methods, data, analyses | Vol. 19(2), 2025, pp. 125-145 DOI: 10.12758/mda.2025.04

Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error

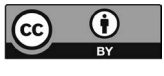
Brady T. West[1], Martin Slawski[2] & Emanuel Ben-David[3]

[1] *Institute for Social Research, University of Michigan-Ann Arbor*
[2] *Department of Statistics, University of Virginia*
[3] *U.S. Census Bureau*

Abstract

Modern predictive modeling tools, such as random forests (and related ensemble methods), have become almost ubiquitous in research applications involving innovative combinations of survey methodology and data science. However, an important potential flaw in the widespread application of these methods has not received sufficient research attention to date. Researchers at the junction of computer and survey science frequently leverage linked data sets to study relationships between variables, where the techniques used to link two (or more) data sets may be probabilistic and non-deterministic in nature. If frequent mismatch errors occur when linking two (or more) data sets, the commonly desired outputs of predictive modeling tools describing relationships between variables in the linked data sets (e.g., variable importance, confusion matrices, RMSE, etc.) may be negatively affected, and the true predictive performance of these tools may not be realized. We demonstrate a new methodology based on mixture modeling that is designed to adjust modern predictive modeling tools for the presence of mismatch errors in a linked data set. We evaluate the performance of this new methodology in an application involving the use of observed Twitter/X activity measures and predicted socio-demographic features of Twitter/X users to accurately predict linked measures of political ideology that were collected in a designed survey, where respondents were asked for consent to link any Twitter/X activity data to their survey responses (exactly, based on Twitter/X handles). We find that the new methodology, which we have implemented in R, is able to largely recover results that would have been seen prior to the introduction of mismatch errors in the linked data set.

Keywords: modern predictive modeling, ensemble methods, record linkage, mismatch error, mixture modeling, linked survey and social media data



In recent years, social media platforms such as Instagram and Twitter/X have provided social scientists with a wealth of user-content data (Agarwal et al., 2011; Bello-Orgaz et al., 2016; Ghani et al., 2019; McCormick et al., 2017). These data are often collected from multiple sources and then combined by probabilistic record linkage; for example, a research team might link two social media data sets, or link one social media data set to survey data (Al Baghal et al., 2021; Conrad et al., 2021; Eady et al., 2019; Karlsen & Enjolras, 2016). Researchers analyzing these linked data sets often apply advanced machine learning techniques, such as random forests, boosting (and related ensemble methods), neural networks, etc., whether the objective of the research project is accurate prediction of categorical survey outcomes (e.g., indicators of survey cooperation) or regression-based prediction of continuous outcomes (e.g., Gautam & Yadav, 2014; Liu & Singh, 2021; Wan & Gao, 2015).

There is, however, a potential pitfall in the widespread application of these modern predictive modeling techniques to linked data sets that needs more research attention. Although linking these types of new data sources provides the required information for novel studies of the relationships between variables, errors in the record linkage process may distort the true relationships between variables that are brought together from different data sources due to *mismatch errors* and *missed-match errors*. Missed-match errors refer to the inability to link a record in one data source to a matching record in a second data source, ultimately preventing that record from being included in an analysis of the relationships between variables from the two data sources. This type of error can lead to a form of selection bias in estimates of relationships, in a setting where the records with missed matches are unique in terms of the relationship of interest (Little & Rubin, 2019). In the setting of linking social media data with survey data, this type of error can arise when survey respondents do not consent to researchers linking their survey data with the information extracted from a Twitter handle or other identifiers (e.g., full names) used for social media accounts (e.g., Al Baghal et al., 2020). In this paper, we do not consider the problem of *missed-match errors*, but we suggest future directions for research in this area in the Discussion.

Mismatch errors, which are the primary focus of the current study, arise when records from different data sources are incorrectly matched (see Figure 1). Several prior studies have demonstrated the attenuating effects of mismatch errors on estimates of relationships in classical parametric regression

Direct correspondence to
Brady T. West, Survey Research Center of the Institute for Social Research, University of Michigan-Ann Arbor, USA.
E-Mail: bwest@umich.edu

modeling settings, and proposed approaches for correcting this attenuation (Dalzell & Reiter, 2019; Han & Lahiri, 2019; Lahiri & Larsen, 2005; Neter et al., 1965; Scheuren & Winkler, 1997, 1993; Slawski et al., 2021; Steorts et al., 2018; Tancredi & Liseo 2015). In the setting of linking social media data with survey data, obtaining consent from respondents to link their survey responses with the social media content that they generate is required (Stier et al., 2020). In this setting, mismatch errors may arise when the names provided by the consenting survey respondents do not match with the names used for social media accounts, the full names provided do not uniquely identify individuals, when social media platform handles corresponding to user accounts are provided with typos that prevent exact matching, or when consenting respondents change their platform handles over time (Beuthner et al., 2021; Stier et al., 2020).

