Inter-American Development Bank

The Longitudinal Linkage of Mexico's Economic Census 1999-2014

Department of Research and Chief Economist

Matías Busso
Oscar Fentanes
Santiago Levy

TECHNICAL NOTE Nº
IDB-TN-1477

September 2018

The Longitudinal Linkage of Mexico's Economic Census 1999-2014

Matías Busso
Oscar Fentanes
Santiago Levy

Inter-American Development Bank

September 2018

IDB
Inter-American Development Bank

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Busso, Matías.
The longitudinal linkage of Mexico's economic census 1999-2014 / Matías Busso, Oscar Fentanes, Santiago Levy.
p. cm. — (IDB Technical Note ; 1477)
Includes bibliographic references.
1. Industrial statistics-Mexico-Longitudinal studies. 2. Mexico-Census-Longitudinal studies. 3. Mexico-Economic conditions-Longitudinal studies. I. Fentanes, Oscar. II. Levy, Santiago. III. Inter-American Development Bank. Department of Research and Chief Economist. IV. Title. V. Series.
IDB-TN-1477

Abstract*

This technical note describes the methodology to construct a longitudinal dataset using the Economic Censuses of Mexico from 1999 to 2014. The procedure is based on an algorithm that links establishments with identical or significantly similar location, legal entity and industry. Since a set of longitudinal identifiers is already available for the 2009 and 2014 Economic Censuses, it is used to validate our results, obtaining 90% accuracy. The paper links 1.44 million establishments for the period 1999-2004, 1.52 million for 2004-2009 and 2.15 million for 2009-2014.

Keywords: Longitudinal data, Economic census, INEGI
JEL codes: C81, D21

* We would like to thank INEGI for the access to the data used to produce this linkage. The linkage presented here is a work undertaken by IDB staff and does not belong to INEGI's official records. All the work took place at the INEGI's Microdata Laboratory under strict surveillance to protect individual businesses. The identifiers created and all data mentioned here can only be consulted at the INEGI facilities after approval. We would also like to thank Jesica Torres Coronado for helpful suggestions and comments. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, or the countries they represent.

1 Introduction

The National Institute of Statistics and Geography of Mexico (INEGI) produces the quinquennial Economic Census since 1989. The census collects data on all business establishments operating in fixed facilities and located in urban localities with more than 2,500 people.

The 2009 and 2014 editions of the Economic Census introduced the identifier CLEE[1], which longitudinally links establishments from both censuses and will, eventually, link subsequent editions. While the CLEE is already used for longitudinal studies, it is not available for preceding editions, limiting the potential of these databases.

In this note we describe the linkage process to create a longitudinal database from 1999 to 2014. Even though the 1989 and 1994 editions could potentially be linked too, they pose difficulties because their industrial classification was discontinued, and their geographical codifications are difficult to harmonize with recent ones.

The rest of the note is structured as follows. First, we describe the databases. Then, we propose an algorithm to link the censuses. Next, we present the results and some exercises of validation as well as measures of job flows by period. Finally, we discuss some caveats of our procedure, and we explain how to access the dataset through INEGI's Microdata Laboratory.

2 The Mexican Economic Census

2.1 Coverage

Our data source is the Economic Census. The temporal coverage is 1999, 2004, 2009 and 2014. Since the censuses collect data at the establishment level, the linkages presented here are also at the establishment level. We use all industries and all regions of Mexico. The number of establishments by census is shown in Table 1.

The number of units of the Economic Census increases mainly by the birth of new establishments, but also, by the expansion of the geographical coverage, namely, new localities surpassing 2,500 inhabitants.

[1]The CLEE (Clave Única de Identificación Estadística) was created by a combination of computer and human observation.

Table 1: Establishments

Census	Total
1999	2,804,984
2004	3,005,157
2009	3,724,019
2014	4,230,745

2.2 Variables

For all censuses, we have very detailed information that allow us to identify establishments; for instance, legal entity, establishment name, location codes, birth year and 6-digit industry. The complete list and codification is shown in Table 2.

Table 2: Variables

INEGI Code	Description
	Location
E03	State Code
E04	Municipality Code
E05	Locality Code
E06	Area (AGEB) Code[2]
E07	Block Code
E10	Street Name
E11	Exterior Number
E14	Neighborhood Name
	Entity
E01	Identification Number (NIC)
E02	Operative Number (NOP)
E08	Establishment Name
E09	Legal Entity/Owner Name
G111	Establishment Birth Year
	Industry
E17	6-Digit Industry

[2]AGEB stands for Basic Geo-Statistical Area.

