Measuring Human Capital with Social Media Data and Machine Learning

Martina Jakob*

University of Zurich
Department of Economics
Zurich, Switzerland
martina.jakob@econ.uzh.ch

and

University of Bern
Department of Social Sciences
Bern, Switzerland
martina.jakob@unibe.ch

Sebastian Heinrich

ETH Zurich
Department of Management, Technology, and Economics
Zurich, Switzerland
heinrich@kof.ethz.ch

January 30, 2026

Keywords: Machine Learning, Social Media Data, Education, Human Capital, Indicators, Natural Language Processing

Classification: Economic Sciences, Social Sciences

***Corresponding Author:** Martina Jakob, martina.jakob@econ.uzh.ch, University of Zurich, Department of Economics, Zurich, Switzerland

Abstract

Timely data on educational attainment at granular geographic levels remains scarce in many countries, limiting evidence-based policy-making. Recent advances in machine learning have enabled the use of non-traditional data sources like satellite imagery and mobile phone records to measure development indicators. While these approaches have been successful in predicting outcomes such as wealth, poverty, or population density, previous attempts to predict educational attainment have achieved only modest accuracy. A key challenge is finding data sources that directly reflect human capital rather than its economic consequences. Here we show that language patterns and user behavior in social media can explain up to 70 percent of the variance in regional educational attainment. Our machine learning framework leverages linguistic features, user behavior, and network characteristics from 25 million geolocated tweets from the United States and Mexico. It performs particularly well in predicting higher education levels and maintains a good performance even with limited data collection periods. These results show that digital communication patterns can serve as reliable proxies for human capital. In light of the rapid expansion of social media use around the globe, this represents a promising approach to tracking educational outcomes in regions lacking granular and timely survey data.

1 Introduction

Reliable data on key socio-economic outcomes enables policy-makers to make informed decisions and promote societal development. However, many countries are plagued by a pervasive lack of such data, limiting their ability to track progress and evaluate policies. To address the problem, a growing body of literature uses alternative data sources such as satellite imagery or phone records to bridge the existing gaps in data availability (Burke et al., 2021). While previous studies have successfully predicted outcomes such as wealth, income or population density, this paper proposes an innovative approach to measuring human capital using geolocated Twitter data (now X).

We focus on educational attainment as our measure of human capital (Becker, 1964; Mincer, 1974), and validate our approach with two countries that differ markedly in language, economic development, and social media penetration: the United States and Mexico. Specifically, we construct a set of interpretable measures of education at low administrative units (municipality in Mexico and county in the United States) based on over 25 million tweets. Our feature matrix includes basic Twitter penetration (e.g., user density) and usage statistics (e.g., tweet length), text-based indicators on spelling mistakes (e.g., frequency of grammatical errors), topics (e.g., share of tweets about science), and sentiments (e.g., share of negative tweets), as well as network indicators (e.g., closeness centrality). We train a stacking regressor combining five machine learning algorithms — elastic net regression, gradient boosting, support vector regression, nearest neighbor regression, and a feed-forward neural network — to predict educational attainment for Mexican municipalities (N = 2,457) and US counties (N = 3,141). We apply grid search to tune the relevant hyperparameters of each model, and evaluate the performance of the final models using five-fold cross-validation.

Our predictions account for 70 percent of the variation in years of schooling in Mexican municipalities and 65 percent in US counties. Where, how and what people tweet is thus highly informative about human capital. Within both countries, Twitter data appears to be particularly well-suited for distinguishing higher levels of education. For example, we achieve an R^2 of 0.70 in predicting county-level shares of US adults holding a bachelor's degree, while the corresponding R^2 for the percentage with a high school degree is only 0.50. We observe a similar though less pronounced relationship for Mexico, with an R^2 of 0.69 for the share with post-basic education and 0.61 for the percentage completing primary education.

