10.5465/AMBPP.2015.69

RESEARCH WITH SECONDARY DATA: DIFFERENT MATCHING METHODS AND IS THERE A DIFFERENCE?

TIM DE LEEUW
Department of Management, Tilburg School of Economics and Management,
Tilburg University, Warandelaan 2, 5037 AB Tilburg, Netherlands

STEFFEN KEIJL
WU Vienna University of Economics and Business.

ABSTRACT

Multiple secondary databases are frequently used. However, papers rarely discuss the method of combining these databases. We report on four methods of matching data across secondary databases and show that the number of observations differ significantly based on the method used. Therefore research results might be dependent on these methods.

INTRODUCTION

Our review of all the studies published in one of the top journals in the field of management (i.e., the *Strategic Management Journal*) in the last six years reveals that 358 of the total 469 studies make use of secondary data, and 297 studies use data derived from multiple secondary databases, which are combined into one dataset for the analyses. Hence, 63 percent of all *SMJ* studies published in the last six years (i.e., 297 of the total 469) use multiple secondary databases. Although these studies have enriched our understanding of many managerial phenomena, only 26 of the 297 studies (11 %) that used multiple secondary databases report on the method that was used to match the data across these different secondary databases (e.g., Csaszar, 2012; Zhou, 2011; Zhu, 2013).

This large-scale absence of reporting regarding the matching methods in quantitative studies that use multiple secondary data sources results in a lack of common methods for matching data across frequently used secondary databases and many researchers are re-inventing the wheel, due to the limited possibilities for learning from prior studies. With regard to different matching methods some studies used a name-matching method (e.g., Arora and Nandkumar, 2012; Heeley and Jacobson, 2008), while others used a common identifier (e.g., Kim *et al.*, 2013; Lim and Mccann, 2013) and still others (software-aided) used name standardization and additional validations (e.g., Zhou, 2011; Zhu, 2013).

An even more important issue concerns whether research findings might be dependent on the method used for matching data across multiple secondary databases. Specifically, possible mismatches (e.g., matching data from firm “Alpha Inc.” in database I with data from firm “Alfa Corp.” in database II) can negatively influence the internal validity of studies that use secondary data. Additionally, a significant amount of information might be overlooked when the matching method is too narrow and firms that are, in reality, the same are not matched (e.g., not matching data from firm “Alpha Inc.” in database I with data from firm “Alpha” in database II). Moreover, due to a lack of attention to how multiple secondary databases can be matched, reviewers and editors might not always know what to look for and ask about during the review process, which leaves a possibly vital part of the research methodology underexposed.

10.5465/AMBPP.2015.69

This study fills this void by specifying and investigating four different matching methods for making connections between two frequently used secondary databases. The four consecutive methods for matching data across four different secondary databases are: (1) Full name matching (e.g., individual or firm "Alpha" does not equal firm "Alfa," and firm "Beta Inc." does not equal firm "Beta Corp."); (2) Common identifier matching (e.g., when two firms have the same CUSIP code); (3) Full name and common identifier matching, including additional checks, such as comparing matching results for both matching procedures; and (4) Full name and common identifier matching, including additional checks and an additional fuzzy name match with additional identification variables such as corporate address. These four matching methods are successive, because each higher hierarchical step will include more positive matches and exclude potential mismatches.

THEORETICAL ELABORATION OF MATCHING METHODS

Two types of errors can emanate from potential mismatches between multiple databases. For example, (financial) information for firm Alpha is obtained from database I and other information (e.g., about Inter-Organizational Relations (IORs)) from database II. An initial mismatch could happen if no or not all "true" observations (e.g., IORs) in database II were matched to firm Alpha, because, say, it was registered differently in the two databases. This would mean that firm Alpha was assumed to have fewer observations (e.g., IORs) than it actually did, which we label a so-called false negative. A second mismatch might happen if more than all of the "true" observations (e.g., IORs) in database II were matched to firm Alpha. This could happen if multiple, but different, firms "Alpha" were to appear in the other database, thereby creating an assumption that firm Alpha had more observations (e.g., IORs) than it actually did, which we label a so-called false positive. In either case, a mismatch would mean that the firms were not aligned with the number of observations (e.g., IORs) they actually had or were wrongly assigned to observations (e.g., IORs) they did not possess.

METHODS OF MATCHING SECONDARY DATA

There are multiple methods of combining databases of which four are described. The first method uses the full name of the unit of analysis (e.g., name of the individual or firm). With this method, the observations from multiple databases can be combined if the full name in one database corresponds to that in another database. Although this seems a relatively easy method for matching data across multiple databases, individuals or firms can have slightly different names in those various databases, so that mismatches are very likely to occur when this method is used exclusively.

