COMBINING FIRM-LEVEL SECONDARY DATA: DIFFERENT MATCHING METHODS DO NOT MATCH

TIM DE LEEUW *
Department of Management, Tilburg School of Economics and Management,
Tilburg University, Warandelaan 2, 5037 AB Tilburg, Netherlands
E-mail: t.deleeuw@tilburguniversity.edu

STEFFEN KEIJL
Institute of Strategy, Technology and Organization, Department of Strategy and Innovation,
WU Vienna University of Economics and Business.
E-mail: steffen.keijl@wu.ac.at

* The authors are listed alphabetically and contributed equally to this paper.

ABSTRACT

Many studies in management base their results on secondary databases, but only a few describe how data were matched. We investigated how of inter-organizational relationships affect innovation, using four different matching methods. Since results differ based on the matching method, we argue that reporting the matching method used in a study is of vital importance.

INTRODUCTION

Our in-depth review of recent studies published in one of the top management journals, reveals that the majority is quantitative and more than half use multiple secondary databases. Only a small number of these studies (<10 percent), reports on the method that was used to match the data across the different secondary databases (e.g.. Csaszar 2012, Zhou 2011, Zhu 2013). There seems to be no common method for making connections between frequently used firm-level secondary databases. 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 used (software-aided) name standardization and additional validations (e.g., Zhou 2011, Zhu 2013). This large-scale absence of reporting with regard to the matching methods used in many of the quantitative studies obfuscates the research process and hampers the accumulation of knowledge on these matching methods. An important question is whether these different matching methods might impact the research findings. Therefore this study aims to address the following research question: which methods of combining firm-level secondary databases are used, do research results differ depending on these different matching methods, and which method should be preferred?

The goal of this study is to investigate different matching methods for making connections between secondary databases containing firm-level data. Four consecutive methods for matching firm-level data across four different secondary databases are identified: (1) Full name matching; (2) Common identifier matching; (3) A combination of the first two, including additional checks; and (4) similar to method 3 with an additional fuzzy name match with supplementary verifications. These four matching methods are consecutive, since each higher hierarchical method will include more real matches and exclude potential. In order to reach our goal, we empirically investigate a key topic within the field of management, which is the

relationship between inter-organizational relationships of a firm and that firm's innovative performance based on the four different matching methods. In the final section, we elaborate on the potential problems and biases that can result from not reporting or, even worse, the use of only one matching method.

THEORETICAL BACKGROUND

Although matching firm-level data from multiple secondary databases is an important step in the process of conducting a study in the field of management, there does not seem to be any common, well-spelled-out method for doing so. Moreover, as our literature review revealed, most studies do not disclose any information on how they retrieved and matched the data from secondary databases. Only a few studies mentioned the method used, with a minor fraction going more into depth on how matches were double, or even triple, verified. This deficiency, (i.e., the lack of reporting on matching methods), has the potential to impact the empirical findings from such data projects. When the data from multiple secondary databases is erroneously combined, this can result in two types of errors, which in turn can influence the findings of the study.

These two types of errors emanate from potential mismatches between two (or more) databases. For example, one could sample all financial information for a firm "Alpha" from one database and search for information about its inter-organizational relations from a second database. A first mismatch will occur if more than all of the "true" inter-organizational relations of "Alpha" in the second database are matched to "Alpha". This can happen if multiple, but different, firms "Alpha" appear across different databases, thereby creating an assumption that "Alpha" had more IORs than it actually did, which we label a so-called "false positive".

A second mismatch will happen if no (or not all "true") inter-organizational relations in this second database are matched to "Alpha", because it is registered under slightly different names in the databases. This would mean that the firm "Alpha" is assumed to have fewer inter-organizational relations than it actually did, which we label a so-called "false negative". In either case, a mismatch means that the firms are wrongly assigned to inter-organizational relations they did not possess, or are not aligned with the inter-organizational relations they actually had. In general this means that data from another database would either wrongly be assigned to the firms or would not be incorporated.