Our focus on a limited number of meaningful features also allows us to study which (groups of) features are most predictive of educational outcomes. In most models, the user density emerges as the single most important predictor of educational outcomes. Twitter penetration features are particularly informative in Mexico, where they alone account for 57 percent of the variation in educational outcomes, compared to 37 percent in the US. Similarly, error and network features appear to be strongly related to education in Mexico ($R^2 = 0.55$ and 0.51, respectively), but less so in the US ($R^2 = 0.42$ and 0.34, respectively). General tweet statistics and topics have consistently high predictive power in Mexico and the United States (R^2 between 0.5 and 0.6).

The main challenge to model performance arises in sparsely populated areas with low Twitter penetration. Accordingly, the population-weighted R^2 for years of schooling is 0.85 for Mexico and 0.70 for the US (compared to 0.70 and 0.65 in our unweighted base model). Similarly, restricting the evaluation sample to areas with at least ten users would increase performance to 0.74 in Mexico and 0.68 in the US. We also explore how model performance evolves depending on the data collection period, finding that we can achieve relatively high predictive power with just three days of tweet data, namely an R^2 of 0.66 for Mexico and 0.58 for the United States.

Using wealth data for Mexico and income data for the US, we further explore how our measure of human capital performs in downstream tasks by comparing regression results based on predicted vs. ground truth measures of education. We find that slope coefficients tend to be biased not only when using the predicted indicator as an independent variable, but also when it acts as the dependent variable. The latter bias results from the typical model tendency to overpredict for low values and underpredict for high values. When using a loss function that penalizes quintile-specific biases (see Ratledge et al., 2022), the bias disappears, and regression coefficients based on our predicted indicator become very similar to their ground truth counterparts. Our simulations show that when appropriately modeled, predicted indicators can produce correct estimates in downstream regression tasks as long as they serve as the outcome and not the treatment variable.

This paper contributes to the recent literature exploring the combined potential of non-conventional data sources and machine learning to measure and understand socio-economic development. While a range of outcomes including wealth (Jean et al., 2016; Blumenstock, Cadamuro, and On, 2015; Yeh et al., 2020; Aiken et al., 2022), population density (Stevens et al., 2015; Wardrop et al., 2018), crop yield (Lobell, 2013; Burke and Lobell, 2017; Sun et al., 2019), informal settlements (Kuffer, Pfeffer, and Sliuzas, 2016; Mboga et al., 2017), electricity access (Ratledge et al., 2022), and disease spread (Wesolowski et al., 2012; Chang et al., 2021) have been accurately predicted using satellite or phone data, previous attempts to infer human capital have been less successful. Head et al. (2017) use satellite data to predict educational attainment in Rwanda, Nigeria, Haiti and Nepal, achieving an average R^2 of $\sim$0.55. The predictive power of other data sources, such as Google Street View images (Gebru et al., 2017) or Wikipedia articles (Sheehan et al., 2019), appears to be even lower, accounting for less than 40 percent of the variation in educational outcomes. A key limitation of these approaches, is that the underlying data does not contain any information on education itself, but only on its economic consequences. We show that by using geolocated Twitter data and natural language processing, we cannot only derive a more accurate indicator of human capital than previous studies but also achieve similar performance to the well-known wealth prediction using satellite data.

We also add to the literature leveraging social media data for social science research. Nearly five billion people worldwide used at least one social media platform in 2023, and another billion are projected to join by 2027, as emerging and developing economies catch up (Poushter, Bishop, and Chwe, 2018; Statista, 2022). Social media data has been used to predict or study diverse outcomes such as migration (Huang et al., 2020; Yin, Gao, and Chi, 2022), social capital (Chetty et al., 2022), censorship (King, Pan, and Roberts, 2013), alcohol consumption (Curtis et al., 2018) or stock market prices (Bollen, Mao, and Zeng, 2011). Moreover, micro-evidence suggests that social media posts are informative about educational characteristics of individual users (Smirnov, 2020; Gómez et al., 2021). This paper goes a step further and shows that despite the high endogenous selection in social media usage (Mellon and Prosser, 2017), the corresponding data can be used to derive accurate education estimates at low administrative units within countries. This underscores the potential of social media data as part of a global tracking system that provides timely and granular information on educational outcomes.

