Using Contacting Information to Derive Employer Name in the Survey of Doctorate Recipients

Quentin Brummet[1], Karen Grigorian[2], Carlann Unger[3]
Wan-Ying Chang[4]
[1]NORC at the University of Chicago, 55 E. Monroe St., Chicago, IL 60603
[2]NORC at the University of Chicago, 55 E. Monroe St., Chicago, IL 60603
[3]NORC at the University of Chicago, 55 E. Monroe St., Chicago, IL 60603
[4]National Center for Science and Engineering Statistics, 2415 Eisenhower Ave., Alexandria, VA 22314

Abstract

We demonstrate a new use of contacting information to derive employer name and employer characteristics in the Survey of Doctorate Recipients. A combination of external data sources on email domains and manual coding procedures was used to assign employer names to email address, work mailing address, and work phone numbers for a random sample of respondents. Our results show significant promise: using email addresses, employer names were coded for 77% of respondents, and 70% of these respondents have a coded employer that aligns with their survey reports. We then develop a least absolute shrinkage and selection operator (LASSO) model to predict the best contact information to use, which we show fits the data well and assists with selecting the most accurate pieces of information. We conclude with a discussion of setting an optimal error rate threshold that allows the model to be operationalized in future SDR operations.

Key Words: Alternative Data Sources, Contacting Information, Predictive Modeling

1. Introduction

Alternative data sources show significant promise in many areas of survey operations, including frame development, weight construction, item imputation, and the enhancement of final survey data (Kreuter, 2013; Stoop et al, 2010). As one example of the promise of these new data sources, we describe the use of contacting information to derive employer name when missing in the Survey of Doctorate Recipients (SDR) and discusses the statistical and practical challenges associated with using this information.

While employer name is not released by the SDR in order to protect survey respondents' confidentiality, it is used to derive many important variables such as Carnegie Class of academic institutions. Therefore, being able to derive employer name from other information captured in the survey would serve to increase the analytical utility of SDR data. Moreover, employer name is not currently collected in an abbreviated version of the survey questionnaire named the "Critical Item Only" (CIO) version. Since these CIO questionnaires account for roughly 10% of completed surveys in recent rounds of the SDR, obtaining an estimate of employer name from an alternative data source such as contacting information could significantly reduce item non-response for the SDR.

To assess the utility of the available contacting information, we start by taking a sample of respondents from the 2015 SDR who reported employer name and, independent of that reported employer information, attempt to derive a coded employer name from the contacting information. Using a combination of external data sources and manual coding procedures, we assign potential employer names based on email domains, work addresses, and work telephones. The results of this process show that we can successfully code an email domain name for the vast majority of respondents. The new coded employer name information also aligns well with self-reported data. Email domains and work addresses align particularly well, while work phone numbers align at lower rates. The process works better for academic respondents, for whom employer names are easier to code and, once coded, are more likely to align with respondent reported data.

Given that this process produces multiple potential employers for a given survey respondent, we then use a machine learning model to predict the likelihood of deriving an accurate employer name from the contacting information. The model fits the data well and can be tailored in a production setting to balance the trade-off between coding as many employer names as possible while maintaining sufficient data accuracy.

In total, these results show promise for using contacting information to derive employer name for SDR respondents in the academic sector. For respondents in the private and government sector, more work is likely needed in order to ensure that contacting information could be used to accurately derive an employer. Out of the pieces of contacting information that we investigated, work telephones performed the worst and may not be worth the cost of coding if this work were undertaken in a production setting. We recommend future research to expand on these results by including information on the cost of coding contacting information in the model or by considering the potential of coding multiple pieces of contacting information per respondent in a production setting.

The remainder of this paper is structured as follows. Section 2 discusses our methodology for coding employer names and developing a model to differentiate between correct and incorrect employers. Section 3 then presents our results, while Section 4 concludes and discusses future research.

2. Methodology

Our study consists of four steps:

1. Select an experiment sample from the 2015 SDR respondents who reported employer name;
2. Assign potential employer names to our sample from contacting information using external data sources and manual coding procedures;
3. Analyze the success of the employer assignments by comparing to existing SDR employee name assignments and employer characteristics from respondent reports;
4. Develop a LASSO model to predict the most accurate coded employer name, given that this process may produce multiple potential employee names for a survey respondent.