The implications of both false positives as well as false negatives for the empirical results might be an erroneous estimation of regression coefficients in terms of direction and/or size, and a wrongfully determination of statistical significance. With regard to the coefficients and variance, this results in a potentially increase of the error terms, making it less likely that the true effects would be statistically significant. Moreover, this would indicate that true curvilinear relationships between two variables might not be discovered, due to a lack of statistical power with increased standard errors, and that instead, statistically significant linear relationships would be wrongly assumed to be true.

In light of these potential matching fallacies and ensuing issues, we investigate the implications of four different matching methods for empirical results by investigating the relationships of inter-organizational relations and firms' innovative performance. Therefore, we combined data on firms from four prominent secondary databases by means of four consecutive matching methods. Specifically, we extent a study performed by Keil et al. (2008) that explored the linear relationships of multiple inter-organizational relationship modes (e.g. alliances, joint ventures,

mergers and acquisitions) on a firm's innovative performance. Additionally to the linear relationships, we incorporate recent theoretical developments (e.g., Jiang et al. 2010) and allow for the possibility of inverted U-shaped relationships between a firms' number of IORs and its innovative performance. Furthermore, we incorporate the distribution across different modes of inter-organizational relations (i.e., the diversity across a firms' licensing agreements, non-equity alliances, corporate venture capital investments (CVC investments), minority investments, joint ventures (JVs), and mergers and acquisitions (M&As)).

EMPIRICAL BACKGROUND

In order to investigate the empirical impact of different matching methods on research outcomes we combined data (four times, each time based on one of four matching methods) from four frequently used secondary databases, which contain information on financial information (Compustat database), different types of IORs (two SDC databases and VentureXpert), and patents (USPTO/NBER database).

Since IORs and innovative performance of firms are key attention areas within the field of management research (Keil et al. 2008), we investigate the impact of these inter-organizational relations of a focal firm on that firms' innovative performance. To realize this we focus on the largest 282 companies in the biopharmaceutical industry.

After defining the sample, four matching methods were used to connect the data of the 282 focal firms in the Compustat database to their data from other databases (i.e., SDC Joint Ventures and Alliances, SDC M&A, VentureXpert) and finally assess their innovative performance (based on the USPTO/NBER patent data). These four methods of matching firms across the different databases are: (M.M. 1) a full name match; (M.M. 2) a match based on a common identifier; (M.M. 3) a match which combines the first two, including additional validations; and (M.M. 4) a match which takes the matching results from the third method supplemented with an additional fuzzy name match based on key-words, including additional validations. Gvkey and cusip identifiers were used as common identifiers. The additional validation checks in the third and fourth method, based on addresses, zip codes, websites, and web searches, were performed by the two authors independently and any small differences that arose were discussed and resolved.

Innovative performance, which is the dependent variable, is measured — in line with previous studies (e.g., Ahuja and Katila 2001, Keil et al. 2008) — as the number of successful patent applications filed each year. The independent variables are measured by the number of IOR modes a firm has in a particular year. Diversity of inter-organizational relation modes in a firm's inter-organizational relation portfolio (i.e., all inter-organizational relation of a focal firm) is also incorporated into the analyses and was operationalized by means of a Herfindahl-Hirschman Index (HHI, i.e., the sum of the squared share of the number of inter-organizational relations per inter-organizational relation mode), also known as Blau's index of heterogeneity. The analyses controlled for the main SIC industry classification of the firm, as well as the years of observation (1990–2005), by introducing dummy variables. Additionally, firm size (the log of annual sales) and R&D intensity (R&D expenditure divided by sales) were controlled for since these variables can have an impact on the value derived from the IOR. Finally, regarding the method of analysis,

negative binomial regression panel data estimations was used, as is common in studies explaining patent counts. A random effects specification has been applied.

RESULTS

Table 1 presents the results of the analyses. Models 1-4, in table 1, test for a linear relationship between the number of inter-organizational relations and a firm's innovative performance. Subsequently, models 5-8 also incorporate squared terms for the different inter-organizational relation modes and the inter-organizational relation portfolio diversity measure. Matching methods (M.M.) 1 through 4 correspond to the four different matching methods described previously, where M.M. 1 is only full name match and M.M. 4 is based on the most comprehensive matching method.