A second method uses a common identifier for the unit of analysis, which then must be available in multiple databases, such as firm identifiers like CUSIP or GVKEY (e.g., Kim *et al.*, 2013; Lim and Mccann, 2013). An important issue is that not all databases have the kinds of alphanumeric identifiers that seem to be common in other databases, and sometimes these common identifiers are designated with a different number of alphanumerical characters in different databases. However, when there are common identifiers, these can be rather easily used to combine observations and match data from multiple databases.

A third method of combining databases represents a combination of the first two, with additional validations. That is, researchers firstly search for observations that can be found in

10.5465/AMBPP.2015.69

multiple databases based on a full name match. Then, independently of the full name match, they search for observations in multiple databases based on a common identifier match. In a third step, they combine the observations from the two matches, which can result in three categories of observations across the databases: (A) observations that are matched based on both the full name match and the common identifier match (i.e., double matched); (B) observations that are only matched based on the full name and not on the common identifier; and (C) matched observations based on the common identifier and not on the full name.

The researchers subsequently have a variety of ways for dealing with these observations. Observations that match in terms of both the full name and the common identifier are "double matched" observations, which are very likely to be the same individual or firm and therefore a true match. The two categories of observations that are matched only in terms of either the full name (B) or the common identifier (C) might either be excluded as a mismatch (Matching Method 3, sub-option I) or manually evaluated based on additional information on the individuals or firms within each database, such as zip codes or websites or as obtained through web searches (Matching Method 3, sub-option II).

A fourth method of combining multiple databases deals with so-called fuzzy name matching, which can be done in addition to the prior matching methods. With matching methods four, sub-option I: The fuzzy name matching procedure starts with the standardization of firm names. Specifically, the names of firms can be standardized by, for example, removing additions at the end of the name, such as "Inc." or "Corp." (e.g., Albany Molecular Research Inc. becomes Albany Molecular Research). This standardization can be achieved relatively quickly, but it does not capture the different ways in which a name might be written (e.g., Albany Molecular Research Inc. might also be registered in a database as AMRI).

Alternatively (Matching Method 4, sub-option II), specific identifying keywords can be derived from a firm name so that the different forms in which it is written can be incorporated into the matching procedure, with multiple keywords derived for every firm name. As an example, the three keywords (including search wildcards, such as the %) for Albany Molecular Research Inc. might be: ALBANY%MOLECULAR%; AMRI%; ALBANY%. A balance needs to be found when generating these keywords between words that are too generic (e.g., "research") and will result in a high number of potential mismatches and words that are too specific and could potentially leave out true matches for a firm that is registered differently across multiple databases.

A way to deal with this is to generate the keywords sequentially in such a way that the first words are more specifically defined and the later words more generic. Combining this ordering of keywords with matching based on the full name and the common identifier decreases the number of observations that need to be manually verified. This is because all the observations that are matched both on full name and common identifier can then be subtracted from the matches based on the first keyword. The matches that remain need to be verified based on other criteria, such as zip codes, websites, or background information found on the internet. Pending verification, these observations can be added to the full name and common identifier matches. Matching the names between multiple databases in this way is practical, provides a lot of insight into the units of analysis, and incorporates the researcher's judgment, but it is also time-consuming.

DIFFERENT MATCHING METHODS: DOES IT MATTER?

10.5465/AMBPP.2015.69

In order to investigate whether different methods of matching data across multiple secondary databases can have an impact on the number of observations, two frequently used databases (i.e., Compustat and SDC) were connected by means of the four matching methods. Compustat contains financial and descriptive information on firms, while SDC contains information on the IORs of firms. The largest (based on number of employees) 282 US public biopharmaceutical firms in Compustat are selected and connected to the SDC database. With regard to the third Matching Method (M.M.) the additional validation of records that were not double matched (i.e., matched both on the full name (M.M. 1) and on a common identifier (M.M. 2)) was used (i.e., sub-option II). For the fourth matching method the development of key words was used (sub-option II).

Table 1 about here

As can be observed from Table 1, the fourth matching method (M.M. 4) results in the highest numbers of matched IORs, while the full name match (M.M. 1) captures 58.07 percent of the IORs compared to the fourth method. This table clearly shows that the number of matched records and information obtained on IORs increases consecutively from the first to the fourth method, although the amount of work also significantly increases.

Based on these differences it is interesting to investigate if these differences in the number of matched IORs are distributed equally among the firms. For this we took the top 15 (with regard to the number of IORs) for each of the four matching methods and compared them. The results are presented in Table 2. This comparison was done per year and Table 2 is based on the year 2000 (randomly selected, the year 2000 shows similar results compared to the other years). The firms are sorted based on the number of IORs based on the fuzzy name match (M.M 4).