2.1 Experiment Sample

We drew a sample of 5,000 cases from the 2015 SDR sample, restricted to cases that completed the full survey and provided a non-missing employer name. This leaves 60,974 (77.9 percent) out of the total sample size of 78,320 respondents eligible for sampling. We utilized systematic sampling to have a sample representative by key variables. The eligible set was sorted on the following variables prior to selection, listed in order:

1. Respondent location on the survey reference date (U.S. or non-U.S.)
2. Employment Sector (Academic, Government, or Non-Academic Private Sector)
3. 8-level field of doctorate degree
4. Years since degree
5. Employer size (Using the following categories: 99 or fewer employees, 100-499 employees, and 500 or more employees)
6. Indicator for whether locating was conducted

2.2 Employer Name Assignment

To assign an employer name, we start with SDR respondent emails from questionnaires and our Case Management System (CMS). Table 1 shows the distribution of email addresses per respondent. In the sample we drew, 84.8 percent of respondents reported at least one email address in the questionnaire, 85.7 percent have at least one email address in the CMS, and 97.2 percent of respondents have either a questionnaire or a CMS email address.

We first used two email domain lookup tables that provide information on employer name for educational institutions and government agencies. The lookup table for educational institutions is taken from an open source table posted on GitHub, and the government agencies is taken from a list of .gov domains maintained by the General Services Administration.[1] For the cases that could not be found using the lookup tables, we conducted a clerical operation that, where possible, assigned an employer name to a given domain. If an email address was clearly personal, we did not attempt to code employer name. For example, if the email address ended in “pg.com”, we coded the employer name as “Proctor and Gamble”. If the email address ended in a generic domain such as “yahoo.com” or “gmail.com”, we noted that it was a portable email address and did not attempt to code an employer name. Table 2 documents the process by which we coded email addresses and shows how many were found in the databases or sent to a clerical review. In the context of our 5,000 case sample, we extracted 1,863 unique email domains from questionnaire responses and 2,883 unique email domains from the CMS. A number of these were coded automatically using these lookup tables, but the majority were sent to the clerical operation. Combining unique domains from both questionnaire and CMS email addresses, we sent 2,573 (75.5 percent) of the original 3,405 email domains to clerical review.

[1] The educational institution table can be found at https://github.com/Hipo/university-domains-list, and the .gov lookup table is available at https://home.dotgov.gov/data/#all-gov-domains. We conducted a small clerical audit of these tables to verify that we believed them to be high quality.

Note that this clerical process allows us to build our own lookup table for email addresses that have been recorded in the SDR. Therefore, this information can be used in the future to conduct automated coding of employer name using email address domains.

In addition to coding email addresses, we extracted physical work addresses and phone numbers from both questionnaires and the CMS in order to perform address- and phone-based employer name lookups. We only attempted to code primary work addresses, and coded at most one questionnaire and one CMS address per respondent. This clerical operation was roughly three times more efficient than the clerical operation for email domains, as an employer can often be coded directly from the contacting information (e.g., "Harvard University" is included in the address). Nonetheless, our coding operation for work addresses and phone number was entirely clerical, so it was on the whole more resource intensive than coding email domains.

2.3 Employer Name Alignment Analysis

After coding contacting information, we compared the coded employer name to the employer name reported by the respondent in the 2015 SDR. This analysis allows us to understand the reliability of using the different types of contacting information to derive employer name. Note that there are reasons for the coded and reported employers to differ other than errors in coding. First, email or work addresses may correspond to the respondent's employer in different time periods either before or after the survey reference date. In addition, respondents may have multiple email addresses reported, and only one relates to their current employer.

This accuracy assessment requires determining whether employer names between two different string variables representing the same employer. In order to account for the fact that names of employers may be written differently in the questionnaire than in the coding operation, we use a Jaro-Winkler string comparator to compare the two strings. This string comparator produces a score ranging from 0 (no match) to 1 (perfect match). Based on this score, we divide up the results into three groups: 1) definite matches, 2) definite non-matches, and 3) undetermined. For the undetermined cases, we ran a brief clerical review. Note that currently this procedure does not utilize a catalog of acronyms and government agency relationships, and therefore we consider the alignment results presented in Section 3.2 to be conservative. Nonetheless, as with the previous clerical operations, this review provides us with information that will allow us to more efficiently assign employer names in the future.