Location codes from E03 to E07 are standardized codifications defined by INEGI. Variables E10, E11 and E14 are self-reported text strings. E08 is the name of the establishment, for instance, "Mini Market Maria". E09 is the name of the legal entity, as "Maria S.A. de C.V.". If the establishment does not belong to a legal entity, E09 reports the owner's name. E17 is the 6-digit sector according to the North American Industrial Classification System (NAICS).

Variables E01 (NIC) and E02 (NOP) are identifiers available for all censuses. They can be used to perform longitudinal linkages before 2009 only for some large establishments. For the majority of units, they can only be used as an identifier within the census and not for linkage.

3 Linkage

Our work consists of five steps, similar to the model presented in Christen (2012). The linkage covers steps I through IV. The fifth step, validation, will be discussed in Section 5. The steps are the following:

I Standardization: We substitute or eliminate special characters like accents and punctuation marks. Also, we standardize establishments descriptions like "Abarrotes" and "Tienda", which stand for the same kind of business. We also correct misspellings in legal entities like "SA CV" instead of "SA de CV". In general, we eliminate, standardize or substitute characters in all self-reported text strings with problems of misspelling or that can be reported in several ways.

II Indexing (pre-matching): We propose candidates for linkage. For instance, if two establishments have the same location in t and $t+5$, we compare the owner's name or establishment's name to decide if it is a good match.

III Comparison: We use different strategies to compare two indexed establishments. In general, we use *STATA* procedures that compare text strings.

IV Matching Classification: We assign an identifier to linked establishments. Then, we tag the linkages to exclude them in further phases (described below). Finally, we assign a number denoting the phase in which it was linked.

V Validation: We measure accuracy by applying the algorithm to the 2009 and 2014 censuses, which were already linked by INEGI (see Section 5).

To cover steps I to IV we define a 10-phase algorithm. All phases follow the continuity rule defined by OECD (2008). The rule considers three continuity factors: legal entity, industry and location. If a unit maintains at least two continuity factors from period t to $t+\Delta t$, that unit is considered the same.

We mainly use *STATA* to perform the 10 phases. In some phases we use the command *matchit*, written by Raffo (2017), which compares text strings and returns a similarity score between 0 and 1. We also use the command *soundex*, which returns a code consisting of the first letter of the text string followed by three digits assigned by *STATA*; these digits are the same for similar chains of consonants.

The phases are:

1. We link establishments with identical combination of Identification Number (NIC) and Operative Number (NOP)[3].

2. We link establishments with the same combination of State, Municipality, Locality, AGEB, Block and Industry.

3. We first index establishments with the same combination of State, Municipality, Locality, AGEB, Block and Building Number. Then we link establishments with similarity of at least 45% in Establishment Name and 75% in Legal Entity[4].

4. We first index establishments with the same combination of State, Municipality, Industry and Legal Entity. Then we link establishments with similarity of at least 30% in Establishment Name.

5. We link establishments with the same combination of State, Municipality, AGEB, and Legal Entity.

6. We first index establishments with the same combination of State, Municipality, Locality, AGEB, Block and Industry. Then we link establishments with the same *soundex* in Establishment Name and Legal Entity.

[3]Some NIC-NOP duplicates are present in 1999 (less than 400) and to a lesser extent in 2004 and 2009 (less than 100). For 2014 there are no duplicates.

[4]If Establishment Name or Legal Entity is empty or reports "SIN NOMBRE" (no name), that establishment is not considered.

7. We link establishments with the same combination of State, Municipality, Locality, AGEB, Block, Industry and Birth Year.

8. We link establishments with the same combination of State, Municipality, Locality, AGEB, Block, Industry and Exterior Number.

9. We first index establishments with the same combination of State, Municipality, Locality, and AGEB or Industry. Then we link establishments with similarity of at least 65% in Establishment Name and Legal Entity.

10. We link establishments with the same combination of Industry, Establishment Name and Legal Entity.

Whenever we match or index establishments according to a sequence of variables, we consider only those with unique combinations within a census. For instance, in phase 2, we match establishments reporting the same location and industry in t and $t+5$; however, if two establishments reported the same location and industry in t, it will not be clear which of them is the one that reappeared in $t+5$. To avoid ambiguities and matching errors, we exclude these cases and try in further phases to link them by using different combinations of variables.

The percentages of similarity required in some phases were determined so they can correctly predict around 90% of their matches. We can be more restrictive with the similarity scores, but the gains in accuracy would not compensate for losses of good linkages. We will further explain the accuracy of each phase in the validation section.

4 Results

After performing the 10 phases we obtain the results showed in Table 3. For any two adjacent censuses, t and $t+5$, we linked at least 50% of the establishments of year t.

Table 4 disaggregates total linkages by phase[5]. Phases 1 to 6 are by far the most important, accounting for least 86% of total matches for any period (and as we will see, they are also the most accurate).