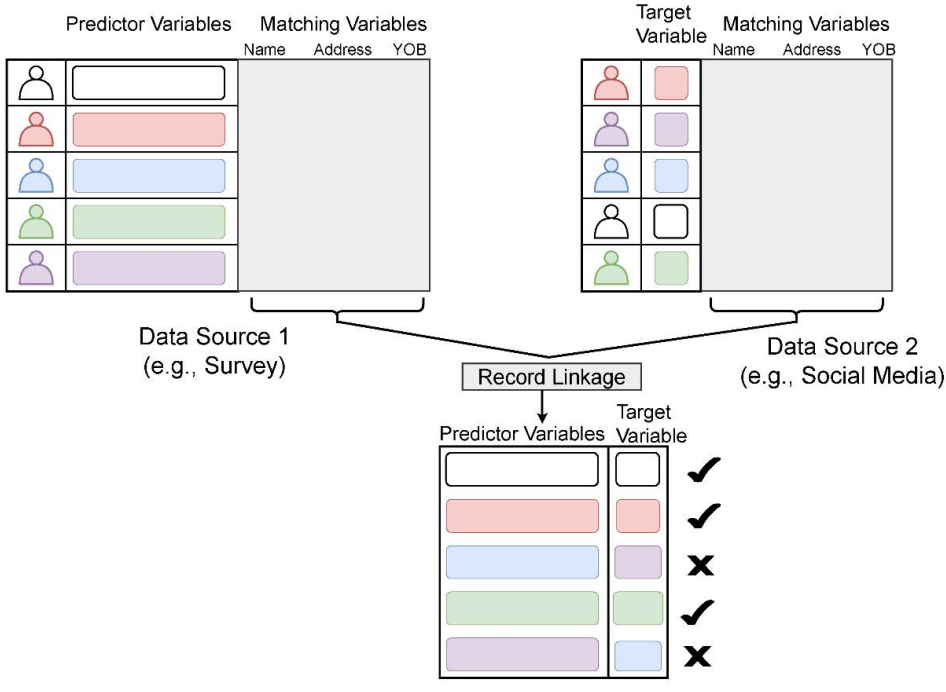


Figure 1 A visual overview of the mismatch error problem. Record linkage produces a linked file from two data sources containing predictor variables (Source 1) and the target (or dependent) variable (Source 2), respectively, based on a set of matching variables common to both data sources. The resulting linked file consists of correct matches (checkmarks) and mismatches (crosses).

This type of "fuzzy matching" can produce record linkages where the probability of a correct match is lower than 1 for certain records in the linked data set. This type of error in record linkage can produce outliers in terms of relationships of interest and may adversely alter the performance and outputs of applied predictive modeling techniques, such as variable importance, confusion matrices, RMSE, etc. Mismatch errors may ultimately prevent the realization of the actual predictive performance of these machine learning techniques, introducing a need for adjustments to the predictions that correct for this problem. Addressing the general absence of such adjustment approaches in the literature, Ben-David et al. (2023) derived and described novel adjustment techniques for the machine learning context based on a general mixture modeling framework (Hof & Zwinderman, 2015; Slawski et al., 2024). Via theoretical development and empirical simulation studies, these authors demonstrated that the proposed adjustment approaches can effectively improve predictions based on selected machine learning algorithms in the presence of various levels of mismatch error.

In this paper, our goal is to apply the methodology presented by Ben-David et al. (2023) to the specific context where 1) survey researchers are interested in linking survey and social media data, 2) fuzzy matching in the record linkage process is likely to introduce mismatch errors, and 3) the researchers wish to apply machine learning techniques to study relationships of interest in the linked data set. We evaluate the performance of this new adjustment methodology in an application involving the use of observed Twitter activity measures and predicted socio-demographic features of Twitter users to accurately predict linked measures of political ideology that were collected in a designed survey, where respondents were asked for consent to link any Twitter activity data to their survey responses (exactly, based on Twitter handles). We aim to demonstrate the use and importance of this new adjustment methodology to survey researchers interested in linking new sources of social media to survey data and ultimately applying machine learning techniques to the resulting linked data sets. We also summarize the limitations of the current adjustment approaches and make recommendations for future work in this area.

Methodology

An Overview of Adjustment Approaches Based on Mixture Modeling

We begin with an overview of our general approaches to adjusting modern predictive modeling algorithms for the presence of mismatch error. This paper focuses on possible adjustment techniques for *ensemble methods*, including bagging (or bootstrap aggregating) and random forests (distinguished from bagging by the selection of a random subset of predictors at each step of decision tree

construction). For brevity, we focus on a heuristic explanation of the approaches and do not provide explicit mathematical or technical details here; interested readers can find these details in Ben-David et al. (2023).

In general, we are interested in using an ensemble method to estimate some general regression function $\mu_{y|x} = E[y|x]$, where y corresponds to a dependent variable of interest and x represents a vector of values on predictor variables of interest. The new adjustment methods introduced in this paper assume that the x variables are measured without error; we revisit this issue in the Discussion section. After a record linkage process, we have values on these variables of interest available for each subject in a study denoted by i, with $i = 1, \ldots, n$. In the *permuted* linked data file that arises due to a record linkage procedure subject to mismatch error (Figure 1), we (unfortunately) observe $\tilde{y}_i$ instead of y_i, where some fraction of the cases in the linked data file have a mismatched value on the dependent variable y. These mismatches are the source of the attenuation in the estimated relationships of interest defined by the regression function.

Following a mixture modeling approach, the overall distribution of the permuted version of y is a *mix* of two distributions: the conditional distribution of y defined by the regression function for those correctly matched cases (which gets a weight of $1 - \alpha$, where α is the probability of a mismatch error, meaning that the weight is the probability of a *correct* match), and the *marginal* distribution of y for the mismatched cases (without conditioning on the covariates), which gets a weight of α. The mixture model is flexible enough to allow a *unique* value of α for each case, denoted by α_i.

This mixture model implies that we can write the regression function as follows (where μ_y is the marginal mean of the variable y):

$$\mu_{y_i|\mathbf{x}_i} = \frac{1}{1-\alpha_i}\mu_{\tilde{y}_i|\mathbf{x}_i} - \frac{\alpha_i}{1-\alpha_i}\mu_y,\ i = 1, \ldots, n \tag{1}$$

When analyzing real data in practice, we would first apply the analyst's favorite predictive modeling algorithm to the linked data including mismatch errors. Given the resulting estimates of $\hat{\mu}_{\tilde{y}_1|\mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n|\mathbf{x}_n}$, along with the sample mean of the observed $\tilde{y}_i$, we can then substitute these quantities in (1). As a result, we can write the overall distribution of the permuted y as a function of α_i alone. Then, we can use maximum likelihood methods (or other optimization methods) to find an optimal $\hat{\alpha}_i^{opt}$ (see Algorithm 1 in Ben-David et al., 2023). This $\hat{\alpha}_i^{opt}$ can then be used in (1) to obtain an *improved* estimate of $\mu_{y_i|\mathbf{x}_i}$. We can also simply work with the mean of the $\hat{\alpha}_i^{opt}$, $\hat{\alpha}^{opt} = \sum_{i=1}^{n}\hat{\alpha}_i^{opt}/n$, in (1). We refer to this as a "mean optimal alpha" approach, which has the potential to save computational time. This is because we can efficiently estimate $\hat{\alpha}^{opt}$, the population mean of

the $\hat{\alpha}_i^{opt}$, using the mean of a small random sample of the $\hat{\alpha}_i^{opt}$, with size much smaller than n.