2 Data and Methods

To develop and validate our approach, we focus on two large countries with high-quality ground truth data on educational outcomes: the United States and Mexico. This pairing allows us to test our approach across different languages (English vs. Spanish), development contexts (high income vs. middle income), and, most importantly, levels of Twitter penetration (see Figure 1). While the United States has one of the highest user densities in the Americas, Mexico ranks in the lower middle overall and has the lowest penetration among the five most populous countries in the region. By testing our approach in these diverse and data-rich contexts, we aim to establish a foundation for potential application in regions where traditional survey data are scarce or outdated.

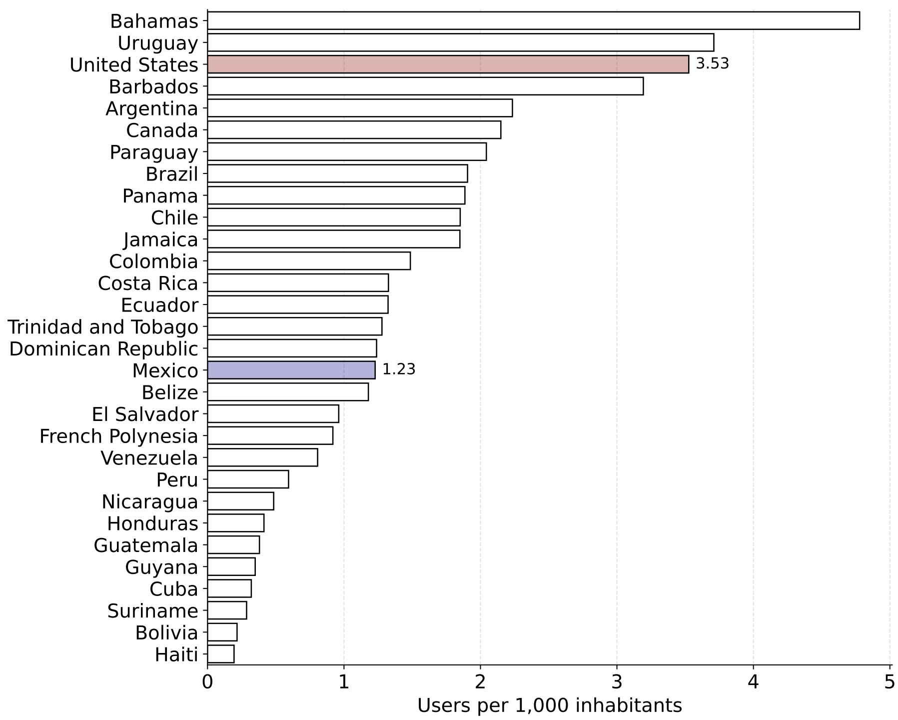


Figure 1: Twitter penetration across the Americas

Geolocated Twitter users per 1,000 inhabitants captured through the Twitter Streaming API in July-August 2021 for countries in North, Central, and South America.

2.1 Collection and Processing of Twitter Data

We used the Twitter Streaming API to compile a large tweet dataset for both countries. Until early 2023, Twitter provided free access to real-time information on 1% of all tweets through its Streaming API, including the text of each tweet and a set of tweet and user characteristics.[1] Our final dataset consists of 2,686,779 geolocated tweets from 123,309 users for Mexico and 22,610,134 tweets from 943,164 users for the United States, collected between July and August 2021.

