target variable, missing values within the two data sets were also imputed, because the missing predictors would cause missing predicted values. The reason for separate initial imputation of the SOCON and NKO variables is that the joint imputation could artificially increase the similarity between the two data sets. In this step, therefore, two independent imputation procedures were performed: a) Imputation of the common variables in the SOCON file (using \textit{ice}); b) Imputation of the common variables in the NKO file (using \textit{ice}). Note that Left-right self-placement variables from the NKO 2002 study (l_r1, l_r2, and l_r3) were not included in the imputation model. Imputation of the L-R variables using the same predictive variables as in the SOCON data file could increase correlation between the NKO L-R variables and the imputed L-R variable.

3. Imputation of the L-R variable – from SOCON to NKO, using \textit{uvis}. The imputed SOCON and NKO files, each containing the original and 5 imputed data sets, were merged together into a combined SOCON-NKO data-file. Then, for each of the 5 imputed SOCON-NKO combinations, a univariate imputation of the L-R variable (from SOCON to NKO) was performed (using \textit{uvis}).

The concept of MI implies construction of a number of new data sets with different versions of the imputed variable. Intended analyses are then performed using each data set, and integrated in the final step using procedures specified in Rubin (1987).

The imputation equation in the SOCON file deserves a bit more attention, since these relationships determine the outcome of the imputation procedure. The common variables explain 40 percent of variance in the L-R scores. Variables with the strongest influence are party preference, education, social class self-assessment, church attendance, and postmaterialism. It is comforting, though not necessarily required for the purpose of imputation, that all the relationships make theoretical sense.

In the preliminary analyses, various interactions were also examined, but none proved significant enough to require modification of the simple linear model. The examination of possible interactions is important since the imputation procedure transfers only those relationships specified by the model. If there are relevant interactions which are not specified by the model they will be lost in transfer.

**Assessing the imputed variable – Comparison with “true” scores**

As shown in Table 1, the basic distributional features of the imputed L-R variable are well preserved. Not only that the imputed variable resembles the source variable (in this
case SOCON L-R), the cross-imputed variable closely resembles the actual responses of the NKO 2002 respondents, especially if we take into account the variations between the three NKO waves. The last row in the table shows the mean and SD pooled over 50 imputations. Note that the SD estimate is more than doubled – a reflection of the added uncertainty associated with the imputed scores.

Figure 1 shows the average distance between the imputed L-R scores and the actual responses. The average distance between the imputed and actual responses in the NKO data set is around 1.80. The distance between respondents’ responses in NKO waves 1 and 2, and 2 and 3 is .97, while between the waves 2 and 3 it is somewhat larger – 1.14, as could be expected given the longer period between the surveys. Thus, the distance between the imputed and actual scores is somewhat larger than between the responses recorded at the three waves of the survey. Yet, the similarity between the imputed and actual responses is quite impressive. For approximately 55 percent of respondents, the scores are either identical or differ only a single point on the 11-point L-R scale.

The next question concerns the correlation between the imputed and “true” scores. The correlation between the actual L-R scores at three waves of the NKO study can serve as a standard for evaluating the comparison. These coefficients range from $r = .71$ (waves 1 and 3) to $r = .78$ (waves 2 and 3). The time difference between the waves is relatively small, so the imputation based on a survey conducted two years earlier should result in somewhat lower coefficients. Table 2 shows the correlations between the “true” and imputed variables (the original NKO and cross-imputed L-R). The correlation coefficients are statistically significant and in the expected direction in all cases, but also substantially lower (around .40) than those observed among the three NKO waves. Thus, the imputed variables are close to the respondent’s actual answers, but the procedure introduces a substantial degree of uncertainty as well.

The next step is to compare the imputed variable (multiple imputations combined following Rubin’s rules) with the true scores, using the imputation equation. In other words,
the actual and imputed L-R scores are regressed on the aforementioned set of common predictor variables. There are two important observations to be made by inspecting Table 3. One is that the coefficients associated with the “true” variables sometimes considerably differ across the waves, thus making it difficult to reach a more general conclusion about the strength of the association. The second is that significant coefficients obtained using the cross-imputed equations are mostly within the range of those obtained using the scores from the three NKO waves. One of the minor exceptions concerns the education variable, where the coefficient is higher for the cross-imputed variable than for the actual scores, although they are all in the same direction and statistically significant.