[5]Phase 1 was not performed for 2009-2014. This is because the phase uses the variables "Identification Number (NIC)" and "Operative Number (NOP)", which are redundant with the CLEE. While for 1999-2004 and 2004-2009 NIC-NOP accounts for 7% of linkages, for 2009-2014 it would be 100%.

Table 3: Linked Establishments

Period	Total Establishments	Linked	%
1999 - 2004	2,804,984	1,444,584	51.5
2004 - 2009	3,005,157	1,522,578	50.7
2009 - 2014	3,724,019	2,154,410	57.9

Note: Total Establishments refers to the first year of the period.

Table 4: Percentage of Linkages by Phase

		Phase									
Period	Linkages	1	2	3	4	5	6	7	8	9	10
1999 - 2004	1,444,584	7.1	38.2	17.2	15.1	6.5	2.1	0.5	12.5	0.8	0.1
2004 - 2009	1,522,578	7.1	40.0	16.9	11.2	8.9	3.0	1.6	10.4	0.8	0.1
2009 - 2014	2,154,410	0.0	49.2	14.1	16.8	9.4	2.7	2.8	4.3	0.5	0.1

The linkage algorithm is performed on pairs of consecutive censuses; however, we can also follow an establishment across several editions. According to Table 5, we can see 8.6 million different establishments across all four censuses. Among these 8.6 million, 63.8% can only be seen in one census, 21% in two, 7.4% in three and 7.8% in four. This last 7.8% forms a balanced panel from 1999 to 2014 of 675 thousand establishments.

Table 5: Total Establishments by Appearance

Row	Years	Number of Years	Establishments	%
1	Only 2014	1	2,076,335	24.0
2	Only 2009	1	1,128,288	13.1
3	2009-2014	2	1,073,153	12.4
4	Only 2004	1	945,987	10.9
5	2004-2009	2	208,708	2.4
6	2004-2014	3	405,878	4.7
7	Only 1999	1	1,360,400	15.7
8	1999-2004	2	536,592	6.2
9	1999-2009	3	232,613	2.7
10	1999-2014	4	675,379	7.8
		Total	8,643,333	100.0

Establishments in Table 5 refer to those that appeared exclusively in the period indicated. However, if we want to know, for instance, how many

establishments appeared from 1999 to 2009, regardless if they reappeared in 2014, we must add the rows 9 and 10. If we want to know how many establishments appeared in both 2009 and 2014, regardless if they appeared before, we have to add rows 3, 6 and 10; and so on for the rest of possibilities.

5 Validation

The quality of the linkage depends on its completeness and accuracy; both can be evaluated by answering the following two questions:

(i) Completeness: How many establishments must be linked for any two consecutive censuses?

(ii) Accuracy: What is the probability that two linked establishments make a good match?

To answer both questions, we compare the linkage performed by INEGI through the CLEE and the one achieved by our algorithm.

5.1 Completeness of the Linkage

Table 6 shows that 58% of all establishments from 2009 can be linked with an establishment of 2014 by using the CLEE. Thus, we expect that any algorithm performing the linkage for the 2009 and 2014 censuses will achieve a similar percentage. As shown in Table 6, our algorithm linked 2,154,410 units, that is 57.9% of the 2009 Census. Total linkages of both methods are nearly the same, as our algorithm matched 99.7% of the amount linked by the CLEE.

Table 6: Linkages by Method 2009-2014

Method	Total Establishments (2009)	Linked	%
CLEE	3,724,019	2,159,804	58.0
Algorithm	3,724,019	2,154,410	57.9

The algorithm attains the expected number of linkages for 2009-2014; however, this does not completely answer question (i), we also need to estimate how many establishments we should link for the periods 1999-2004 and 2004-2009.

One way to answer this question is by using the self-reported birth year. If the establishment's age is 5 or greater in $t + 5$, it could potentially be observed in t. According to Table 7, in 2009, 1.8 of the 3.7 million establishments reported operations in 2004 or before (age equal or greater than 5). Since the 2004 Census captured 3 million establishments, we expect to link 59.74% of them with establishments of 2009. Similarly, 1.4 of the 3 million establishments of 2004 reported operations in 1999 or before. Since the 1999 Census captured 2.8 million, we expect to link 51.47% of them with establishments from 2004.

Table 7: Establishments by Age

Age	2004	2009	2014
Less than 5	1,561,466	1,928,674	1,806,638
5 or More	1,443,691	1,795,345	2,424,107
Total	3,005,157	3,724,019	4,230,745

However, there are a couple of reasons why the expected number of matches might be overestimated if we only use the self-reported age as reference. The first reason is because the Economic Censuses expand their geographical coverage with each edition. Some localities that were not covered before are now included as they surpassed 2,500 inhabitants[6]. Second, since the birth year of an establishment is self-reported, it might have some degree of misreporting. We can use the CLEE to measure how different is the number of linkages made by INEGI and the number of establishments that in 2014 reported operations the preceding census.