The tweets included in our final dataset were selected based on three criteria:

1. *Geographical location*: We excluded all tweets that were not posted from within the geographic territory of the respective country. In the case of the United States, we

[1]In February 2023, the free Streaming API was replaced with a paid service. Twitter described the 1% sample provided by its free Streaming API as random, though the exact mechanism was proprietary, not fully transparent, and shown to exhibit systematic biases (Morstatter et al., 2013). Importantly, the generalizability of our approach does not require the sample to be representative of all Twitter users but only consistent across countries and time. While Twitter provided a single global API endpoint, the proprietary nature of the sampling mechanism means that regional differences in coverage cannot be entirely ruled out. However, Morstatter et al. (2013) found that when using geographic bounding boxes as collection parameters — as we do — the Streaming API captured approximately 90% of geotagged tweets rather than the nominal 1% sampling rate, leaving limited room for region-specific variation in sampling coverage.

use all tweets from the mainland, Alaska and Hawaii, but not from unincorporated territories such as Puerto Rico or the Virgin Islands. We also exclude tweets without precise location information (i.e., less precise than municipality/county level). Our final sample comprises tweets with exact coordinates (MX: 3%, US: 3%), neighborhood or point of interest (poi)-level precision coordinates (MX: 2%, US: 2%), and city-level precision coordinates (MX: 95%, US: 94%).

2. *Language*: For each country, only tweets written in the primary native language (i.e., Spanish for Mexico and English for the United States) are included.

3. *Source*: A major concern regarding the reliability of Twitter data is that many tweets are automatically disseminated through APIs rather than individually created by a human user. We thus restrict our sample to content that is posted through the four main channels for human users: iPhone, Android, iPad, and Instagram.[2] This excludes tweets generated through third-party APIs from platforms such as Foresquare or CareerArc (approximately 1 percent of geolocated tweets in Mexico and 7 percent in the United States).

To compute municipality- or county-level statistics, we follow a three-step procedure. First, each tweet is assigned to a geographical unit (i.e., municipality or county) based on its coordinate data. While this is straightforward for exact coordinates, we apply different types of consistency checks to find the correct unit when coordinate information consists of a bounding box at the city, poi, or neighborhood level.[3]

Next, we approximate the home municipality or county for each user. If users tweet from more than one geographical entity (MX: 33% of users, US: 35% of users), we assign all their tweets to the entity from which they tweeted the most. For users whose tweet counts are equally divided among two or more entities (MX: 1%, US: 2%), we use the number of tweets posted during non-work hours on weekdays as a tiebreaker. This procedure results in the reassignment of 14 percent of tweets in Mexico and 12 percent of tweets in the United States. Tweets that cannot be unambiguously assigned to a municipality through this procedure are dropped (MX: 0.4%, US: 0.2%).

[2]For tweets posted through Instagram, we exclude all tweets that use the default text ("Just posted a photo @...") rather than a message specified by the user. Tweets posted through the Twitter website are not included in our sample becuase they do not have any associated coordinates.

[3]In most cases, assignment to the geographical unit containing the centroid of the tweet's bounding box yielded correct results. However, particularly in the Mexican case, where the location precision for tweets tends to be lower (and city-level precision as defined by Twitter refers to municipalities rather than places within municipalities), we combine spatial joins with name matching to ensure that all tweets are assigned to the correct entity.