To summarize, the regression equations using the same specification as in the imputation model, but using the original NKO L-R variables as the DV reveal quite similar relationships in the two data sets. In fact, the differences between the three NKO waves seem to be larger than between the SOCON and NKO surveys taken together. This step is

![Figure 1. Distance between the imputed and actual L-R scores](image)

**Table 2. Correlation between the imputed LR scores and the actual responses**

<table>
<thead>
<tr>
<th>Actual scores</th>
<th>L-R imp. 1</th>
<th>L-R imp. 2</th>
<th>L-R imp. 3</th>
<th>L-R imp. 4</th>
<th>L-R imp. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-R 1st wave</td>
<td>.33</td>
<td>.39</td>
<td>.35</td>
<td>.38</td>
<td>.38</td>
</tr>
<tr>
<td>L-R 2nd wave</td>
<td>.40</td>
<td>.42</td>
<td>.39</td>
<td>.45</td>
<td>.42</td>
</tr>
<tr>
<td>L-R 3rd wave</td>
<td>.40</td>
<td>.42</td>
<td>.40</td>
<td>.45</td>
<td>.42</td>
</tr>
</tbody>
</table>

Note: All coefficients are statistically significant, \( p < .001 \).
a useful check to verify that the relationships that provided the basis for the imputation (based on SOCON data) are similar to those in the new data set.

### Relationships with variables not included in the imputation model

The most relevant test of the method for data transfer is to compare the relationships between the actual and imputed L-R scores, on one hand, with a set of variables from the receiver (NKO) file that were not included in the imputation model, on the other. One way to look at this problem is to calculate correlation coefficients between the NKO variables, and the imputed and “true” L-R scores.

For this purpose, more than 60 attitudinal variables that differ according to their expected degree of association with L-R self-placement are selected. They are rather typical representatives of variables often found in survey analyses of political behavior. Examples include satisfaction with democracy and government, sympathy scores for political parties, measures of political efficacy, political cynicism, trust, and so on.

The evidence to look for is whether one would reach different conclusions about the relationships between the L-R variable and the included attitudinal variables if one would use the imputed L-R variable rather than the actual responses. Table 4 shows a portion of the examined associations.

The coefficients associated with the imputed variable are lower than those for the actual scores. For instance, correlation between the sympathy scores for the GroenLinks party and “true” L-R scores is between −.47 and −.49. The imputed L-R scores are associated with the same variable to a smaller degree ($r = -.31$). However, the coefficients for the imputed variable are without exception in the same direction, and very