In addition to reviewing the success of correctly coding employers, we also analyzed the success of using derived employers to code employer characteristics by comparing IPEDS for matched academic employers. This process used standard IPEDS coding process for the SDR, which typically attaches characteristics of postsecondary institutions based on the institution name. Note that this process can only be applied to academic employers. If the SDR decides to use derived employer name for non-academic employers in future operations, it will require new alternative data on firm characteristics.

2.4 LASSO Model to Predict Correct Employer Name

For the experiment sample, we coded all email domains for both CMS and questionnaire emails as well as any potential work addresses and phone numbers. Of these 5,000 respondents, 3,286 respondents have at least one piece of contacting information that links to their current employer. In a survey production setting there is no way of telling which of these pieces of contacting information actually pertain to the correct employer. Therefore, we developed a model to distinguish which contacting information should be coded in order to provide correct employer information. Given the information that would be observed in a survey production setting, the model chooses a single piece of contacting information to send to a coding operation in order to maximize the chance that we code the correct employer.

We run a LASSO model where the dependent variable takes a value of '1' if the piece of contacting information can be correctly coded, and '0' otherwise. The model includes the following predictors from the 2015 survey frame:

- Years since PhD
- Age
- Race
- Field of Degree
- Sex
- US Citizenship
- Indicator for Completing Prior Wave of Survey
- Indicator for Whether Locating was Conducted
- Disabilities Indicator
- Postdoctoral Status

In addition, the model includes predictors that are characteristics of the contacting information:

1. For emails only, a set of indicators for domain extension (.com, .org, .net, etc…)
2. Type of contacting information (academic, government, business, etc…)
3. Source of contacting information (locating, questionnaire, etc.)

3. Results

3.1 Employer Name Alignment across Pieces of Contact Information

We begin by presenting the coding rates by type of contacting information. Table 3 summarizes the source and success of coding at the respondent level. The vast majority of our coded email domains came from the educational data base and the clerical operation. 44.60 percent of respondents had an email coded using the educational data base and 42.88 percent had an email coded through the clerical operation. Taking all sources together, 77.10 percent of respondents had at least one email address coded.

Address and phone coding was similar, but slightly less successful: 62.60 percent of the 5,000 respondents had at least one work address coded, and 41.10 percent had a work phone number coded successfully. Partially, these lower rates reflect the fact that we only coded primary work addresses that were most likely to reflect the current employer.[2]

[2] Only 68.56 percent and 52.22 percent of respondents had a primary work address or phone number in the 2015 SDR, respectively.

Table 4 shows the fraction of respondents for which we were able to assign an employer name to email addresses broken apart by respondent characteristics. Of the 5,000 respondents, 77.1 percent had at least one questionnaire or CMS email address coded. Importantly, this is higher for respondents working in academia, for whom over 90 percent could have at least one email address coded. This is particularly important, as these email domains are less time intensive to code given the availability of lookup tables. Note also that CMS email domains tend to be slightly easier to code, particularly for individuals for whom locating was conducted.

Moving to our address analysis, Table 5 presents statistics on the success of coding employer addresses. Recall that we only code at most one questionnaire and one CMS address per respondent, so this table can be interpreted as a respondent-level analysis. Overall, we see similar patterns to the email coding results presented in Table 4. 62.6 percent of respondents have a work address coded to an employer. This figure goes up to 77.7 percent for individuals working in academia. Again, CMS addresses tend to be coded at higher rates, particularly when locating was conducted that might provide us with more up to date contacting information.

Table 6 presents statistics on the success of coding employer phone number. Overall, 41.1 percent of respondents had a phone number coded. This figure goes up 53.8 percent of those working academia.