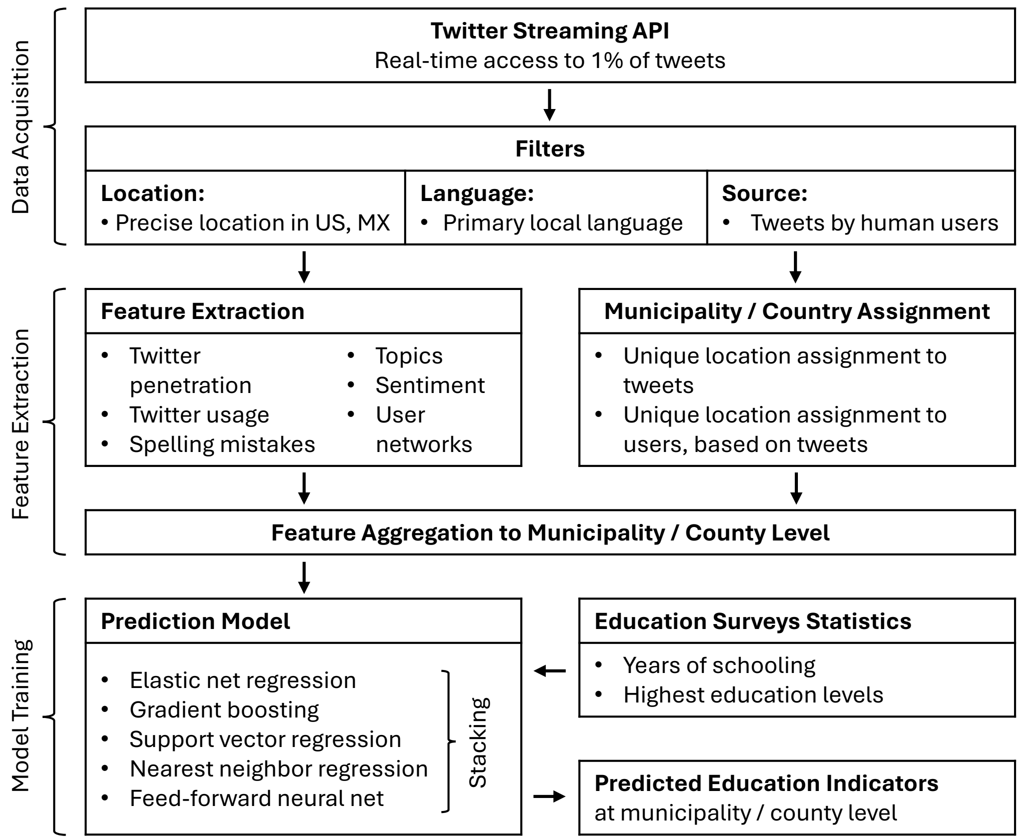


Figure 2: Data Extraction and Model Training Pipeline

Finally, data is aggregated at the municipality or county level using the unit-level sum, mean, or median depending on the distribution of the underlying variables (for details, see Section 2.3 and Appendix C). To give equal weight to all users irrespective of their degree of activity, all tweet-level variables are first aggregated at the user level.

2.2 Survey Data

While many countries lack timely and spatially disaggregated information on educational outcomes, such data are available for both Mexico and the United States, allowing us to train and test a prediction algorithm in two different settings. Our main outcome variable is years of schooling for both countries, but we also look at the share of adults holding different educational degrees to better understand where in the educational distribution our models work best (see Table D.1). We use data from the 2020 census for Mexico and from the American Community Survey (2017–2021, 5-year estimates) for the United States, meaning that outcomes for both countries are temporally closely aligned with our input features from 2021.[4] Following Barro and J. W. Lee (2013), we approximate county-level years of schooling for the US based on the proportion of the population holding different educational degrees and the average years of schooling associated with these degrees.[5]

Section C in the Appendix presents summary statistics on all outcome variables. In the average Mexican municipality, 28 percent of the population holds a post-basic degree, 54 percent has graduated from secondary school, 76 percent has finished primary school, and the average person has completed 7.8 years of schooling. The corresponding figures in US counties are 23 percent with a bachelor degree, 54 percent with some college, 88 percent with a high school degree, and 13.3 years of schooling.[6]