---

**Table 3. Regression of the imputed and actual scores on the common set of predictor variables**

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Coef.</th>
<th>t</th>
<th>Coef.</th>
<th>t</th>
<th>Coef.</th>
<th>t</th>
<th>Coef.</th>
<th>t</th>
<th>Coef.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urbanization</td>
<td>.075</td>
<td>2.95</td>
<td>.085</td>
<td>2.77</td>
<td>.086</td>
<td>2.59</td>
<td>.037</td>
<td>1.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>−.185</td>
<td>−1.47</td>
<td>−.340</td>
<td>−4.26</td>
<td>−.180</td>
<td>−2.06</td>
<td>−.099</td>
<td>−1.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.001</td>
<td>.26</td>
<td>−.008</td>
<td>−2.58</td>
<td>−.007</td>
<td>−2.06</td>
<td>−.004</td>
<td>−1.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class</td>
<td>.385</td>
<td>5.36</td>
<td>.118</td>
<td>1.86</td>
<td>.109</td>
<td>1.55</td>
<td>.163</td>
<td>2.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income</td>
<td>.002</td>
<td>.05</td>
<td>.058</td>
<td>1.34</td>
<td>.011</td>
<td>.22</td>
<td>.013</td>
<td>.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>−.145</td>
<td>−6.50</td>
<td>−.098</td>
<td>−4.62</td>
<td>−.077</td>
<td>−3.32</td>
<td>−.046</td>
<td>−2.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Church attendance</td>
<td>−.217</td>
<td>−3.34</td>
<td>−.227</td>
<td>−6.06</td>
<td>−.339</td>
<td>−8.35</td>
<td>−.265</td>
<td>−8.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Party</td>
<td>.391</td>
<td>15.77</td>
<td>.348</td>
<td>22.80</td>
<td>.423</td>
<td>25.33</td>
<td>.460</td>
<td>34.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employed</td>
<td>.012</td>
<td>.12</td>
<td>.175</td>
<td>1.82</td>
<td>.138</td>
<td>1.31</td>
<td>.035</td>
<td>.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-materialism</td>
<td>−.359</td>
<td>−4.55</td>
<td>−.186</td>
<td>−2.56</td>
<td>−.313</td>
<td>−3.92</td>
<td>−.321</td>
<td>−5.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political interest</td>
<td>.197</td>
<td>1.94</td>
<td>.136</td>
<td>1.83</td>
<td>.187</td>
<td>2.31</td>
<td>.098</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proud to be Dutch</td>
<td>−.244</td>
<td>−4.56</td>
<td>−.048</td>
<td>−1.04</td>
<td>−.072</td>
<td>−1.42</td>
<td>−.202</td>
<td>−5.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>4.61</td>
<td>8.55</td>
<td>4.76</td>
<td>11.80</td>
<td>4.83</td>
<td>10.99</td>
<td>4.70</td>
<td>13.50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$R^2$ | .39 to .43 | .31 | .40 | .41
close in terms of statistical significance. The direction and significance of the association is often correctly estimated even for the weak associations, such as concerning the satisfaction with democracy. When the imputed scores suggest an insignificant association, it is either insignificant also according to the actual answers, or unclear, as with the attitude towards the necessity of political parties.

More specifically, identical conclusions, in terms of direction and statistical significance, would be reached in 56 out of 63 examined associations, regardless of whether one used the imputed or true L-R scores (that is in 88.9% of the cases). In most cases (51), this meant concluding that the association is significant, but the insignificant associations are also generally correctly estimated (in 5 remaining cases). Different conclusions would be reached in 11.7 percent of cases (7 cases). Most of them (6 out of 7) concern concluding the insignificant association using the imputed scores, when the relationship based on the true scores seems to be significant.

In only a single instance the imputed variable would lead to concluding a significant association when the actual scores indicate an insignificant relationship (item iii/1293). Even in that case, one of the coefficients for the actual scores is higher, but is below statistical significance because of the smaller number of cases.

Thus, the typical “error” one would make when using the imputed scores would be to conclude that a relationship is insignificant when it is not. However, this tendency is shown only for the relationships that are rather weak. The highest “missed” correlation was $r = -.11$ (with political knowledge; the coefficient represents average for the three NKO L-R variables). To summarize, the coefficients associated with the imputed variables are lower in magnitude, but generally correctly estimate the direction and

### Table 4. Correlations between the L-R scores (imputed and actual) and selected attitudinal variables in the receiver (NKO) data-set