The improvement in estimates of $\mu_{y_i \mid \mathbf{x}_i}$ based on this approach thus depends on (1) $\hat{\alpha}^{opt}$ being a good estimate of α, (2) $\hat{\mu}_{\tilde{y}_i \mid \mathbf{x}_i}$ being a good estimate of $\mu_{\tilde{y}_i \mid \mathbf{x}_i}$ (i.e., the regression function is specified correctly), and (3) the mixture model being a good fit for the overall distribution of the permuted y values. We note that this "optimal alpha" adjustment method would generally be applied *after* any other predictive modeling algorithm has been used to generate initial predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ for all cases in the linked data file.

Extending this idea to the more general context of the ensemble methods that are the focus of the current study, the α values described above can play the role of *weights* in the algorithms used to build the decision trees. We distinguish between two different approaches to using weights in the construction of decision trees: *adj-trees*, where differential case weights are used at each step of the tree construction process to determine optimal splits, and *adj-rf*, where differential case weights are used when the bootstrap samples are selected for the ensemble method (and cases with a higher weight would have a higher probability of selection).

Given no prior information about the mismatch probabilities, we would assign a weight of 1 to each case and set $\alpha_i = 0.5$ for all cases. We can then take, say, 100 bootstrap samples from the data (this number could be modified). For each sample, we first obtain $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ from a decision tree, or random forests, with our initial weights. We can then use methods described in Ben-David et al. (2023) to compute the *posterior probability* of a mismatch given the predicted values according to the regression function, and then update the weight of each observation i as $1 - \alpha_i$. We then re-run the decision tree, or random forests, with these updated weights (which again either affect how the bootstrap samples are selected *or* how the tree is split at each node) to compute a new set of predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$. We repeat this procedure, updating the weights and then updating $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$, until there is no numerical evidence of a significant improvement in the predictions obtained with the new weights. In the end, we average over the $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$ obtained from the final set of bootstrap samples and report this as the adjusted predictions $\hat{\mu}_{\tilde{y}_1 \mid \mathbf{x}_1}, \ldots, \hat{\mu}_{\tilde{y}_n \mid \mathbf{x}_n}$.

Ben-David et al. (2023) refer to this general approach as a *weighting-reweighting* adjustment method (Algorithm 2). Figure 2 visualizes this general approach. In theory, this adjustment procedure that assigns greater weight to cases with higher estimated probabilities of being a correct match will yield ensemble predictions with improved accuracy; simulations reported by Ben-David et al. (2023) provide empirical support for this concept.

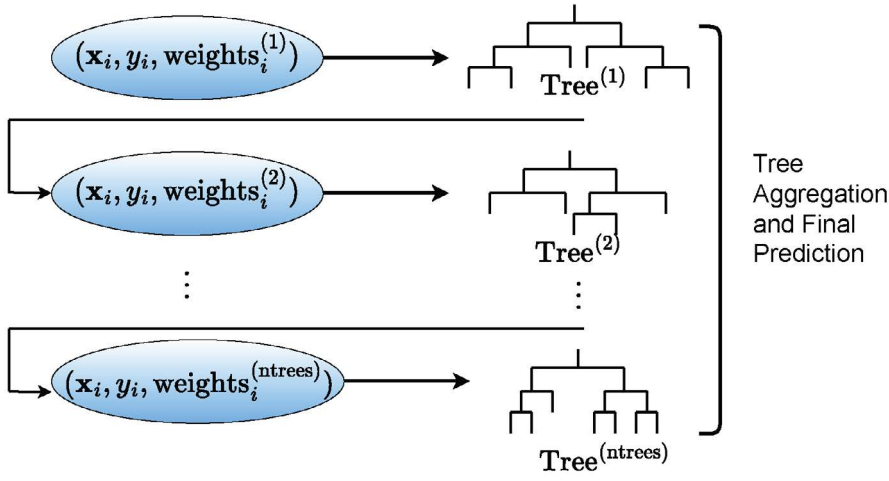


Figure 2 A visual overview of the weighting-reweighting adjustment method in the context of ensemble methods such as bagging and random forests.

There are, therefore, several possible combinations of approaches that one could use when applying these ensemble methods to a linked data set. We distinguish between four methods that do not include the computation of optimal alpha values (referred to as basic bagging, basic random forests, adj-trees, and adj-rf) and four methods that do include the subsequent computation of optimal alpha values (optimal-alpha-bagging, optimal-alpha-rf, optimal-alpha-adj-trees, and optimal-alpha-adj-rf). In our analyses, we evaluate the performance of these eight alternative methods, summarized below in Table 1.

Data Source

We conduct secondary analyses of a linked data set (n = 448) that includes data from web survey respondents and aggregated measures of social media activity based on their linked Twitter profiles (we refer to Twitter, rather than X, as this data collection occurred prior to the change in the name of that platform). The web survey data, capturing measures of social media use, political attitudes and knowledge, and other related topics, were collected from a random sample of the Ipsos KnowledgePanel in January and February of 2020 (response rate = 76%); see Mneimneh (2022) for the original study design details. The record linkage was based on actual Twitter handles for those respondents who consented to this linkage, meaning that the record linkage was largely deterministic, exact, and error-free.