The 2014 Economic Census captured 4.2 million establishments, of which INEGI linked 2.2 million through the CLEE. At the same time, in 2014, 2.4 million reported operations in 2009 according to their age. In other words, INEGI linked only 89.1% of the establishments that reported operations in 2009 and 2014 according to their age. If we assume that this degree of discrepancy has been constant over time, we would expect to link around 89.1% of the establishments that in 2009 reported operations in 2004 according to their age (similarly for 2004 and 1999).

[6]From 2009 to 2014, urban localities went from 2,000 to 3,600; those 1,600 new urban localities added around 12,000 new establishments in 2014.

As Table 8 shows, 1.4 million establishments operated in 1999 and 2004, according to their age; however, considering the discrepancy discussed above, we should link around 1.3 million of them (89.1%). The actual number of linkages by the algorithm was 1.4 million, that is 12.3% more than expected, suggesting an overestimation of the survival rate of establishments. For the period 2004-2009 we linked 5% less than expected, implying a possible overestimation of establishment mortality. For 2009-2014 we linked almost the same amount as expected.

Table 8: Expected and Actual Linkages

Period	Age ≥ 5	Exp. Linkages	Actual Linkages	% of Exp.
1999-2004	1,443,691	1,286,329	1,444,584	112.3
2004-2009	1,795,345	1,599,652	1,522,578	95.2
2009-2014	2,424,107	2,159,879	2,154,410	99.7

Note: Expected Linkages are 89.1% of those that reported age 5 or greater.

5.2 Accuracy of the Linkage

Question (ii) can also be answered by using the 2009-2014 linkage defined by INEGI through the CLEE. As we mentioned before, the algorithm links almost the same number of establishments as does the CLEE. However, this does not mean that both methods match exactly the same establishments. Actually, we can have the following four types of match:

1. True Positive: Matched both with the algorithm and the CLEE.
2. False Positive: Matched with the algorithm but not with the CLEE.
3. True Negative: No matched with either the algorithm or the CLEE.
4. False Negative: No matched with the algorithm but matched with the CLEE.

Table 9 presents the share of establishments of 2009 and 2014 according to these four categories. The last column shows that 90% of the establishments in 2009 linked by the algorithm were also linked by the CLEE 89.8% for 2014). In general, we can interpret the percentage of True Positives as the algorithm's accuracy. It can also be read as the probability that two linked establishments are actually the same.

Table 9: Type of Match by Year

	Not Linked			Linked		
Year	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
2009	86.0	14.0	100	10.0	90.0	100
2014	89.1	10.9	100	10.2	89.8	100

On the other hand, False Positives were 10% for 2009 and 10.2% for 2014. The percentage of False Positives is the price that we have to pay to obtain a high amount of linkages by using the algorithm. One way to decrease this percentage is by increasing the restrictions of some of the 10 phases of the algorithm; for instance, requiring higher degrees of similarity for variables like the establishment name or legal entity. The drawback is that by doing so, we will also increase the percentage of False Negatives since many of establishments that we correctly predicted as matches will no longer be linked.

The other side of the coin when we talk about accuracy is the percentage of True and False Negatives, that is, those establishments that were not linked. Leaving out some establishments that should have been linked implies an overestimation of establishments exit. According to Table 9, 14% of the establishments not linked in 2009 were actually present in 2014. In absolute numbers, we are leaving out around 220 thousand establishments that we should have linked; but at the same time, we have almost the same number of False Positives, so total linkages of the Algorithm and the CLEE is nearly the same (see Table A1 in the Appendix).

Percentages shown in Table 9 are overall numbers that include all establishments from all sizes, industries and regions of Mexico. They also consider all phases of the algorithm, and not all of them are equally accurate. We can disaggregate these percentages to see if there is a systematic difference in accuracy given size, industry, state or phase.

Accuracy by Size

Tables 10 and 11 report that the accuracy of the algorithm increases with size (higher percentage of True Positives). This means that if the algorithm predicts that two large establishments are the same, it is almost certain that they are. The risk of mismatching is greater for smaller units. On the other hand, the percentage of False Negatives is increasing with size. This means

that the algorithm might overestimate the number of large establishments that leave the market. Note that overall percentages are always extremely close to those of small establishments (0-10 workers); this is because they account for almost 95% of all establishments.