<table>
<thead>
<tr>
<th>Variables</th>
<th>Imputed L-R</th>
<th>“True” L-R scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wave 1</td>
</tr>
<tr>
<td>I/141 General satisfaction with government.</td>
<td>.10*</td>
<td>.17*</td>
</tr>
<tr>
<td>I/ Policy satisfaction score 2002</td>
<td>-.08*</td>
<td>-.12*</td>
</tr>
<tr>
<td>I/142 Satisfaction with democracy</td>
<td>.06*</td>
<td>.10*</td>
</tr>
<tr>
<td>I/281 Sympathy score: CDA</td>
<td>.17*</td>
<td>.33*</td>
</tr>
<tr>
<td>I/281 Sympathy score: PvdA</td>
<td>-.24*</td>
<td>-.36*</td>
</tr>
<tr>
<td>I/281 Sympathy score: VVD</td>
<td>.21*</td>
<td>.39*</td>
</tr>
<tr>
<td>I/281 Sympathy score: D66</td>
<td>-.23*</td>
<td>-.26*</td>
</tr>
<tr>
<td>I/281 Sympathy score: GroenLinks</td>
<td>-.31*</td>
<td>-.47*</td>
</tr>
<tr>
<td>I/281 Sympathy score: Leefbaar Nederland</td>
<td>.11*</td>
<td>.23*</td>
</tr>
<tr>
<td>I/281 Sympathy score: Lijst Pim Fortuyn</td>
<td>.21*</td>
<td>.40*</td>
</tr>
<tr>
<td>II/355 Politics is too complicated</td>
<td>-.06*</td>
<td>-.08*</td>
</tr>
<tr>
<td>III/1295 Parties necessary for functioning of democracy</td>
<td>.01</td>
<td>.07*</td>
</tr>
<tr>
<td>iii/1293 Views of MP’s are good reflection of voters</td>
<td>.04*</td>
<td>.04</td>
</tr>
</tbody>
</table>

Note that the table shows coefficients based on a single imputation. * $p < .05$. 

Downloaded from bms.sagepub.com at Central European University on October 26, 2012
significance of the association. Using the imputed variable one is in danger of making Type II error, but much less of making Type I error.

The findings just presented are based on a single imputation, so the coefficients with specific variable may differ between different imputations. The differences are very small, and do not change the percentages reported in the table. Yet, more in the spirit of the MI approach is to calculate summary measures based on multiple imputations (following Rubin’s rules). The following exercise does that by calculating bivariate regressions over multiple imputations (in this case 100 imputations), where independent variables are the 65 attitudinal variables, and the dependent variable is the L-R variable (imputed and true scores). Thus, both the point estimate and the standard errors are adjusted for the additional variability introduced by the multiple imputations.

In this case, because pooling over multiple imputations increases standard errors, the probability of wrongly assuming that the relationship is insignificant is greater than in the exercise with correlation coefficients. The following scatter graphs show the difference in the outcomes when using the imputed variable versus the actual scores.

Figure 2 shows the relationship between the $b$ coefficients for the imputed and true left-right scores (vertical axis) and the average $b$ for the “true” scores (horizontal axis). We can clearly observe the two expected tendencies. One is that the coefficients are underestimated if one uses the imputed variable. The other is that $b$’s for the imputed variable are in close linear relationship with the coefficients for the true scores. In fact, the correlation coefficients between the $b$ coefficients for the imputed L-R and the true original L-R variables are above .97.

The reduction in the size of the $b$ coefficient for the imputed variable is shown by the regression coefficient when $b$ coefficients for the imputed variable are regressed on the coefficients for the “true” L-R scores. In all three cases, this second-level $b$-coefficient is around .60 (.62, .58, .57 for waves 1, 2 and 3, respectively), showing the drop in the coefficients estimated on the basis of the imputed variables. This drop reflects two related underlying processes. One is the problem of conditional independence – the applied imputation procedure implies conditional independence of the imputed variable from all variables not included in the imputation model, given the variables included in the imputation model. The second process concerns the unreliability of the imputed scores. This is a direct reflection of the size of R-squared obtained applying the imputation equation.

In broad terms, the deviation of the imputed variable’s slope from 45° angle in Graph 1 reflects the unreliability of the estimates, while the deviations of the individual estimates from the slope line predominantly reflects the problem of conditional independence. The latter deviations reflect conditional independence concerning omitted variables relevant for the specific variable in question. To the extent that there are variables excluded from the imputation model that influence the relationship between the target variable and all the other variables included in this exercise, the slope also reflects their exclusion.