Insert table 1 about here

A comparison of the matching methods in the first four models reveals that the results for the licensing agreements, non-equity alliances, joint ventures, and overall portfolio diversity seem to be stable independent of the matching method used for data from multiple databases (e.g., licensing agreements: $\beta = 0.03$, $p < 0.05$ for all four matching methods). However, the results for the CVC, minority investments, and M&A vary depending on the matching method used. For example, the number of M&A does not have a statistically significant effect ($p < 0.05$) on innovative performance when the dataset is based on a full name match (M.M. 1: $\beta = 0.00$, n.s.), a common firm identifier match (M.M. 2: $\beta = 0.02$, $p < 0.10$), and the fuzzy name match (M.M. 4: $\beta = 0.02$, n.s.), while the name and common identifier combined matching method show a significant relationship (M.M. 3: $\beta = 0.02$, $p < 0.05$).

A comparison of the results for Models 5-8, which include the squared coefficients to account for the inverted U-shaped relationships, shows that licensing agreements do not significantly affect a firm's innovative performance when the dataset is based on a full name (M.M. 1: $\beta = 0.05$, $p < 0.10$) or common identifier matching methods (M.M. 2: $\beta = 0.04$, $p < 0.10$), whereas the effect does become significant when the combined name and common identifier (M.M. 3: $\beta = 0.05$, $p < 0.05$) and fuzzy name matching methods are used (M.M. 4: $\beta = 0.04$, $p < 0.05$). Also the relationship between non-equity alliances and innovative performance is not significant based on the full name matching method (M.M. 1: $\beta = 0.01$, n.s.; $\beta 2 = 0.00$, n.s.), whereas it is significant based on the other matching methods (e.g., M.M. 4: $\beta = 0.07$, $p < 0.01$; $\beta 2 = 0.00$, $p < 0.001$). With regard to the IOR portfolio diversity, the datasets based on full name matching indicates a positive relationship to innovative performance (M.M. 1: $\beta = 1.48$, $p < 0.001$; $\beta 2 = -0.93$, n.s.), while the expected inverted U-shaped relationship is found based on the other matching methods (e.g., M.M. 4: $\beta = 2.21$, $p < 0.001$; $\beta 2 = -1.62$, $p < 0.01$).

DISCUSSION AND CONCLUSIONS

The aim of this study was to investigate whether the different matching methods might have an impact on empirical research findings. We focused on four particular matching methods for combining firm-level secondary databases, as well as a number of frequently used databases

containing firms' financial information, inter-organizational relations, and patent information. Subsequently, to investigate the potential impact these different matching methods can have, we investigated the relationships between the inter-organizational relations and innovative performance.

The results of this study show that empirical findings can indeed differ depending on the method that was used to match multiple databases. For instance, there does not seem to be a relationship between non-equity alliances and innovative performance when the data on the firms are combined based exclusively on full name matching. However, when the matching method incorporates additional matching data, the relationship between alliances and innovation becomes a statistically significant, inverted U-shaped relationship. In a similar vein, the overall diversity of the inter-organizational relation modes seems to be positively related to innovative performance when a full name match is used. And yet, this relationship turns out to be an inverted U-shape, as theoretically expected, when the inter-organizational relations obtained though the other matching methods are incorporated. We thus showed the impact of different matching methods on the relationships between different inter-organizational relation modes and innovative performance and found differences between the datasets developed with these different matching methods.

We started this paper with our observation that secondary databases are frequently combined for research purposes (e.g., 63 percent of all SMJ studies published in the last six years), but that only a small amount of these studies (e.g., 8 percent in SMJ) report on the method that was used to match the data across these different secondary databases. Since our detailed review of the SMJ papers revealed that all matching methods are used, some of these previous findings might be different when another matching method would have been used. For the vast majority of existing studies in many top journals the exact method of matching the data between secondary databases could not be determined, but since the large differences found between the different matching methods some of these results could also be dependent on the matching method used.

Based on the found differences between the different matching methods we would therefore 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.