2.3 Features

Our feature matrix comprises municipality-level information on *(i)* Twitter penetration, *(ii)* Twitter usage, *(iii)* spelling mistakes, *(iv)* topics, *(v)* sentiment, and *(vi)* user networks (for a detailed overview, see Section C and D in the Appendix). In addition, we also include population density estimates.[7] To advance our understanding of the aspects of people's online behavior that are most predictive of human capital, we deliberately focus on a limited number of interpretable features rather than, for example, using tweet text embeddings. Importantly, the included features vary in how directly they capture education. In contrast to previous satellite-based approaches, several of our indicators are direct expressions of education (e.g., spelling mistakes, grammatical errors) or closely linked to educational attainment (e.g., different topics) rather than reflecting only the economic consequences of education. Nevertheless, other features such as Twitter penetration, usage patterns, and network characteristics are related to both education and broader socioeconomic conditions, meaning that our predictions partly capture regional

[4]The Mexican census data is publicly available at `https://www.inegi.org.mx/datosabiertos/`, while data from the American Community Survey can be accessed at `https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/`.

[5]Average years of schooling for a given county are computed using $\sum_j h_j \, Dur_j$, where h_j indicates the the proportion of the population that has attained education level j and Dur_j indicates the corresponding duration to attain level j. We use data from the *Current Population Survey*, specifically the *2021 Annual Social and Economic (ASEC) Supplement*, to compute estimates for Dur_j. In Mexico, this approximation is not necessary because average years of schooling are included in the census data.

[6]MX: Estimates for years of schooling, primary and secondary completion are provided for the population aged 16 or more, while the share with post-basic education is defined for adults (i.e., over 18). US: All education statistics refer to the population aged 25 or older.

[7]Population data is globally available; thus, its inclusion does not limit the external validity of our approach. Population data is also necessary for the computation of tweet and user densities. A model using only population estimates will serve as our benchmark against which the performance of our approach is compared.

Table 1: Summary statistics by education level for selected features

	Mexico			United States		
	Bottom 25%	Top 25%	All	Bottom 25%	Top 25%	All
User density	0.23	0.86	0.47	0.79	2.49	1.45
Tweet density	1.94	16.70	7.00	12.03	45.14	24.40
Tweet length	68.75	72.88	69.94	77.09	82.05	80.86
Account age	5.03	6.34	5.67	6.67	7.51	7.06
Tweets per year	1,306.55	362.71	841.93	648.93	351.58	495.19
Favorites per tweet	5.02	1.34	3.76	1.52	2.14	1.73
Error total	24.60	23.54	25.28	15.23	13.14	13.87
Error grammar	0.17	0.15	0.17	0.65	0.47	0.55
Error typos	12.18	10.66	12.47	7.48	6.92	7.19
Topic science	1.84	1.92	1.87	1.58	1.82	1.69
Topic relationships	6.66	5.72	6.27	5.31	4.42	4.76
Sentiment positive	0.39	0.37	0.38	0.50	0.50	0.50
Offensive language	0.15	0.16	0.15	0.17	0.16	0.16
Network clos. centr.	0.06	0.31	0.16	0.28	0.42	0.34
Number of Areas	430	429	1,714	723	723	2,889

Municipality (MX) or county (US) averages for selected features by educational outcome. The bottom 25% and top 25% refer to the municipalities/counties in the lowest or highest quartile of years of schooling. Only areas with at least one tweet are included. Features are not log transformed.

socioeconomic outcomes correlated with education. The implications of this are discussed in detail in Section 3.5.

Twitter penetration data (4 features) consists of the total number of tweets and users as well as the number of users and tweets relative to the population (referred to as user and tweet densities). We further include general information on *Twitter usage* (11 features), such as the average tweet length, the number of followers, the user mobility, the account age, the number of emojis per tweet, or the share of tweets posted during work hours or from an iPhone. To obtain estimates for the frequencies of different *spelling mistakes* (MX: 23 features, US: 16 features), we use a Python wrapper for "LanguageTool", an open source grammar, style, and spell checker. LanguageTool is available in over 25 languages, including English and Spanish, and classifies the detected errors into different categories such as grammar, typos, casing, punctuation, or style.[8] We include the total number of errors per 1,000 characters and the corresponding numbers for each category. To determine the *topics* of each tweet (19 features), we use a pre-trained multi-label tweet classification model (Ushio and Camacho-Collados, 2022). This allows us to estimate the probability a given tweet is about a particular topic, such as news, celebrity, sports, or science. Since no pre-trained tweet classification models are available in Spanish, we