3.2 Coded Employer Name Alignment across Pieces of Contacting Information

Table 7 presents statistics on the alignment of employer names coded from email domains and addresses at a respondent level. Almost 70 percent of respondents with a successfully coded email have at least one correct employer from either survey or CMS email domain coding.[3]

Table 8 shows the alignment of employer name coded from addresses with respondent-reported employer name. All fractions reported refer to the fraction of coded employers that correctly matched the respondent-reported value. Overall, the addresses are fairly accurate, and the alignment rates are even higher than those for employer names coded from email domains. 83.1 percent of all respondents with a coded employer name have an employer name that agrees with what they reported. Individuals working in academia have particularly accurate coded names, with employer names aligning roughly 87 percent of the time.

Table 9 summarizes the alignment of employer name coded from phone numbers with respondent-reported employer name. Overall, 69.0 percent of the coded names are accurate. This is not as high quality as email domains or addresses, but still provides valuable information. Of respondent characteristics, being in the academic sector is one of the main predictors of successful coding. Phone numbers obtained from locating are also easier to code and match to the true employer at higher rates.

[3] While questionnaire emails were less likely to be coded than CMS emails, they tend to be more accurate conditional on being coded: 76.9 percent of respondents with a coded questionnaire email have at least one correct employer. Also it should be noted that the vast majority of our inability to code an email address is driven by respondents reporting portable email addresses.

3.3 Comparison to IPEDS for Matched Employer Results

Table 10 shows the results of our comparison to IPEDS. In general, match rates for Carnegie Class are similar to match rates based on employer name. Public/private matched at higher rates, but this is unsurprising given that this variable contains fewer categories.

3.4 Model Results and Potential Uses in Future Survey Production

To assess the predictive power of our LASSO model, we randomly split our sample of 5,000 respondents so that 60 percent of respondents fall in a "training" sample used to fit the model and 40 percent of respondents fall in a "test" sample. All results below are calculated from the test sample, meaning they measure out-of-sample performance of the model.

We found that characteristics of the contacting information itself are most important in determining whether a piece of contacting information should be coded. In particular, the type of email or address is extremely important as is the source of the information for information derived from the CMS. For the most part, frame characteristics of the respondents are less important, particularly for demographics such as age, race/ethnicity, and sex. Time since degree is the most important of the frame variables, likely reflecting the individuals who are more established in their careers are more likely to have stable contacting information attached to employers.

Figure 1 shows the distribution of predicted scores arising from the model. There are two large humps corresponding to pieces of contacting information that are clearly not worth coding to an employer, and pieces of contacting information that may be of value. In order to make the best use of this information, we must now determine where would be an appropriate cut point on this distribution to decide that the contacting information was potentially useful.

We present five potential uses of this model to identify contacting information to be coded. The "Ideal" scenario would be knowing beforehand whether a piece of contacting information would lead to the correct employer or not. If this were the case, we would accurately decide to code contacting information for the 1,264 respondents (reflected in the blue bar) in our test sample for whom we had contacting information leading to an employer, and we would not code the remainder since they would lead to incorrect employer information. However, this is clearly infeasible since we do not observe the truth. Instead, we consider five scenarios for using the model described above to determine which piece of contacting information to code:

1. For each respondent, code the piece of contacting information with the highest predicted probability of matching the current employer, regardless of how high that predicted probability is.
2. For each respondent, code the piece of contacting information with the highest predicted probability, provided the predicted probability of a correct employer is above ~39.4 percent. This number is chosen based on maximizing the product of sensitivity and specificity, following the suggestion of Liu (2012).
3. For each respondent, code the piece of contacting information with the highest predicted probability, provided the predicted probability of a correct employer is at least 70 percent.

4. For each respondent, code the piece of contacting information with the highest predicted probability, provided the predicted probability of a correct employer is at least 80 percent.
5. For each respondent, code the piece of contacting information with the highest predicted probability, provided the predicted probability of a correct employer is at least 90 percent.

Moving from (1) to (5), the procedure becomes more selective with which piece of contacting information should be coded. The more selective it becomes, the less chance of making a mistake and coding the incorrect employer. However, a more selective procedure will code employer name for fewer respondents, so we must make a decision to balance this tradeoff.