We can also observe that in some cases differences between the coefficients for three waves of the 2002 survey are not much smaller than the difference between them and the imputed variable, which is particularly visible in the third quadrant. This observation can be taken as a reminder of the relativity of what is understood as the “true” relationship.
Figure 3 illustrates another tendency that is in-built in the adopted imputation procedure. Namely, according to Rubin’s rule, standard errors for multiply imputed variables reflect variation within imputations (average SE over multiple imputations) and variations between imputations. Therefore, the overall estimated standard error based on multiply imputed variables is higher than for the actual L-R scores. This is reflected by the second-stage regression coefficients, with SE’s for the imputed variable as the dependent variables, and SE’s for the original scores as dependent variables, which are 1.27, 1.21, and 1.10 for waves 1, 2 and 3 respectively. Thus, SE for the imputed variable is systematically increased between 10 and 20 percentages. What is equally important is that the two sets of SE estimates are again very closely related (r’s are above .98).

The problem of conditional independence

The applied imputation procedure implies conditional independence of the imputed variable from variables not included in the imputation model, given the variables included in the imputation model. This means that if any variable is added to the regression equation
predicting the imputed variable, in addition to all the variables included in the imputation equation, the estimated relationship will be insignificant. Rässler (2003: 60) says “The association of the variables never jointly observed is unidentifiable and cannot be estimated by means of likelihood inference.”

To ameliorate the effects of the implied conditional independence, Rässler (2003) discusses two approaches. One is to use additional, perhaps of smaller magnitude, auxiliary data set that actually includes variables not jointly observed in the main data sets. The additional data set could, therefore, provide for the estimated conditional relationship. Another method is “based on informative prior distributions in the Bayesian context” (Rässler, 2003: 60). Prior information can be used to estimate the range of possible outcomes.

Here, I present a method that combines the two approaches and is based on data simulation. The auxiliary data set could be created by simulating the variables not jointly observed. Various hypotheses about the consequences of different degrees of conditional relationships could be examined by generating a series of such “third” samples. This approach is easy to apply, and is not dependent on a particular software’s ability to incorporate the prior information.

Figure 3. Standard errors associated with bivariate regression coefficients (b’s) for the imputed and “true” L-R scores
If a researcher is interested in the relationship between variables $Y$ and $Z$, but they are not observed together, $Y$ can be imputed into the B data set using the set of common variables $X$. Normally, the imputed $Y$ would be independent of all other variables, including $Z$, conditioned upon the common variables $X$. However, it is possible to add a simulated “third” data set, containing the $Y$ and $Z$ variables, and in that way manipulate the degree of association between $Z$ and $Y$ in the resulting imputed data set. Thus, this can be seen as an alternative way to deal with the conditional independence problem.

The following example illustrates this approach. The aim is to estimate the relationship between L-R self-placement and sympathy for political parties (in this case CDA and VVD). When the two sympathy scores are entered into the regression equation in the first data set (SOCON), both variables appear significantly related with the L-R scores (see the top panel in Table 5). However, when the L-R variable is transferred to the NKO dataset via the MI procedure, the sympathy scores appear unrelated to the dependent variable, as a consequence of the conditional independence since these variables are not part of the imputation model (second panel, Table 5).

To compensate for the conditional independence, a “third” data set ($N = 100$) is simulated. The intercorrelations between the three variables are simulated so to be close to those actually observed in the donor data-file. The simulated dataset is then appended to the SOCON data-file, and the imputation procedure is performed with the additional 100 simulated cases. The result is that the conditional relationships are now preserved – despite the fact that the target variables did not exist in the data that served to construct the imputation model (see the third panel in Table 5).

### Table 5. Conditional relationships between L-R self-placement and party sympathy scores

<table>
<thead>
<tr>
<th></th>
<th>SOCON data (donor file)</th>
<th>NKO data (receiver file)</th>
<th>NKO data (receiver file)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b$</td>
<td>SE</td>
<td>$t$</td>
</tr>
<tr>
<td>Sympathy for CDA</td>
<td>.014</td>
<td>.002</td>
<td>6.32</td>
</tr>
<tr>
<td>Sympathy for VVD</td>
<td>.019</td>
<td>.002</td>
<td>8.63</td>
</tr>
</tbody>
</table>

Note: Complete equations include the common variables in SOCON and NKO data sets. Coefficients for these variables are not shown.
Using additional information about conditional relationships may be a worthwhile effort. In addition to the methods proposed by Rässler (2003), the method of a simulated data set provides a simple way to examine consequences of different assumptions about various patterns and degrees of association between the variables of interest.