Given the objectives of our study, we randomly permuted the linked social media data to simulate mismatch errors (as the actual record linkage process used was unlikely to result in mismatch errors). As we noted in the Introduction, these types of mismatch errors may arise for several reasons when linking survey and social media data, but this mismatch error scenario may be even more common in other applications that involve linking survey data and administrative data (e.g., Patki & Shapiro, 2023).

Table 1 Alternative adjustment methods under consideration (none = no adjustment).

Adjustment method	Description
Bagging (none)	This is a standard application of bootstrap aggregating (bagging) using the original linked data and no random selection of predictors at each step of the tree construction.
Random forests (none)	This is a standard application of random forests similar to bagging but including the random selection of possible predictors at each step of the tree construction.
Adj-trees	The weighting-reweighting adjustment method, starting with default values of alpha (0.5) for all cases (and equal weights of 1), and then proceeding iteratively with applying weights to cases when splits are determined to construct individual trees. Improved estimates of the regression function are based on the mixture model.
Adj-rf	Like adj-trees, but applying the weights in the selection of the bootstrap samples (rather than in the formation of splits).
Optimal-alpha-bagging	A modification of bagging including a subsequent application of the optimal alpha algorithm to improve adjusted estimates based on the mixture model. Given our results and the additional computational burden introduced by using a unique optimal alpha for each case (without apparent benefits of this approach), we focus on the mean optimal alpha value for all "optimal alpha" approaches.
Optimal-alpha-rf	This is a modification of random forests, including the application of the optimal alpha algorithm to improve adjustment estimates based on the mixture model.
Optimal-alpha-adj-trees	This is a modification of adj-trees to include the optimal alpha algorithm.
Optimal-alpha-adj-rf	This is a modification of adj-rf to include the optimal alpha algorithm.

Measures

In our analysis, we focus on applying predictive modeling where we wish to predict a dependent variable representing an ordered measure of political ideology collected in the web survey. This question asked, "In general, do you think of yourself as..." and provided the following response options: 1 = extremely liberal; 2 = liberal; 3 = slightly liberal; 4 = moderate, middle of the road; 5 = slightly conservative; 6 = conservative; and 7 = extremely conservative. Given the roughly symmetric distribution of this variable among the survey respondents, we treated the variable as a continuous outcome in our analyses. Candidate predictors of this survey measure were all derived from the linked Twitter data. These included predictions of the person's gender (male vs. female) and age (>45 or <= 45) based on a neural network model (Liu & Singh, 2021), along with predictions of gun ownership (yes or no) and political party (Democrat or Republican) based on a random forest classifier using features of tweets and Twitter biographies. We also included as a predictor the overall number of tweets generated by the survey respondent (based on actual Twitter activity for the linked Twitter handle). We assume that all of these measures derived from the Twitter data are error-free; we return to this issue in the Discussion section.

Analytic Approach

In our evaluation of the eight alternative adjustment approaches described in Table 1, we first applied each of the eight approaches to the exactly matched Twitter and survey data (i.e., a 0% mismatch rate), evaluating the mean squared error (MSE) of the predictions for political ideology based on the correctly linked data. This initial analysis provided a benchmark for evaluating the success of the adjustment methodology after varying levels of mismatch error were introduced via random permutations (10%, 15%, ..., 35%, 40%). We then evaluated the ability of the eight different approaches to recover this "ideal" MSE of the predictions based on the correctly-linked data. We constructed 100 trees based on bootstrap replicate samples for each ensemble method. We repeated these analyses 100 times and averaged the estimated MSE values across these 100 iterations.

Because ensemble methods may also be computationally expensive depending on the size of the data set and the number of predictors under consideration, we also compared the computational times associated with executing each adjustment procedure (based on a single run of each procedure). We provide separate computational times for each of the two algorithms described earlier, given that the use of optimal values of alpha for the weighting-reweighting adjustment approach also requires execution of the first algorithm to identify optimal values of alpha (possibly for each individual case). We weigh the comparisons of the procedures in terms of MSE based on the computational run

times to identify an optimal adjustment procedure that is also computationally efficient.

Results

Table 2 compares the alternative adjustment methods in terms of the average estimated MSEs of the predictions of political ideology, averaged across the 100 iterations of each analysis and separately for different simulated mismatch rates.

Table 2 Relative performance of each adjustment procedure in terms of average estimated MSE (across 100 iterations) of the predictions for political ideology (best performance indicated in boldface).

	Mismatch rate							
Adjustment method	0%	10%	15%	20%	25%	30%	35%	40%
Bagging	**1.63**	1.68	1.74	1.76	1.80	1.82	1.87	1.93
Random forests	2.02	2.05	2.10	2.13	2.13	2.14	2.19	2.23
Adj-rf	1.99	2.00	2.01	2.02	2.03	2.04	2.06	2.09
Adj-trees	**1.63**	1.68	1.73	1.76	1.80	1.82	1.87	1.93
Optimal-alpha-bagging	1.64	**1.64**	**1.68**	**1.70**	**1.74**	**1.76**	**1.80**	**1.86**
Optimal-alpha-rf	2.09	2.07	2.10	2.12	2.12	2.12	2.16	2.20
Optimal-alpha-adj-rf	2.15	2.09	2.09	2.09	2.10	2.11	2.11	2.13
Optimal-alpha-adj-trees	1.64	**1.64**	**1.68**	**1.70**	**1.74**	**1.76**	**1.80**	**1.86**

The performance of each procedure when all matches are correct (i.e., when analyzing the original linked Twitter data) can be found in the Mismatch rate column of Table 2 labeled "0%." In this setting, basic bagging and adj-trees have the best predictive performance (MSE = 1.63), and we use this as a benchmark to evaluate the performance of the alternative adjustment procedures when mismatches are introduced in the linked data. Examining the other columns of Table 2 corresponding to increasing mismatch rates (introduced by randomly permuting the values of the dependent variable for the indicated percentage of cases in the linked data set), we observe that the optimal-alpha-bagging and optimal-alpha-adj-trees approaches yield predictions that are consistently closest to the benchmark performance, with larger deviations from the benchmark as mismatch rates increase (as would be expected).