Table 2 about here

The results show for instance that Safeguard Scientifics inc has 67 IORs based on the fuzzy name match (M.M. 4) and is positioned as the firm with the third most IORs, while there are 22 IORs found based on the other three matching methods, resulting in positions 15 (M.M. 3), 13 (M.M. 2), and 12 (M.M. 1). Glaxosmithkleine plc for instance has 47 IORs based on matching methods 4, 3 and 2, while the full name match (M.M. 1) finds no IORs, resulting in a large difference with regard to the position in the top 25 (i.e., position 107 for the full name match (M.M. 1) and position 5 based on the fuzzy name match (M.M. 4.).

DISCUSSION AND CONCLUSIONS

The aim of this study was to investigate whether the different matching methods that are used to combine multiple secondary databases for developing datasets can have an impact on research findings. We focused on four particular matching methods for combining secondary databases, as well as a number of frequently used databases containing information on IORs, as

10.5465/AMBPP.2015.69

described above. The results of this study show that empirical findings might differ depending on the method that was used to match (i.e., combine) multiple databases.

We started this paper with our observation that secondary databases are frequently combined for research purposes (i.e., 63 percent of all SMJ studies published in the last six years), but that only a small amount of these studies (i.e., 11 %) report on the method that was used to match the data across these different secondary databases. Based on the found differences between the different matching methods we would recommend scholars to precisely and consciously describe the matching procedure as part of the methods section of a paper. Additionally, we would recommend reviewers and editors to explicitly ask for this. This does not only allow for an accumulation of knowledge on the different matching methods used, but also increase the clarity of the matching method conducted and enables others to replicate research findings.

Additionally, based on the results, we would recommend scholars to use a rather thorough and extensive matching method, in future studies, such as the above-described matching methods which combines the full name match with the common identifier match including the additional validation (M.M. 3) or even the additional fuzzy name match (M.M. 4). If the additional fuzzy name matching is used we can recommend the development of the hierarchical key words (i.e., fuzzy names, sub-option II), by multiple researchers independently. Based on the results we would advise scholars to refrain from using only a full name matching method.

This study is not without its limitations. First, the different matching methods were investigated based on firm data, whereas future research on the individual level (e.g. managers or inventors) could also consider the impact of different matching methods. Second, we do not claim that the proposed matching methods are superior and invite other scholars to share other creative methods to combine data across multiple secondary databases. Third, in line with other studies (e.g., Ahuja, 2000; Beckman, Haunschild, and Phillips, 2004; Hitt *et al.*, 1996; Keil *et al.*, 2008), we have only focused on large firms, whereas the differences in the number of observations might be different for privately owned or smaller firms.

Altogether, this study shows that there are large differences between difference matching methods and that research findings can differ based on the method that is used to match data across multiple (secondary) databases and therefore calls for more attention to at least the description of the matching method used in a study.

ENDNOTES

1. The authors are listed alphabetically and contributed equally to this paper. This paper benefited from a grant from the WU Vienna University of Economics and Business.

REFERENCES AVAILABLE FROM THE AUTHORS

Table 1: Number of IORs per matching method (M.M.)

	M.M. 1.	Compared to M.M.4	M.M. 2.	M.M. 3.	M.M. 4.
Non-equity alliances	1,482	68.93 %	1,782	1,921	2,150
Joint ventures	370	61.67 %	453	484	600
Minority investments	286	70.79 %	388	358	404
M&A	1,606	67.96 %	1,856	2,014	2,363
Total # of IOR	4,961	58.07 %	5,991	6,410	8,543
# of focal firms with 1 or more IOR	185	75.82 %	215	234	244

Table 2: Top 15 firms with the most IORs per matching method (M.M.)

Firm name	Position M.M. 1	# IORs M.M. 1	Position M.M. 2	# IORs M.M. 2	Position M.M. 3	# IORs M.M. 3	Position M.M. 4	# IORs M.M. 4
JOHNSON & JOHNSON	5	40	6	40	6	40	1	100
AVENTIS SA	1	76	1	76	1	76	2	76
SAFEGUARD SCIENTIFICS INC	12	22	13	22	15	22	3	67
NOVARTIS AG	2	48	2	49	2	48	4	64
GLAXOSMITHKLINE PLC	107	0	3	47	3	47	5	47
ELAN CORP PLC	4	41	5	41	5	41	6	46
ABBOTT LABORATORIES	3	42	4	42	4	42	7	43
WYETH	6	36	7	36	7	36	8	36
ROCHE HOLDING AG	7	34	140	0	8	34	9	36
PFIZER INC	8	29	8	33	9	33	10	33
BAXTER INTERNATIONAL INC	9	28	11	28	12	28	11	32
LILLY (ELI) & CO	108	0	9	31	10	30	12	31
BRISTOL-MYERS SQUIBB CO	10	28	10	31	11	30	13	31
ASTRAZENECA PLC	13	21	12	27	13	27	14	27
BAYER SCHERING PHARMA AG	119	0	52	5	55	5	15	25