Figure 2 shows the results under each of these five scenarios and the "Ideal scenario". The grey bars show respondents who do not have contacting information coded to an employer, the blue bars show respondents who have contacting information coded to the correct employer, and the red bars show respondents who have contacting information coded to the incorrect employer. When we do not have a minimum threshold for choosing contacting information to code in scenario (1), we code many pieces of contacting information that lead to incorrect employers. As we get to the relatively selective cutoffs in scenarios (4) and (5), we are coding relatively less information, but are making very few mistakes: with the most selective cutoff in (5), we only make mistakes for 1.2 percent of respondents (24 respondents).

4. Conclusion

Making full use of data collected in the course of survey operations (such as contacting information) requires overcoming a number of practical challenges. In this paper, we show that contacting information may provide a valuable research for creating employer information in the SDR. We are able to successfully code the vast majority of academic, government, and business email address domains with employer names. Especially promising is the fact that academic emails for questionnaire domains can be coded near 100 percent of the time with relatively little effort and are correct at high rates. Our process is also able to successfully code addresses at very high rates. This is particularly true for academic addresses, for which we are able to code near 100 percent of addresses, and they are correct roughly 85 percent of the time. While we are less successful at coding work phone numbers, we still find that they provide useful information.

We then develop a model to distinguish whether pieces of contacting information would be useful to code in a future production setting. We find that our model performs well, and discuss five different scenarios where the model could be used depending on the level of accuracy desired by the SDR. Deciding on the appropriate level of accuracy is a policy decision that is left for further discussion and research.

We envision at least two future pathways to build on this research. First, the LASSO model does currently not take into account relative cost effectiveness of coding. Coding of email addresses from .edu and .gov sources is relatively costless given the availability of databases. If a piece of contacting information goes to a clerical review, we have found that survey assistance can code ~40 email domains an hour or ~100 addresses an hour. In addition, the current approaches discussed here select a single piece of contacting information from a given respondent to code. It would also be possible to code multiple pieces of contacting information for the same respondent. This would be more resource intensive in production and would require a more complicated modeling approach, but may serve to increase the utility of the contact data.

Acknowledgements

We are grateful to Daniel Foley at NCSES for helpful comments and suggestions. This work also benefitted from the valuable suggestions and contributions of Shana Brown, Zachary Gebhardt, Don Jang, Adam Ryckaert, Zachary Seeskin, Lance Selfa, and Christopher Wong at NORC at the University of Chicago. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the National Center for Science and Engineering Statistics, or NORC at the University of Chicago.

References

X. Liu, “Classification accuracy and cut point selection,” *Statistics in Medicine*, vol. 31, no. 23, pp. 2676–2686, 2012.

Jaro, M. A., "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida", *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–20, 1989.

Kreuter, F. Ed. *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, New Jersey: Wiley Series in Survey Methodology, 2013.

Stoop, I. A.L., J. Billiet, A. Koch, and R. Fitzgerald. *Improving Survey Response: Lessons Leared from the European Social Survey*. Chichester, West Sussex, PO19 8SQ, United Kingdom: Wiley, 2010.

Winkler, W. E., "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage" (PDF), *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pg. 354–359, 1990.

Table 1: Frequency of Email Addresses per Respondent by Source

Number of Email Addresses per Respondent	Questionnaire	CMS
0	760	717
1	2,666	2,390
2	1,574	1,240
3	-	367
4	-	149
5	-	88
6	-	27
7	-	18
8	-	4

Table 2: Email Domain Coding Results

Email Domain Coding Result	Questionnaire	CMS	Either Questionnaire or CMS
Number of unique email domains	1,863	2,883	3,405
Number found in educational database	626	727	793
Number found in government database	24	36	39
Number sent to clerical operation			2,573

Table 3: Overview of Coding Success

Coding Resource	Email	Address	Phone
Educational data base	44.60%	-	-
Government data base	3.64%	-	-
Clerical operation	42.88%	62.60%	41.10%
All	77.10%	62.60%	41.10%