Given the results in Table 2, we next consider the computational run times associated with each procedure. Table 3 presents run times in seconds for the various components of the adjustment procedures.

Table 3 Run times in seconds for the various components of the adjustment procedures.

Optimal alpha	Mean optimal alpha	Bagging	Random forests	Adj-trees	Adj-rf
6.349	0.864	0.502	0.016	36.208	0.001

We note that a particular adjustment procedure may introduce the run times associated with each of the two algorithms. For example, the optimal-alpha-adj-trees approach requires subsequent execution of the optimal-alpha algorithm (6.349 seconds) following the adj-trees algorithm (36.208 seconds). Table 2 shows that the adj-trees approach tends to be computationally expensive. Combining these results with those in Table 2, it therefore seems that the optimal-alpha-bagging approach has the best overall performance in the setting considered here.

We have included the R code needed to carry out these analyses in the GitHub repository https://github.com/ehb2126/Data-Analysis-after-Record-Linkage.

Discussion

Summary of Contributions

Mismatch errors are common in probabilistic record linkage procedures. In the specific setting of linking survey data with social media data, these errors can arise for several reasons, including names provided by the consenting survey respondents that do not match with the names used for social media accounts, full names provided by consenting survey respondents that do not uniquely identify individuals, social media platform handles corresponding to user accounts containing typos that prevent exact matching, or consenting respondents changing their platform handles over time (Stier et al., 2020; Beuthner et al., 2021). At the same time, machine learning methods are becoming increasingly popular for studying complex relationships in the analyses of linked data sets from different sources (e.g., social media and survey data, or survey data and administrative data).

Much of the record linkage literature has focused on adjustment procedures for mismatch errors in classical parametric regression modeling. Recently, Ben-David et al. (2023) addressed an important gap in this area, focusing on optimal methods for adjusting for mismatch errors when applying modern prediction tools (specifically bagging and random forests) and describing alternative adjustment procedures for ensemble prediction methods within a mixture modeling framework. This paper applies these new adjustment methods to a case

study linking survey data with social media (specifically Twitter/X) data, and demonstrates that these methods improve the performance of modern predictive modeling methods that were applied to this linked data set under various simulated rates of mismatch error.

We find that in the presence of these various rates of mismatch error, an adjustment methodology that combines bagging with optimal estimation of the probability of correct linkage for each case tends to have the best predictive performance, from the perspectives of both MSE of predictions and computational runtime. This procedure is straightforward to implement using available software, and we have implemented it using the R software (see the GitHub repository https://github.com/ehb2126/Data-Analysis-after-Record-Linkage).

Limitations and Directions for Future Research

We note that studies linking social media data with survey data generally use exact platform handles or other types of unique identifying information in the record linkage, and do not attempt the linkage at all if respondents do not consent to provide these handles or other user account information, such as full names (e.g., Al Baghal et al., 2021). This introduces the possibility of missed-match errors, a type of selection bias that could affect the performance of predictive modeling methods. Selection bias due to missed-match errors could affect machine learning algorithms that are focused on prediction in three ways (Quiñonero-Candela et al., 2022):

> 1) *covariate shift*, where the distribution of the predictors x would differ across successfully linked cases and missed matches;
>
> 2) *label shift*, where the distribution of the dependent variable *y* would differ across successfully linked cases and missed matches; or
>
> 3) *concept drift*, where the distribution of *y* conditional on x would differ across successfully linked cases and missed matches, and the classification rule would depend on the successfully linked cases.

If we assume that an indicator of successful linkage is independent of *y* when conditioning on x, then *concept drift* does not hold, but this is a strong assumption that needs to be evaluated in future simulation studies. Adjustment approaches accounting for these types of missed match errors and allowing for violations of this assumption are still needed in the machine learning context; we only focused on mismatch errors in the current application.

The new methodologies illustrated in this paper also assume that the mismatch errors occur *completely at random*, using the terminology of Little and Rubin (2019) in the missing data context. This strong assumption may not hold in real applications, since the probability of a mismatch error may at least depend on the values of observed covariates. We designed an additional simulation study to evaluate the performance of the methodology in a setting where the probability of a mismatch depends on the value of the covariate that had the strongest relationship with political ideology in a linear regression model fitted to the political ideology outcome in the original data: the binary prediction of preferring the Republican party (1 = yes, 0 = no). The supplemental materials describe the design of this additional simulation study and the corresponding results.

Summarizing those results here, we find that the methods identified as having the best performance in the "mismatch completely at random" scenario have equally strong performance in this informative mismatch error scenario. Despite these positive results, additional theoretical development is still needed to understand why the current methodologies also seem to work well in this informative mismatch error setting; they are presently designed for mismatches occurring completely at random. Future research on this methodology should also aim to accommodate more complicated types of informative mismatch error scenarios.

We also did not quantify variable importance in our application of the adjustment methods. We have not yet developed a procedure for identifying the most important predictors that emerge from one of these adjustment approaches, and work on the development of adjusted variable importance measures is ongoing. This is another worthwhile direction for future research.

We also note that we assumed that all of the social media measures were of sufficiently high quality. These variables computed from the Twitter/X data were either predictions of user characteristics or counts of tweets that may themselves be subject to prediction error and sampling error. Future applications involving predictive modeling of linked survey and social media data need to carefully consider potential sources of error in derived variables from social media activity and ensure that these errors are either corrected, adjusted for, or transparently described in written summaries of the modeling applications.