Table 10: Type of Match by Size in 2009

	Not Linked			Linked		
Workers	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
0-10	86.3	13.7	100	10.1	89.9	100
11-50	79.6	20.4	100	7.6	92.4	100
50-100	69.4	30.6	100	5.3	94.7	100
$>$ 100	63.3	36.7	100	5.4	94.6	100
Total	86.0	14.0	100	10.0	90.0	100

Table 11: Type of Match by Size in 2014

	Not Linked			Linked		
Workers	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
0-10	89.4	10.6	100	10.4	89.6	100
11-50	83.8	16.2	100	8.2	91.8	100
50-100	74.8	25.2	100	5.8	94.2	100
$>$ 100	69.0	31.0	100	4.9	95.1	100
Total	89.1	10.9	100	10.2	89.8	100

Accuracy by Industry

Tables A2 and A4 in the Appendix show the types of match by industry. In 2009, the industry with the lowest percentage of True Positives is *55 Management*; however, it is made up of only 204 establishments so it has no impact on overall accuracy (see A3 and A5). Manufacturing, a commonly studied sector, is very accurate, with less than 10% False Positives. Industries 11, 21, 22, 23 and 55 show unusually high levels of False Negatives, which means that exit in those industries might be overestimated by the algorithm; however, they account for only 1.2% of total establishments, so their impact in overall accuracy is also limited. The remaining industries

show little variation in their percentages with respect to the overall.

Accuracy by State

Tables A6 and A7 in the Appendix show the types of match by state. Percentages of True Positives are not so dispersed by state, ranging from 87% to 92%. This means that the accuracy of the algorithm is almost the same for all regions. On the other hand, the percentage of False Negatives presents more dispersion. Particularly, the algorithm overestimates the exit of establishments from Mexico City (CDMX). This could happen because it is a densely populated region and identifying units becomes harder since very similar establishments are located in small areas, creating ambiguities that are difficult to resolve.

Accuracy by Phase

As we anticipated, not all phases of the algorithm have the same accuracy (percentage of True Positives). As shown in Table 12, accuracy is above 90% for phases 2 to 6 (accuracy of phase 1 is 100% but it was not performed for 2009-2014 because it is redundant with the CLEE). The last 4 phases have lower levels of True Positives but they account for 7.8% of all establishments so they have little impact in overall accuracy. Those phases were nevertheless included since they link establishments with great similarity in location, legal entity and industry, satisfying OECD requirements.

It is worth mentioning that our 10% of false positives is not a random linkage of establishments; that is, we will not link Walmart with a small grocery store. Even though the algorithm links some establishments that the CLEE does not, they share features related to location, legal entity and industry. Even if they are not the same in reality, it might not be problematic for statistical analysis since these establishments are extremely similar. Unfortunately, there is no way to know (at least through a computational procedure) which establishments of that 10% are indeed the same or not before 2009.

6 Job Flows

Another way to assess the quality of the algorithm is by estimating measures of job flows, entry and exit with the identifiers generated by the algorithm and compare them to those obtained by INEGI's CLEE. To compute these measures we follow Jarmin and Miranda (2002). They are defined as follows:

Table 12: Type of Match by Phase

Phase	False Pos.	True Pos.	Total	% True Pos.
1	0	0	0	-
2	85,527	973,927	1,059,454	91.9
3	17,673	287,094	304,767	94.2
4	29,680	331,617	361,297	91.8
5	16,996	186,513	203,509	91.6
6	5,290	51,866	57,156	90.7
7	6,358	54,913	61,271	89.6
8	45,834	47,076	92,910	50.7
9	6,819	4,376	11,195	39.1
10	777	2,074	2,851	72.7
Total	214,954	1,939,456	2,154,410	90.0

Job Creation and Job Destruction Rates:

$$JCR = \frac{JC}{X}$$

with:

$$JC = E_{t+5} - E_t$$

where E denotes the employment of expanding establishments and startups. And X is the average employment of t and $t+5$. The Job Destruction Rate (JDR) is computed analogously but E is the employment of contracting and shutting down establishments.

Establishment Entry and Exit Rates:

$$Entry\ Rate = \frac{ENTRY}{AVG}$$

where $ENTRY$ is the number of new establishments in $t+5$ and AVG is the average number of establishments between t and $t+5$. The Exit Rate is analogous, but we replace startups with shutting down establishments ($EXIT$).

$$Exit\ Rate = \frac{EXIT}{AVG}$$

The last two rows of Table 13 show that entry and exit rates are almost the same for the two methods of linkage. However, the algorithm slightly overestimates both job creation (JC) and job destruction (JD).

It is worth noticing that, for periods with equal Δt, for instance 1999-2004 and 2004-2009 ($\Delta t = 5$), or 1999-2009 and 2004-2014 ($\Delta t = 10$), the rates keep the same order of magnitude.