**Substantive research Illustration - Party preference, group identification and personality**

For analysts of Dutch politics, the success of Pym Fortuyn remains an interesting research topic. Given the rhetoric of this and parties following its suit, one may ask a number of questions about the sources of ethnocentric attitudes in politics.

From the individual-differences perspective, it may be argued that individual dispositions, such as personality characteristics, explain ethnocentric preferences and attitudes. The best-known approach in this tradition is the one based on the authoritarian personality tradition (Adorno et al., 1950; Meloen, 1993). On the other side, according to the inter-group relations perspective, ethnocentrism is a reflection of the relationships between groups. Social identity theory (SIT), for instance (Tajfel, 1981; Tajfel and Turner, 1986), would argue that ethnocentrism depends on the strength of in-group identification.

Finally, according to the integrated model, which would be close to recent attempts to integrate SIT and individual disposition approaches (Duckitt, 2001), it could be expected that authoritarianism is stronger predictor under condition of weaker group identification. Under condition of strong in-group identification, ethnocentrism should be high regardless of one’s personality disposition, but in case of weak identification, ethnocentrism should reflect one’s personality to a greater extent.

It would be interesting to examine the implication of this hypothesis in connection with the attitudes of Dutch voters in 2002, at the time of the LPF electoral breakthrough. The problem is that data needed to perform such analysis are not available. To be more precise, the data are not available within a single data set.

However, the relevant data can be found scattered across two data sets: the Dutch Parliamentary Election Study from 2002 (NKO 2002), and the SOCON 2000 study. The NKO study contains various measures of political behavior, including a measure of ethnocentric orientation, and was conducted at the time of the LPF greatest electoral success. However, it does not contain social-psychological variables, so the hypothesis about the interaction between individual disposition and group identification cannot be examined. The SOCON study, however, includes an authoritarianism scale, in addition to various variables that are also available in the NKO study. The Authoritarianism scale (see Table 6) consists of 6 items, and individual scores are created by calculating the mean score on the 6 items. Sample item: *What we need are less laws and less institutions, and more courageous, indefatigable, and devoted leaders, in whom the people can put their faith.* Reliability of the summarized scale is $\alpha = .79$.

To address the outlined research problem, it is possible to impute the authoritarianism scores into the NKO data set, using a set of mostly socio-demographic variables available in both data sets (the same set of variables as in the earlier examples, plus the L-R scores). The R-square for the imputation equation was $R^2=.37$. Again, the predictive
equation is quite good in predicting the target variable, though still far from perfect. The same imputation procedure is applied as in the earlier example.\textsuperscript{6}

Once the authoritarianism scores are cross-imputed from SOCON to NKO, it becomes possible to examine the interaction between group identification and authoritarianism. The association between authoritarianism (imputed from SOCON to NKO) and ethnocentrism is stronger in the low identification group ($b = .53$, s.e. = .15), than in the high-identification group ($b = .27$, s.e. = .08), thus confirming the general hypothesis.\textsuperscript{7} If an interaction term (authoritarianism and group identification) is added to the equation, it shows that the interaction is statistically significant. Moreover, although both authoritarianism and group identification are positively related with ethnocentrism, their interaction is significant and goes in negative direction.

The hypothesis about the effect of the interaction between dispositions and group identification could be further elaborated. Namely, authoritarianism should predict party preference for ethnocentric parties depending on the degree of in-group identification. The association should be weaker in the case of strong in-group identification. The hypothesis refers primarily to preferences for parties with ethnocentric orientation (in this case LPF).

However, it may also apply to parties whose primary emphasis is not on policies relevant for intergroup relationships, but whose broader ideological orientation has relevant implications. In the Dutch case, it seems reasonable to expect that parties on the left will exhibit the opposite relationships than the parties from the right. Figure 4 displays regression coefficients indicating the relationship between authoritarianism and party preference, under different degrees of in-group identification, for the major Dutch parties.

The obtained findings clearly support the hypothesis: under condition of low in-group identification, authoritarianism is stronger predictor of preference for LPF than among the strong identifiers. In both waves, the absolute difference between the coefficients is the highest for LPF.