[8]See `https://dev.languagetool.org/languages` for information on language availability.

translate all Spanish tweets into English using a pre-trained model based on the Marian NMT framework (Junczys-Dowmunt et al., 2018) to determine the topic distributions of our Mexican tweets.[9] A further group of inputs comprises features related to *sentiments* (4 features), such as the share of tweets with negative or positive sentiments, offensive language, or hate speech. They are generated using pre-trained classification models for Spanish and English tweets.[10] Finally, we also add *network indicators* (4 features), such as degree and closeness centrality. For this purpose, we use quotes and mentions to construct a user-to-user network and subsequently aggregate this network to the municipality or county level. We take the log of right-skewed features and standardize all features before training.[11]

To address potential problems related to sparse or noisy data in areas of low population density, we develop a procedure that allows our model to learn from spatial neighbors. For each unit (i.e., municipality or county), we create a cluster consisting of the focal unit and all its spatial neighbors, and compute cluster-level estimates for each of our features. We use this information about Twitter usage in the broader area around each unit in three ways: First, we add the cluster-level estimates as additional inputs to our feature matrix (i.e., for each unit and measure, we include both unit- and cluster-level values). Second, we use cluster-level features to impute missing values in units without tweets using an elastic net regression model. This provides estimates for features that cannot be observed in the absence of tweets, and is necessary as most machine learning algorithms cannot handle missing values. Third, in units with fewer than 5 tweets, we replace extreme outliers with imputed values using the same imputation procedure.[12]

Table 1 shows the mean of selected features by educational level for both countries (see Section C in the Appendix for complete summary statistics). This simple inspection already reveals a strong correlation between Twitter features and educational outcomes. In both countries, user and tweet density is markedly higher in places with more educated populations. Similarly, users in more educated areas tend to write longer tweets, make fewer errors, and talk about different topics (e.g., science rather than relationships). On the other hand, users in less educated areas tweet more actively.

[9] The model is provided via the HuggingFace library: `https://huggingface.co/docs/transformers/model_doc/marian`.

[10] The classification models are provided by the same library used for the topic classification above.

[11] Appendix D documents which variables are log-scaled. Following Stahel (2000), we use $log(x + c)$ to deal with zeros, with x as the values of a particular feature and $c = Q_{0.25}^2 \,/\, Q_{0.75}$, where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles based on feature values $x > 0$.

[12] Extreme outliers are defined as values that are lower than $Q_{0.25} - 3\,IQR$ or higher than $Q_{0.75} + 3\,IQR$, with $Q_{0.25}$ and $Q_{0.75}$ as the first and third quartiles and IQR as the interquartile range.

2.4 Training and Evaluation

To train our models, we use a stacking regressor combining five machine learning algorithms: *(i)* elastic net regression, *(ii)* gradient boosting, *(iii)* support vector regression, *(iv)* nearest neighbor regression, and *(v)* a feed-forward neural network (i.e., a multilayer perceptron). We use cross-validated grid search to tune the hyperparameters of each model. The performance of the final stacking regressor is evaluated using five-fold cross-validation. This procedure is known as nested cross-validation. We report the cross-validated R^2 for each fold as well as an overall R^2 obtained by combining all cross-validated predictions, where R^2 represents the coefficient of determination.[13]