Table 13: Annual Rates of Entry, Exit, Job Creation and Job Destruction

Matching	Period	Entry	Exit	JCR	JDR
IDB's Algorithm	1999 - 2004	9.2	15.4	9.3	13.6
	1999 - 2009	6.6	9.1	6.4	8.4
	1999 - 2014	4.8	6.6	4.7	6.4
	2004 - 2009	11.6	12.7	11.3	11.8
	2004 - 2014	6.9	8.2	6.6	9.8
	2009 - 2014	9.4	11.2	11.0	13.5
INEGI's CLEE	2009 - 2014	9.4	11.1	9.8	12.7

7 Discussion

One caveat of our procedure is that we do not take into account the reorganization of units, like mergers or splits. If a large number of mergers occurred from t to $t+5$, we could be overestimating the exit of establishments. Unfortunately, there is too little information about this phenomenon to consider it in the linkage process.

Also, we do not consider establishments that were present in non-adjacent censuses. If an establishment was present in 1999 and reappeared in 2009 (due to inactivity or non-response in 2004), it will not be linked. This issue could lead to an overestimation of exit.

Another issue is that we do not use equivalence tables to harmonize recodification of 6-digit industries and location codes. Although the changes are minimal, we may be leaving out some establishments that should have been linked.

8 INEGI's Microdata Laboratory

The Economic Censuses at the establishment level are considered confidential information by INEGI. The only way to work with these data is to attend the facilities of the Microdata Laboratory located in Mexico City. If researchers are interested in using the identifiers described in this document, they have to make a special request to INEGI and request the *IDB identifiers* to be included in the databases.

References

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering 24*(9), 1537–1555.

Jarmin, R. S. and J. Miranda (2002). The longitudinal business database. *Center for Economic Studies, Working Paper*, 02–17.

OECD (2008). *Eurostat-OECD manual on business demography statistics*. Organisation for Economic Co-operation and Development.

Raffo, J. (2017). Matchit: Stata module to match two datasets based on similar text patterns.

A Appendix Tables

Table A1: Type of Match by Year

	Not Linked			Linked		
Year	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
2009	1,349,264	220,345	1,569,609	214,954	1,939,456	2,154,410
2014	1,850,319	226,016	2,076,335	220,625	1,933,785	2,154,410

Table A2: Type of Match by Industry, Shares in 2009

	Not Linked			Linked		
Industry	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
11 Agriculture	74.6	25.4	100	1.6	98.4	100
21 Mining	71.0	29.0	100	3.5	96.5	100
22 Utilities	54.0	46.0	100	0.3	99.7	100
23 Construction	86.8	13.2	100	4.9	95.1	100
31 Manufacturing	86.4	13.6	100	8.1	91.9	100
32 Manufacturing	87.9	12.1	100	9.6	90.4	100
33 Manufacturing	88.2	11.8	100	8.9	91.1	100
43 Wholesale	86.4	13.6	100	9.5	90.5	100
46 Retail	86.1	13.9	100	11.1	88.9	100
48 Transportation	87.8	12.2	100	11.7	88.3	100
49 Transportation	92.8	7.2	100	13.4	86.6	100
51 Information	89.8	10.2	100	10.6	89.4	100
52 Finance	92.3	7.7	100	14.7	85.3	100
53 Real Estate	93.9	6.1	100	7.6	92.4	100
54 Professional	91.6	8.4	100	13.1	86.9	100
55 Management	64.6	35.4	100	16.4	83.6	100
56 Support	92.7	7.3	100	14.8	85.2	100
61 Educational	92.0	8.0	100	7.6	92.4	100
62 Health Care	92.6	7.4	100	10.7	89.3	100
71 Entertainment	93.3	6.7	100	9.7	90.3	100
72 Food	74.4	25.6	100	4.0	96.0	100
81 Other	90.2	9.8	100	10.1	89.9	100
Total	86.0	14.0	100	10.0	90.0	100