Table 4: Respondent-Level Email Coding Success

Key Variables	Cases	Fraction with At Least 1 Questionnaire Email Coded	Fraction with At Least 1 CMS Email Coded	Fraction with Either Questionnaire or CMS Email Coded
Overall	5,000	0.537	0.650	0.771
Location				
US	4,434	0.528	0.652	0.762
Non-US	566	0.610	0.634	0.841
Sector				
Academic	2,580	0.695	0.804	0.905
Non-Academic Private Sector	1,901	0.332	0.391	0.588
Government	519	0.499	0.397	0.776
Field of Degree				
Computer and Information Sciences	139	0.432	0.635	0.763
Mathematics and Statistics	245	0.608	0.718	0.816
Biological, Agricultural, and Environmental Life Sciences	1,355	0.540	0.647	0.789
Health	235	0.579	0.683	0.791
Physical Sciences	955	0.554	0.609	0.766
Social Sciences	638	0.613	0.723	0.837
Psychology	565	0.487	0.642	0.696
Engineering	868	0.475	0.610	0.732
Employer Size				
99 or fewer	711	0.321	0.423	0.547
100-499	468	0.541	0.586	0.763
500 or more	3,821	0.577	0.683	0.814
Locating				
Locating was not conducted	1,426	0.497	0.487	0.731
Locating was conducted	3,574	0.553	0.704	0.787

Table 5: Respondent-Level Address Coding Success

Key Variable	Surveys	CMS Data	Fraction with Survey Address Coded	Fraction with CMS Address Coded	Fraction with Survey or CMS Address Coded *
Overall	5,000		0.427	0.459	0.626
Location (survey only)					
US	4,434	-	0.408	0.426	0.599
Non-US	566	-	0.569	0.716	0.837
Sector (survey only)					
Academic	2,580	-	0.555	0.595	0.777
Non-Academic Private Sector	1,901	-	0.263	0.292	0.436
Government	519	-	0.387	0.393	0.574
Field of Degree					
Computer and Information Sciences	139	139	0.324	0.396	0.532
Mathematics and Statistics	245	245	0.514	0.551	0.702
Biological, Agricultural, and Environmental Life Sciences	1,355	1,355	0.43	0.453	0.627
Health	235	235	0.421	0.485	0.651
Physical Sciences	955	955	0.46	0.443	0.629
Social Sciences	638	638	0.464	0.541	0.697
Psychology	565	565	0.412	0.391	0.591
Engineering	868	868	0.359	0.447	0.578
Employer Size (survey only)					
99 or fewer	711	-	0.309	0.309	0.488
100-499	468	-	0.429	0.464	0.637
500 or more	3,821	-	0.448	0.486	0.65
Locating					
Locating was not conducted	1,426	1,426	0.374	0.25	0.457
Locating was conducted	3,574	3,574	0.448	0.543	0.693

Table 6: Respondent-Level Phone Coding Success

Key Variable	N	Fraction Coded
Overall	5,000	0.411
Location		
US	4,434	0.394
Non-US	566	0.539
Sector		
Academic	2,580	0.538
Non-Academic Private Sector	1,901	0.239
Government	519	0.403
Field of Degree		
Computer and Information Sciences	139	0.281
Mathematics and Statistics	245	0.453
Biological, Agricultural, and Environmental Life Sciences	1,355	0.41
Health	235	0.472
Physical Sciences	955	0.382
Social Sciences	638	0.473
Psychology	565	0.457
Engineering	868	0.359
Employer Size		
99 or fewer	711	0.319
100-499	468	0.395
500 or more	3,821	0.429
Locating		
Locating was not conducted	1,426	0.196
Locating was conducted	3,574	0.496