Finally, while the adjustment approaches in this paper were evaluated in the context of mismatch error in linked social media and survey data, we anticipate that they will also have widespread application in other substantive settings where probabilistic record linkage is used (e.g., Patki & Shapiro, 2023) and researchers are interested in predictions based on machine learning procedures.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30–38). Association for Computational Linguistics. https://dl.acm.org/doi/abs/10.5555/2021109.2021114

Al Baghal, T., Sloan, L., Jessop, C., Williams, M., & Burnap, P. (2020). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review, 38* (5), 517–532. https://doi.org/10.1177/0894439319828011

Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: Asymmetry in quantity and its impact. *EPJ Data Science, 10* (1), 32. https://doi.org/10.1140/epjds/s13688-021-00286-7

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion, 28*, 45–59. https://doi.org/10.1016/j.inffus.2015.08.005

Ben-David, E., West, B. T., & Slawski, M. (2023). A novel methodology for improving applications of modern predictive modeling techniques to linked data sets subject to mismatch error. In *Big Data Meets Survey Science (BigSurv)* (pp. 1–8). IEEE. https://doi.org/10.1109/BigSurv59479.2023.10486610

Beuthner, C., Breuer, J., & Jünger, S. (2021). *Data linking – Linking survey data with geospatial, social media, and sensor data* (GESIS Technical Report, Version 1.0). https://doi.org/10.15465/gesis-sg_en_039

Conrad, F. G., Keusch, F., & Schober, M. F. (2021). New data in social and behavioral research. *Public Opinion Quarterly, 85* (S1), 253–263. https://doi.org/10.1093/poq/nfab027

Dalzell, N., & Reiter, J. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics, 27*(4), 728–738. https://doi.org/10.1080/10618600.2018.1458624

Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open, 9* (1), 1–21. https://doi.org/10.1177/2158244019832

Gautam, G., & Yadav, D. (2014). Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In *Seventh International Conference on Contemporary Computing (IC3)* (pp. 437–442). IEEE. https://doi.org/10.1109/IC3.2014.6897213

Ghani, N. A., Hamid, S., Targio Hashem, I. A., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior,* 101, 417–428. https://doi.org/10.1016/j.chb.2018.08.039

Han, Y., & Lahiri, P. (2019). Statistical analysis with linked data. *International Statistical Review, 87*(S1), 139–157. https://doi.org/10.1111/insr.12295

Hof, M., & Zwinderman, A. (2015). A mixture model for the analysis of data derived from record linkage. *Statistics in Medicine, 34*(1), 74–92. https://doi.org/10.1002/sim.6315

Karlsen, R., & Enjolras, B. (2016). Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with Twitter data. *The International Journal of Press/Politics, 21*(3), 338–357. https://doi.org/10.1177/1940161216645335

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association, 100*(469), 222–230. https://doi.org/10.1198/016214504000001277

Little, R. J., & Rubin, D. (2019). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

Liu, Y., & Singh, L. (2021). Age inference using a hierarchical attention neural network. In *Proceedings of the ACM International Conference on Information & Knowledge Management* (pp. 3273–3277). Association for Computing Machinery. https://doi.org/10.1145/3459637.3482055

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., & Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research, 46* (3), 390–421. https://doi.org/10.1177/0049124115605339

Mneimneh, Z. (2022). Evaluation of consent to link twitter data to survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(Supplement 2), S364–S386. https://doi.org/10.1111/rssa.12949

Neter, J., Maynes, S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association, 60*(312), 1005–1027. https://doi.org/10.1080/01621459.1965.10480846

Patki, D., & Shapiro, M. D. (2023). Implicates as instrumental variables: An approach for estimation and inference with probabilistically matched data. *Journal of Survey Statistics and Methodology, 11*(3), 597–618. https://doi.org/10.1093/jssam/smad005

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds). (2022). *Dataset shift in machine learning*. The MIT Press. https://mitpress.mit.edu/9780262545877/dataset-shift-in-machine-learning/

Scheuren, F., & Winkler, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology, 19*(1), 39–58. https://www150.statcan.gc.ca/n1/pub/12-001-x/1993001/article/14476-eng.pdf

Scheuren, F., & Winkler, W. (1997). Regression analysis of data files that are computer matched – Part II. *Survey Methodology, 23*(2), 157–165. https://www150.statcan.gc.ca/n1/pub/12-001-x/1997002/article/3613-eng.pdf

Slawski, M., Diao, G., & Ben-David, E. (2021). A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics, 30*(4), 991–1003. https://doi.org/10.1080/10618600.2020.1870482

Slawski, M., West, B. T., Bukke, P., Wang, Z., Diao, G., & Ben-David, E. (2024). A general framework for regression with mismatched data based on mixture modelling. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae083. https://doi.org/10.1093/jrsssa/qnae083

Steorts, R. C., Tancredi, A., & Liseo, B. (2018). Generalized Bayesian record linkage and regression with exact error propagation. In J. Domingo-Ferrer & F. Montes (Eds.), *Privacy in statistical databases. PSD 2018. Lecture Notes in Computer Science* (Vol. 11126, pp. 295–306). Springer. https://doi.org/10.1007/978-3-319-99771-1_20

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review, 38*(5), 503–516. https://doi.org/10.1177/0894439319843

Tancredi, A., & Liseo, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica, 75*(1), 19–35. https://www.proquest.com/docview/1765123569?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals

Wan, Y., & Gao, Q. (2015). An ensemble sentiment classification system of Twitter data for airline services analysis. In *IEEE International Conference on Data Mining Workshop (ICDMW)* (pp. 1318–1325). IEEE. https://doi.org/10.1109/ICDMW.2015.7

Appendix

Improved Ensemble Predictive Modeling Techniques for Linked Social Media and Survey Data Sets Subject to Mismatch Error

Simulation Study of Informative Mismatch Error

To compare the MSE of methods where mismatches correlate with a given predictor x (in this case, the binary prediction of preferring the Republican party) versus ones where the mismatches occur completely at random, we need to consider the average percentage of cases where x = 1 is swapped with x = 0 when mismatches are introduced completely at random. Table A1 below shows this average for various mismatch rates using the same simulation approach described in the paper, with each average computed numerically based on 100,000 replications.