Table A3: Type of Match by Industry in 2009

Industry	True Negatives	False Negatives	Total Not Matched	False Positives	True Positives	Total Matched	Total
11 Agriculture	5,994	2,040	8,034	184	11,225	11,409	19,443
21 Mining	1,057	431	1,488	51	1,418	1,469	2,957
22 Utilities	154	131	285	6	2,298	2,304	2,589
23 Construction	9,159	1,389	10,548	395	7,694	8,089	18,637
31 Manufacturing	77,899	12,234	90,133	11,819	133,410	145,229	235,362
32 Manufacturing	31,899	4,401	36,300	4,664	43,701	48,365	84,665
33 Manufacturing	44,322	5,949	50,271	5,945	60,608	66,553	116,824
43 Wholesale	44,323	6,953	51,276	6,363	60,389	66,752	118,028
46 Retail	587,281	94,567	681,848	117,966	940,708	1,058,674	1,740,522
48 Transportation	6,662	926	7,588	900	6,776	7,676	15,264
49 Transportation	1,891	147	2,038	54	349	403	2,441
51 Information	7,254	824	8,078	348	2,928	3,276	11,354
52 Finance	8,224	690	8,914	1,442	8,350	9,792	18,706
53 Real Estate	24,875	1,605	26,480	2,107	25,601	27,708	54,188
54 Professional	36,727	3,364	40,091	5,865	38,739	44,604	84,695
55 Management	53	29	82	20	102	122	204
56 Support	42,372	3,339	45,711	5,220	29,991	35,211	80,922
61 Educational	17,443	1,508	18,951	1,860	22,475	24,335	43,286
62 Health Care	48,133	3,873	52,006	10,084	84,442	94,526	146,532
71 Entertainment	20,765	1,501	22,266	1,893	17,662	19,555	41,821
72 Food	162,661	56,028	218,689	6,944	166,609	173,553	392,242
81 Other	170,116	18,416	188,532	30,824	273,981	304,805	493,337
Total	1,349,264	220,345	1,569,609	214,954	1,939,456	2,154,410	3,724,019

Table A4: Type of Match by Industry, Shares in 2014

	Not Linked			Linked		
Industry	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
11 Agriculture	75.9	24.1	100	1.4	98.6	100
21 Mining	72.3	27.7	100	3.6	96.4	100
22 Utilities	68.1	31.9	100	0.3	99.7	100
23 Construction	85.4	14.6	100	4.5	95.5	100
31 Manufacturing	91.3	8.7	100	8.5	91.5	100
32 Manufacturing	89.6	10.4	100	9.6	90.4	100
33 Manufacturing	91.7	8.3	100	9.3	90.7	100
43 Wholesale	89.8	10.2	100	10.1	89.9	100
46 Retail	88.6	11.4	100	11.3	88.7	100
48 Transportation	87.8	12.2	100	11.7	88.3	100
49 Transportation	85.8	14.2	100	9.6	90.4	100
51 Information	92.0	8.0	100	10.2	89.8	100
52 Finance	94.4	5.6	100	14.7	85.3	100
53 Real Estate	94.9	5.1	100	7.9	92.1	100
54 Professional	93.8	6.2	100	13.4	86.6	100
55 Management	83.3	16.7	100	24.2	75.8	100
56 Support	94.1	5.9	100	14.9	85.1	100
61 Educational	93.6	6.4	100	7.8	92.2	100
62 Health Care	94.3	5.7	100	11.1	88.9	100
71 Entertainment	94.5	5.5	100	9.7	90.3	100
72 Food	82.3	17.7	100	4.1	95.9	100
81 Other	92.2	7.8	100	10.6	89.4	100
Total	89.1	10.9	100	10.2	89.8	100

Table A5: Type of Match by Industry in 2014

Industry	True Negatives	False Negatives	Total Not Matched	False Positives	True Positives	Total Matched	Total
11 Agriculture	6,831	2,168	8,999	165	11,243	11,408	20,407
21 Mining	1,130	432	1,562	53	1,417	1,470	3,032
22 Utilities	284	133	417	8	2,296	2,304	2,721
23 Construction	7,667	1,307	8,974	361	7,728	8,089	17,063
31 Manufacturing	117,919	11,241	129,160	12,377	132,948	145,325	274,485
32 Manufacturing	36,285	4,219	40,504	4,591	43,265	47,856	88,360
33 Manufacturing	54,733	4,986	59,719	6,212	60,754	66,966	126,685
43 Wholesale	60,178	6,857	67,035	6,400	56,913	63,313	130,348
46 Retail	753,618	96,562	850,180	120,426	941,687	1,062,113	1,912,293
48 Transportation	7,721	1,073	8,794	903	6,791	7,694	16,488
49 Transportation	958	158	1,116	37	348	385	1,501
51 Information	5,532	482	6,014	339	2,985	3,324	9,338
52 Finance	13,207	788	13,995	1,438	8,328	9,766	23,761
53 Real Estate	33,251	1,793	35,044	2,182	25,589	27,771	62,815
54 Professional	42,038	2,770	44,808	5,969	38,477	44,446	89,254
55 Management	194	39	233	30	94	124	357
56 Support	53,551	3,340	56,891	5,190	29,530	34,720	91,611
61 Educational	21,629	1,473	23,102	1,866	21,914	23,780	46,882
62 Health Care	71,964	4,368	76,332	10,485	84,120	94,605	170,937
71 Entertainment	28,564	1,664	30,228	1,954	18,210	20,164	50,392
72 Food	270,039	57,902	327,941	7,192	166,315	173,507	501,448
81 Other	263,026	22,261	285,287	32,447	272,833	305,280	590,567
Total	1,850,319	226,016	2,076,335	220,625	1,933,785	2,154,410	4,230,745