In addition, it is also clear that the interaction between dispositions and party preferences in this case goes beyond a single party – regression coefficients are generally higher in the group of low identifiers, although the effects go in the opposite direction.

\begin{table}[h]
\centering
\caption{Authoritarianism scale items, data set SOCON 2000 (P 1556)}
\begin{tabular}{ll}
\hline
SOCON code name & Item \\
\hline
v0623 & There are two sorts of people: the strong and the weak. \\
v0626 & Our social problems would be largely solved, if we could somehow get rid of immoral and dishonest people. \\
v0627 & What we need are less laws and less institutions, and more courageous, indefatigable, and devoted leaders, in whom the people can put their faith. \\
v0634 & In spite of what some people keep saying, the lot of the average man is getting worse. \\
v0636 & These days a person does not really know whom he / she can count on. \\
v0637 & Criticizing the government is useless, because the government simply does what it considers to be proper. \\
\hline
\end{tabular}
\end{table}
at the opposed sides of the political spectrum. The effects are weaker for those parties for which the issue of ethnocentrism is less politically relevant. This applies to the centrist D66, and to the smaller religious parties (CU and SGP). The preference for the latter parties is positively associated with authoritarianism, but there is no interaction with group identification.

The two presented analyses show that the integrated model is supported by the data: group identification modifies the influence of authoritarianism onto both the degree of ethnocentrism and preference for parties whose ideological orientation contains elements relevant for intergroup relations.

There are two important qualifications concerning the obtained findings. First, the obtained coefficients are estimated minimum associations due to the characteristics of the variables transferred via the MI procedure. Second, the conclusions are based on the previous research, namely the previously observed association between authoritarianism and predictors common for the two data sets. In other words, the conclusions rest on the assumption that the relationships present in the SOCON 2000 study are also valid in the NKO 2002. It is not difficult to support this assumption. Both data sets are intended to be representative for the Dutch population. The two studies were conducted close in time, and employed similar field methods. The relationships used for the imputation are theoretically grounded (the relationship of authoritarianism with education, age, etc., see Verberk et al., 2002), exhibit considerable cross-cultural validity and stability (Meloen 1993), and are verified using different Dutch samples (Verberk et al., 2002, Scheepers et al., 1990).

**Discussion and Conclusions**

The previous sections showed that the MI approach to statistical matching results in variables that are close to the actual respondents’ answers, and is applicable to substantive research problems. Although the statistical foundation of the method is already well established (Rubin, 1986; Gelman et al., 1998; Rässler, 2002, 2003; Moriarity and...
Scheuren, 2003), its application to substantive problems has still been rare outside of marketing research and official statistics (although see Rässler, 2002; He et al., 2007).

Obviously, certain non-statistical considerations influence the perception of statistical matching procedures. From the perspective of traditional data collection and analysis, the outlined approach might indeed appear as creating inauthentic data. However, what the procedure actually does is use the extant (“previous”) research to address new research problems (Aluja-Banet and Thiò, 2001). Thus, from the perspective of finding new ways to use the existing research, this is a promising approach. The traditional method of “taking the previous research into account” is to refer to previously published findings in a brief narrative form. Rare are introductions to research articles that do not contain phrases such as “It has been found...”. Unfortunately, such narrative references are often vague, usually cover only a fraction of the relevant research, and occasionally do not refer to findings that are directly related to the research problems in question.

Quantitative meta-analysis is a more advanced approach, although not without its own potential shortcomings. For instance, the available meta-analyses may focus on problems that do not exactly fit one’s research questions. Publication bias is another source of imperfection (Rosenthal, 1979; Scargle, 2000). In any case, both the narrative method and quantitative meta-analysis involve information loss.

MI approach to statistical data matching, however, uses previous research in a much more thorough and neutral, or unbiased, way. Instead of including results of previous analyses, MI approach incorporates previously collected data. Much of the wealth of information contained in the data is typically lost to the other approaches. This refers especially to distributional features and complex relationships between variables. Relationships in the already collected data, even if present only implicitly (not reported in literature), can be seen as the best guesses about the relationships in comparable data sets. Statistical matching represents a method to exploit this, by integrating the previous research directly into the research design.