Table 7: Respondent-Level Alignment from Coded Email Domains

Key Variable	Cases (All)	Fraction with At Least 1 Correct Coded Employer Name (Survey)	Fraction with At Least 1 Correct Coded Employer Name (CMS)	Fraction with At Least 1 Correct Coded Employer Name (All)
Overall	3,855	0.769	0.650	0.699
Location				
US	3,379	0.777	0.652	0.703
Non-US	476	0.716	0.634	0.672
Sector				
Academic	2,335	0.883	0.804	0.854
Non-Academic Private Sector	1,117	0.581	0.391	0.469
Government	403	0.440	0.397	0.434
Field of Degree				
Computer and Information Sciences	106	0.717	0.635	0.679
Mathematics and Statistics	200	0.812	0.718	0.760
Biological, Agricultural, and Environmental Life Sciences	1,069	0.788	0.647	0.695
Health	186	0.794	0.683	0.72
Physical Sciences	732	0.730	0.609	0.676
Social Sciences	534	0.803	0.723	0.758
Psychology	393	0.760	0.642	0.697
Engineering	635	0.745	0.610	0.66
Employer Size				
99 or fewer	389	0.570	0.423	0.478
100-499	357	0.692	0.586	0.636
500 or more	3,109	0.799	0.683	0.734
Locating				
Locating was not conducted	1,042	0.736	0.487	0.607
Locating was conducted	2,813	0.781	0.704	0.733

Table 8: Respondent-Level Alignment from Coded Addresses

Key Variable	Cases (All)	Fraction with At Least 1 Correct Coded Employer Name (Survey)	Fraction with At Least 1 Correct Coded Employer Name (CMS)	Fraction with At Least 1 Correct Coded Employer Name (All)
Overall	3,130	0.837	0.801	0.831
Location				
US	2,656	0.853	0.821	0.845
Non-US	474	0.748	0.709	0.755
Sector				
Academic	2,004	0.872	0.85	0.878
Non-Academic Private Sector	828	0.788	0.721	0.757
Government	298	0.716	0.652	0.725
Field of Degree				
Computer and Information Sciences	74	0.8	0.836	0.811
Mathematics and Statistics	172	0.817	0.83	0.849
Biological, Agri., and Environmental Life Sciences	849	0.82	0.793	0.826
Health	153	0.859	0.798	0.843
Physical Sciences	601	0.838	0.825	0.839
Social Sciences	445	0.878	0.814	0.863
Psychology	334	0.798	0.751	0.787
Engineering	502	0.865	0.789	0.827
Employer Size				
99 or fewer	347	0.755	0.664	0.72
100-499	298	0.801	0.774	0.792
500 or more	2,485	0.852	0.82	0.852
Locating				
Locating was not conducted	652	0.867	0.764	0.85
Locating was conducted	2,478	0.828	0.808	0.826

Table 9: Respondent-Level Alignment from Coded Phone Numbers

Key Variable	Cases	Fraction Correct Coded Employer Name
Overall	2,053	0.690
Location		
US	1,748	0.713
Non-US	305	0.554
Sector		
Academic	1,389	0.759
Non-Academic Private Sector	455	0.626
Government	209	0.368
Field of Degree		
Computer and Information Sciences	39	0.821
Mathematics and Statistics	111	0.712
Biological, Agricultural, and Environmental Life Sciences	555	0.699
Health	111	0.604
Physical Sciences	365	0.701
Social Sciences	302	0.682
Psychology	258	0.647
Engineering	312	0.708
Employer Size		
99 or fewer	227	0.608
100-499	185	0.697
500 or more	1,641	0.700
Locating		
Locating was not conducted	279	0.659
Locating was conducted	1,774	0.694

Table 10: Comparison with IPEDS Characteristics

Coding Source	Cases	Carnegie Class Matches	Public/Private Matches
Questionnaire Emails			
Questionnaire Emails Overall	1,513	0.709	0.770
CMS Emails			
CMS Emails Overall	2,303	0.608	0.685
Questionnaire Addresses			
Addresses Overall	1,113	0.774	0.828
CMS Addresses			
CMS Addresses Overall	1,140	0.76	0.819
Phones			
Phones Overall	1,393	0.680	0.744