Table A1 x = 1 permutation rates introduced by mismatches occurring completely at random.

Percentage of mismatches (completely at random)	Average percentage of x = 1 swapped with x = 0
10%	10%
15%	17%
20%	21%
25%	23%
35%	31%
40%	35%

For this supplemental simulation study, we selected probabilities of changing values from x = 1 to x = 0 that were consistent with the table above, to ensure that overall mismatch rates were similar to those already evaluated in the paper.

Table A2 below shows the MSEs of bagging, random forests, and the new adjustment methods for these informative mismatch error scenarios. We used the same simulation approach described in the paper, but allowed the probability of a mismatch error to change for cases with x = 1. Each column of Table A2 below shows the MSEs of these methods in a different mismatch scenario, where $P(x' = 0 \mid x = 1)$ varies according to percentages comparable with the "completely at random" mismatch rates given in Table A1 above. In addition, the probability of a mismatch error was set to be larger for cases with x = 1: $P(x' = 1 \mid x = 0) = (71/377) \times P(x' = 0 \mid x = 1)$, which shows how the covariate was related to the probability of mismatch error. The MSEs in the table are averaged over 250 iterations.

Table A2 Informative mismatch error simulation results (MSEs).

P(x' = 0 | x = 1)	0	0.1	0.17	0.21	0.23	0.25	0.31	0.35
P(x' = 1 | x = 0)	0	0.019	0.032	0.040	0.043	0.047	0.058	0.066
Bagging	1.63	1.64	1.66	1.67	1.69	1.69	1.74	1.76
Random forests	2.02	2.05	2.08	2.10	2.11	2.12	2.17	2.20
Adj-rf	1.99	1.99	2.00	2.00	2.01	2.01	2.03	2.05
Adj-trees	1.63	1.64	1.66	1.67	1.69	1.69	1.74	1.76
Optimal-bagging	1.64	1.60	1.59	1.59	1.60	1.60	1.63	1.65
Optimal-rf	2.09	2.08	2.09	2.10	2.10	2.11	2.15	2.17
Optimal-adj-rf	2.15	2.12	2.10	2.09	2.09	2.08	2.08	2.09
Optimal-adj-trees	1.64	1.60	1.59	1.59	1.60	1.60	1.63	1.65

Overall, we see performance quite similar to that in the mismatch completely at random scenario that was analyzed in the paper. The adj-trees, optimal-bagging, and optimal-adj-trees approaches all tend to have the best performance, and while the MSEs increase somewhat as the "conditional" mismatch probabilities increase, these methods consistently have the best performance in this scenario.

Reflective Appendix

1. If you had been required to pre-register your methodological approach, in advance of conducting your research, how would you have described it?

The main objective of the paper was to evaluate the ability of a new methodology, presented by the authors in previously published work, to improve the accuracy of predictions of respondents' self-reported political ideology using their Twitter account information. For purposes of the analysis, we obtained a secondary dataset of US Ipsos KnowledgePanel participants (n = 448) who responded to a web survey measuring social media use, political attitudes and knowledge, and other related topics. All 448 web survey respondents had consented to let the study designers link their survey responses to their Twitter/X accounts, and all 448 had provided their correct Twitter handles for the linkage. Given the 100% matching of respondents to their Twitter accounts, we had to simulate mismatch error to test our methodology. The methodological approach used bagging and random forest techniques to predict an ordinal dependent variable measuring political ideology from the web survey. The predictors of interest included predicted socio-demographic information and aggregated measures of activity from the linked accounts on the Twitter/X platform.

2. To what extent did you make any modifications to your plans as described above in the course of producing the final version of the paper?

Since we were using secondary data that was selected explicitly for the purposes of this analysis, we were able to specify clearly and precisely the methodology adopted in advance and we did not deviate from what was originally proposed. Having a dataset with no mismatches provided us with an important and valuable benchmark against which to assess the performance of the varying methods for predicting survey answers on ideology from individuals' Twitter data. However, having a pre-cleaned dataset generates some "costs" to the analysis in terms of limiting the diagnostics we were able to perform and our capacity to make adjustments or modifications to the analysis. Having access to the larger original dataset from which this subset of 448 correctly matched individuals were drawn would have provided the opportunity to draw more generalizable conclusions. Specifically, prior work using this larger dataset (Mneimneh, 2022) revealed that of the 58.6% of survey respondents that consented to the Twitter linkage, less than half (48%) provided a useable handle. Had the dataset included the correctly matched and larger sample of mismatched respondents it would have been possible to conduct diagnostics on the former sample to see how closely they resembled the latter, and whether mismatches were more likely for certain types of individuals. For example, were our sample respondents more active on Twitter than the average user? The counts of tweets and retweets ranged from 21 to 75,077 for the fully matched sample of 448 we investigated and no individuals had counts of zero or missing data. If bias did exist, this type of "informative" mismatch error would have been useful for adjusting our hypothetical simulated mismatch error to correspond to "organic" mismatch error, thereby enhancing the robustness of its application to larger samples that are more at risk of the "organic" mismatch error.

Additionally, having access to the larger sample that included organically mismatched survey respondents would have allowed us to replicate the analysis performed here, and assess the results of the predictive algorithms employed for the correctly matched subsample in the larger (more representative) sample of Twitter users. It is possible that our methodology is more effective than is demonstrated here in the context of stronger predictive models (where the mismatch error is likely to have larger attenuating effects on predictive performance). For this we would need to have access to aggregate information regarding the characteristics of the Twitter-using population at the time when this survey was collected, and microdata from non-consenting respondents to determine whether the power of the predictive models was lower in our subsample compared to a larger set of Twitter users.

3. Can you list up to 3 practical steps that you would recommend, based on what you learned doing this research project, future researchers take into

account when working with similar data sources? These would ideally be relevant to the methods, data and analysis chosen.