The selection of variables for the imputation model and the conditional independence problem are two related issues that deserve particular attention when statistical matching is considered. Success in data transfer depends on the availability of suitable variables that occur across data sets and which provide ground for data matching. The best policy in variables selection seems to be to follow the general guidelines for MI: imputations should be based on as many variables as feasible, and the model should be as complex as needed. For instance, no relevant interactions should be omitted. According to Rässler (2003: 60):

if the common variables are (carefully) chosen in a way that establishes more or less conditional independence among the variables not jointly observed given these common variables, then inference about the actually unobserved association is valid. In terms of regression analysis this implies that the explanatory power of the common variables is high concerning the specific variables.

The quoted paragraph suggests that well-selected variables that enter the imputation equation already attenuate the problem of conditional independence. The problem is also
attenuated if there are theoretical reasons not to include all the predictor variables into the analysis, as demonstrated in the substantive research example above. The research question concerned the differences in the degree of the relationship under different conditions, and therefore there was no need to include many control variables. However, since the number and character of commonly occurring variables is typically limited, this paper presented a technique to deal with the conditional independence by simulating a “third” data set.

Another aspect of the variable selection problem is more difficult to specify in general terms. Namely, the essential requirement is that the relationships observed in the donor data set are valid in the receiver data set. Concerning the example presented above, it is easy to argue that the negative association between education and authoritarianism is a stable relationship, and that there are no reasons to believe that it changed between 2000 and 2002. It is easy to imagine examples where this would be more difficult to establish. Thus, any conclusion based on the MI, or any other, approach to data matching has to be qualified in the sense that it is based on the assumption that the relationships that provided ground for the imputation hold across the involved data sets.

To summarize, the theoretical foundation of statistical matching seems well established in literature (Rubin, 1986; Gelman et al., 1998; Rässler, 2002, 2003). The first part of this paper demonstrated some of the essential features of the procedure. The exercises showed that the imputed scores tend to be very close to what respondents actually answered. The main point is not only that one can predict respondents’ actual L-R scores on the basis of the relationships obtained with completely different individual respondents two years earlier. The most important point is that the variable transferred via MI shows comparable relationships with various other variables, whether or not they were included in the imputation equation. The substantive example showed how statistical matching enabled the examination of a hypothesis that would not be possible otherwise.

Because of the imperfect ability to predict target variables, a considerable degree of uncertainty is introduced in the course of statistical matching. As a result, the relationships between the imputed and variables not included in the model tend to be underestimated, while standard errors are overestimated. It was also shown that the addition of a simulated ‘third’ data set could be useful in dealing with the conditional independence problem.

Not only that MI approach is statistically superior to alternative data matching methods, but it is also a promising method for “taking the previous research into account”. Science is supposed to have a cumulative nature. Therefore, it is important to maximally use the results of previous research. It is clear that the published research does not necessarily represent all the (potentially) known truth about the reality. The MI approach to statistical matching is unique in that it allows the incorporation of the existing data directly into the research design. Further research is required in order to improve the method and make it more applicable. Simulation studies should be helpful in examining and quantifying the consequences of varying degrees of reliability and conditional independence for the relationships in question.

Notes


4. *Ice* performs multiple imputation for a set of variables, and is able to adjust to different levels of measurement. In addition to *Ice*, I also used Stata procedure *uvis*, which imputes missing values in the single variable based on multiple regression on a list of predictors. *Uvis* is called repeatedly by *ice* in a regression switching mode to perform multivariate imputation.

5. The first wave (pre-election) interviews were conducted between 18 April and 14 May 2002. The post-election (second wave) wave began on 16 May, the day following the election, and lasted until 27 June 2002. The third wave interviews were held between January and March, 2003.

6. No substantive differences were observed if the scale items were first transferred (cross-imputed to NKO data) and then used to calculate the scale scores.

7. According to Wald’s test, the coefficients from the two equations are not equal ($F(2, 1266) = 9.82, p < .001$).

References