Table A6: Type of Match by State in 2009

	Not Linked			Linked		
State	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
Aguascalientes	87.6	12.4	100	10.7	89.3	100
Baja California	87.5	12.5	100	12.8	87.2	100
Baja California Sur	88.3	11.7	100	9.5	90.5	100
Campeche	85.8	14.2	100	8.5	91.5	100
Coahuila	91.5	8.5	100	10.0	90.0	100
Colima	85.9	14.1	100	8.5	91.5	100
Chiapas	84.5	15.5	100	11.8	88.2	100
Chihuahua	89.1	10.9	100	8.8	91.2	100
CDMX	79.6	20.4	100	10.7	89.3	100
Durango	88.7	11.3	100	7.6	92.4	100
Guanajuato	84.9	15.1	100	10.2	89.8	100
Guerrero	86.4	13.6	100	11.5	88.5	100
Hidalgo	86.4	13.6	100	10.0	90.0	100
Jalisco	86.4	13.6	100	9.8	90.2	100
México	85.1	14.9	100	10.2	89.8	100
Michoacán	87.4	12.6	100	10.3	89.7	100
Morelos	88.2	11.8	100	10.8	89.2	100
Nayarit	84.7	15.3	100	7.1	92.9	100
Nuevo León	90.1	9.9	100	10.4	89.6	100
Oaxaca	84.4	15.6	100	9.2	90.8	100
Puebla	85.6	14.4	100	10.2	89.8	100
Querétaro	85.2	14.8	100	10.5	89.5	100
Quintana Roo	86.2	13.8	100	11.7	88.3	100
San Luis Potosí	88.3	11.7	100	9.1	90.9	100
Sinaloa	87.0	13.0	100	7.5	92.5	100
Sonora	88.0	12.0	100	8.1	91.9	100
Tabasco	85.4	14.6	100	11.9	88.1	100
Tamaulipas	89.5	10.5	100	9.0	91.0	100
Tlaxcala	87.1	12.9	100	10.6	89.4	100
Veracruz	87.0	13.0	100	9.3	90.7	100
Yucatán	84.4	15.6	100	7.9	92.1	100
Zacatecas	89.8	10.2	100	7.8	92.2	100
Total	86.0	14.0	100	10.0	90.0	100

Table A7: Type of Match by State in 2014

	Not Linked			Linked		
State	True Neg.	False Neg.	Total	False Pos.	True Pos.	Total
Aguascalientes	90.7	9.3	100	10.9	89.1	100
Baja California	90.9	9.1	100	13.1	86.9	100
Baja California Sur	91.9	8.1	100	9.7	90.3	100
Campeche	88.0	12.0	100	8.6	91.4	100
Coahuila	92.0	8.0	100	10.2	89.8	100
Colima	88.6	11.4	100	8.8	91.2	100
Chiapas	89.0	11.0	100	12.4	87.6	100
Chihuahua	90.7	9.3	100	8.9	91.1	100
CDMX	82.7	17.3	100	11.1	88.9	100
Durango	91.0	9.0	100	7.9	92.1	100
Guanajuato	90.2	9.8	100	10.8	89.2	100
Guerrero	88.1	11.9	100	11.5	88.5	100
Hidalgo	90.6	9.4	100	10.4	89.6	100
Jalisco	90.4	9.6	100	10.1	89.9	100
México	89.0	11.0	100	10.6	89.4	100
Michoacán	90.0	10.0	100	10.4	89.6	100
Morelos	89.8	10.2	100	10.6	89.4	100
Nayarit	89.4	10.6	100	7.5	92.5	100
Nuevo León	91.2	8.8	100	10.2	89.8	100
Oaxaca	89.8	10.2	100	9.6	90.4	100
Puebla	89.5	10.5	100	10.6	89.4	100
Querétaro	90.2	9.8	100	10.8	89.2	100
Quintana Roo	89.5	10.5	100	11.7	88.3	100
San Luis Potosí	90.8	9.2	100	9.2	90.8	100
Sinaloa	90.5	9.5	100	7.7	92.3	100
Sonora	89.8	10.2	100	8.4	91.6	100
Tabasco	88.5	11.5	100	12.2	87.8	100
Tamaulipas	90.2	9.8	100	8.8	91.2	100
Tlaxcala	90.6	9.4	100	11.0	89.0	100
Veracruz	88.4	11.6	100	9.5	90.5	100
Yucatán	88.4	11.6	100	8.1	91.9	100
Zacatecas	91.3	8.7	100	8.0	92.0	100
Total	89.1	10.9	100	10.2	89.8	100