Four practical steps that we would recommend when working with linked survey and social media data and using predictive modeling techniques to study relationships among variables from the two linked data sources:

1. Evaluate the quality of the record linkage process. Were all linkages correct, or was there evidence of problems in the process that would lead to potential mismatches? For example, are the useable Twitter handles provided actually those of the survey respondents, or those of *other* individuals? Providing the Twitter handle of another individual could lead to another type of mismatch error (above and beyond the sources mentioned in the paper), and our methodology would still be able to accommodate this alternative type of linkage error. Asking consenting survey respondents for their two most recent Twitter handles to facilitate record linkage is one possible approach for dealing with non-useable handles or handles of other users that survey respondents have provided. All else being equal, we would avoid the collection of names, addresses, and other personal identifying information in an attempt to resolve these problems, as this raises new ethical concerns. Our methodology is designed to address the mismatch errors that may result from this "less than ideal" type of respondent behavior.

2. If alternative variables aside from the Twitter handles are ultimately used to perform the linkage for selected cases, try to obtain information on the correctness of each link, when possible, taking into account ethical considerations regarding the protection of respondent confidentiality. For example, this could involve calculating the predicted probability of correct linkage arising from a probabilistic record linkage procedure. Such information will be helpful in informing adjustment approaches like the one evaluated in this paper. Alternatively, if such information is not available, anticipated mismatch rates or block-wise mismatch rates can still be helpful. Blocks define different groups of cases based on discrete observable characteristics (sex, race, age, etc.) within which linkage is considered.

3. If there is a risk of mismatch error and the analyst ultimately wants to use a modern predictive modeling technique to predict values of a variable in one data source with predictors from the other data source, consider using the R software provided with this paper to perform the bagging and random forests (rather than standard procedures implementing these techniques, which would be adversely affected by the mismatch error). We find in this paper that an adjustment methodology that combines bagging with optimal estimation of the probability of correct linkage for each case tends to have the best predictive performance.

4. Compare the performance of the adjusted predictive modeling methods with that of the standard predictive modeling methods to quantify the potential effects of the mismatch error on the performance of the techniques. If notable differences in performance are observed, report predictions based on the adjusted methods, as the standard techniques were likely affected by the mismatch errors engendered by the record linkage process. If a fraction of the linked data set has records that were linked deterministically (i.e., with no probability of linkage error), predictions based on a machine learning algorithm applied to those "exactly matched" cases could be used as a benchmark, and predictions based on an application of our methodology to the *full* data set (including mismatch errors) could be compared in a validation analysis to assess the effectiveness of our methodology.

More generally, we believe that the methodology illustrated in this case study can be generalized and transported to other applications involving the linkage of records in novel data sources, and we also believe that several extensions of our approach are possible. First, mismatch errors may not necessarily cause estimation problems (for example, in a binary classification problem, if the mismatched record has the same value on the binary variable, there is no impact). Second, we may want to use information on "local" mismatch rates (e.g., changing with covariates or with information about the quality of the record linkage) rather than using a global mismatch rate model (as was the case in this study). The availability of metadata in the specific setting of linking social media information (e.g., geographic location, age of account, employment history, etc.) may be helpful for improving estimation of these "local" mismatch rates, in turn improving the adjustments engendered by our methodology. In the simulation study that we considered to look at "informative" mismatch errors, we found that the methods identified as having the best performance in the "mismatch completely at random" scenario have equally strong performance in the informative mismatch error scenario. However, we also found that the conditional mismatch rate did have an impact on predictive performance overall, meaning that additional enhancements of our methodology to implement larger adjustments for certain subgroups of cases with larger associated mismatch rates may be important.

Third, missed matches may be even more problematic than mismatch errors, given that we might be training a machine learning model on a non-representative sample. Fourth, our approach can be connected to robustness (Slawski et al., 2021). In this setting, other sources of measurement error or outliers can be handled simultaneously. At the same time, mismatch errors and other measurement errors cannot be distinguished, which makes sense, since their impact is often identical. For example, an incorrect link or a data entry error in terms of computed Twitter activity are equivalent in terms of impact. Fifth, the size of the data set may be important when choosing the specific adjustment method. The

optimal-alpha method may be prohibitive from the point of view of computational runtime. Finally, the general methodology illustrated here can be applied to other types of variables of interest; regression functions based on specified link functions in generalized linear models could be used as part of the algorithm to accommodate other types of dependent variables of interest (binary, count, etc.) in other linked data sets.

We believe that it would be easy to apply our methodology (designed for the secondary analysis setting) in other settings where novel data sources have been linked, a data analyst who did not perform the linkage is working with the linked data file, and mismatch errors are suspected. As noted in the paper, having the probability of a correct match for each case can be helpful for the algorithms, but is *not required*. The methodology can therefore proceed in the absence of quality indicators in the linked data file, making applications of the methodology easy for data users (no matter what types of data sources have been linked). We are confident that the optimal adjustment approach identified in this case study would also emerge in other larger data sets or other applications involving the linkage of other types of novel data sources. However, as we note above, the optimal alpha adjustment method may be computationally prohibitive in larger data sets, in which case the "next best" methods may need to suffice.

Additional research involving applications of our methodology in other settings would shed more light on these practical and computational issues, and likely lead to additional refinements of the new methodology presented here. As we note in the Discussion, additional research examining the performance of the adjustment methodology in more complex settings of *informative* mismatch error is necessary. We performed a relatively simple simulation study, and depending on the variables of interest in the analysis and the nature of the informative mismatch error as a function of these variables, other more optimal approaches may emerge. There is a possibility that linking alternative types of data sources could result in more complex patterns of informative mismatch error (e.g., certain socio-demographic subgroups are less likely to provide correct or truthful information related to handles for a particular type of social media platform), and this should be a focus in future research that seeks to refine our adjustment methodology.