Figures

Figure 1: Distribution of Predicted Probabilities

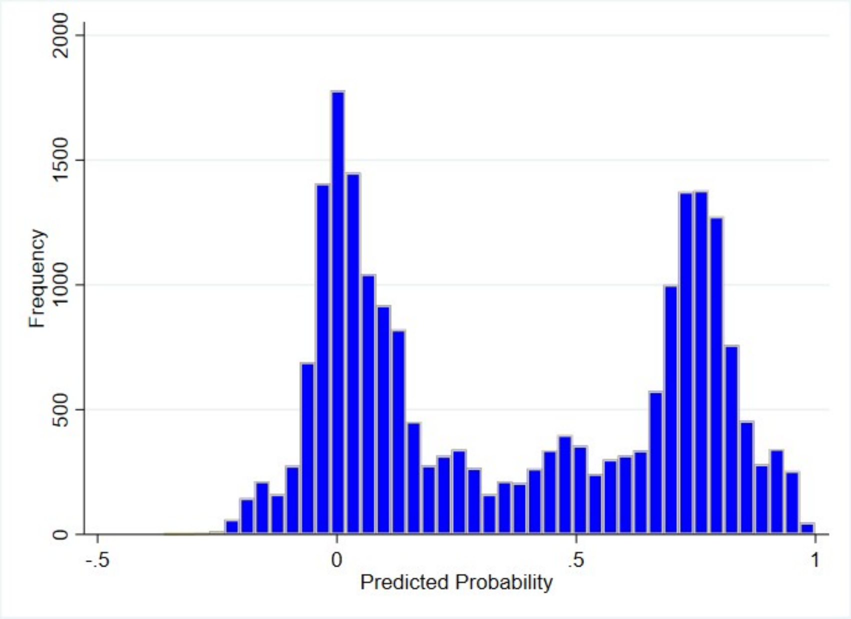
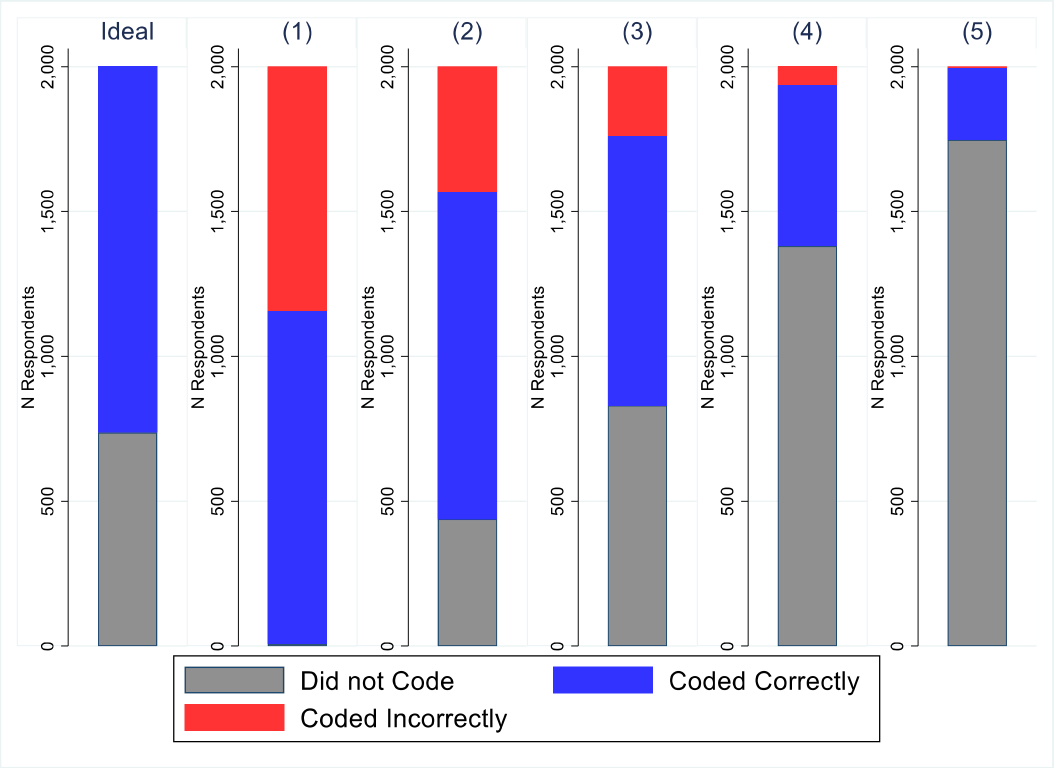


Figure 2: Comparison of Model Performance under Varying Strictness Conditions


Ideal
(1)
(2)
(3)
(4)
(5)
N Respondents
Did not Code
Coded Correctly
Coded Incorrectly