Altogether, this study describes four different matching methods and shows that there are large differences between them. Next to large differences in the matched number of IORs, research findings can differ based on the method that is used to match data across multiple (secondary) databases. As such we call for more attention to at least the description of the matching method used in a study and advices scholars to use a rather thorough and extensive matching method, which ideally combines a full name match, with a common identifier match, followed by an additional fuzzy name match based on key-words, including additional verifications (i.e., M.M. 4).

REFERENCES AVAILABLE FROM THE AUTHORS

Table 1: IORs and innovative performance, based on different matching methods (M.M.)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Matching method	M.M. 1	M.M. 2	M.M. 3	M.M. 4	M.M. 1	M.M. 2	M.M. 3	M.M. 4
Licensing agreements	0.03*	0.03*	0.03*	0.03**	0.05+	0.04+	0.05*	0.04*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)
Licensing agreements^2					-0.00	-0.00	-0.00	-0.00
					(0.00)	(0.00)	(0.00)	(0.00)
Non-equity alliances	0.00	-0.02	-0.01	-0.01	0.01	0.05*	0.05*	0.07**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.02)	(0.02)	(0.02)
Non-equity alliances^2					-0.00	-0.00***	-0.00***	-0.00***
					(0.00)	(0.00)	(0.00)	(0.00)
CVC investments[a]	-0.14*		-0.14*	0.00	-0.13		-0.07	-0.01
	(0.06)		(0.06)	(0.00)	(0.18)		(0.18)	(0.01)
CVC investments^2					-0.00		-0.01	0.00
					(0.03)		(0.03)	(0.00)
Minority investments	-0.04	-0.05+	-0.05+	-0.09***	0.05	-0.04	-0.05	-0.06
	(0.03)	(0.03)	(0.03)	(0.03)	(0.07)	(0.06)	(0.06)	(0.06)
Minority investments^2					-0.00	0.00	0.00	0.00
					(0.01)	(0.01)	(0.01)	(0.01)
Joint ventures	-0.08***	-0.08***	-0.09***	-0.06***	-0.16***	-0.13***	-0.14***	-0.13***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.02)
Joint ventures^2					0.00***	0.00***	0.00***	0.00***
					(0.00)	(0.00)	(0.00)	(0.00)
M&A	0.00	0.02+	0.02*	0.02	0.07*	0.08**	0.09**	0.06*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.03)	(0.03)
M&A^2					-0.00*	-0.01**	-0.01**	-0.00*
					(0.00)	(0.00)	(0.00)	(0.00)
Portfolio diversity	0.97***	1.06***	1.12***	1.28***	1.48***	1.81***	1.85***	2.21***
	(0.13)	(0.13)	(0.13)	(0.13)	(0.44)	(0.41)	(0.42)	(0.40)
Portfolio diversity^2					-0.93	-1.48**	-1.44*	-1.62**
					(0.62)	(0.57)	(0.58)	(0.54)
Firm size	0.10***	0.10***	0.09***	0.09***	0.10***	0.09***	0.09***	0.08***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R&D intensity	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	-1.72***	-1.75***	-1.73***	-1.74***	-1.74***	-1.79***	-1.76***	-1.74***
	(0.41)	(0.42)	(0.42)	(0.42)	(0.41)	(0.42)	(0.42)	(0.42)
# of observations	2,652	2,652	2,652	2,652	2,652	2,652	2,652	2,652
# of firms[b]	265	265	265	265	265	265	265	265
Log likelihood	-5491	-5486	-5477	-5468	-5478	-5462	-5452	-5439

Dependent variable: Annual count of successful patent applications. Models 1 and 5 use matching method 1 (full name), Models 2 and 6 use matching method 2 (common identifier), Models 3 and 7 use matching method 3 (methods 1 and 2 including additional validations) and Models 4 and 8 use matching method 4 (method 3, including fuzzy name matching). Standard errors in parentheses. *** p<0.001, ** p<0.01, * p<0.05, + p<0.10, based on two-sided tests. Controlled for the SIC industries and years in all models.

[a] For matching method 2 (common firm identifier), there was no match possible with regard to the CVC investments database (i.e., this database does not have a common firm identifier).

[b] 17 firms dropped out of these analyses due to missing observations.