3 Results

3.1 Main Results

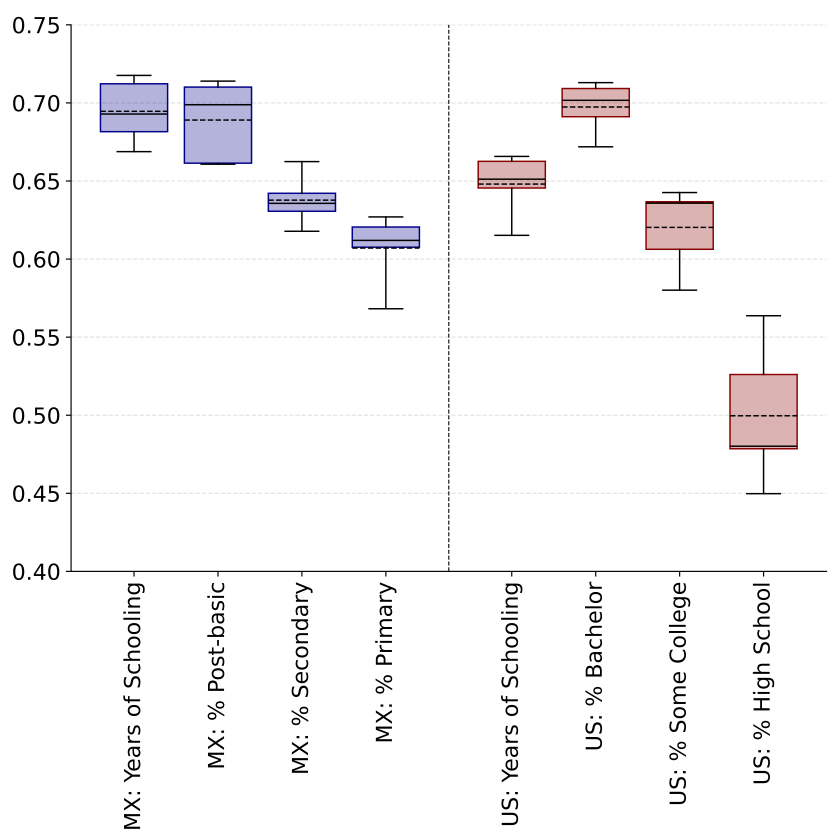
Our final model is able to account for 70 percent of the variation in years of schooling in Mexican municipalities and 65 percent in US counties (see Figure 3). Population-weighted performance estimates are even higher, reaching an R^2 of 0.85 in Mexico and of 0.70 in the United States.[14] A closer look at the predictive power for different educational degrees reveals substantial variation in model performance in both countries.

In Mexico, we report an R^2 of 0.69 for the share of the population holding a post-basic degree (i.e., high school or more), an R^2 of 0.64 for the corresponding share with a secondary degree, and an R^2 of 0.61 when aiming to predict the prevalence of primary school completion. Differences are even more pronounced in the United States, where our model captures 70 percent of the variation in the percentage of adults that hold a bachelor's degree, 62 percent for the share that went to college, and 50 percent when focusing on high school completion. This suggests that Twitter data is particularly informative about higher education levels and less sensitive to differences at the lower end of the education distribution. This pattern is likely driven by selection into platform usage, as more educated populations are more likely to use Twitter and generate sufficient data for reliable predictions.

Among the five included models, gradient boosting and support vector regression perform best and, accordingly, receive the highest weights in the final stacking regressor (see Figure A.1 and Table A.1 in the Appendix). The neural network and the nearest neighbor regressor, on the other hand, perform rather poorly, achieving a lower predictive

[13] $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$, where y_i are observed values, $\hat{y}_i$ are predictions, and $\bar{y}$ is the observed mean.

[14] Population weights are not taken into account during training.



MX	R^2
Years of schooling	0.695
	(0.847)
% Post-basic	0.690
	(0.820)
% Secondary	0.638
	(0.815)
% Primary	0.608
	(0.809)

US	R^2
Years of schooling	0.648
	(0.697)
% Bachelor	0.698
	(0.753)
% Some College	0.621
	(0.683)
% High School	0.499
	(0.589)

Figure 3: Performance for different educational outcomes

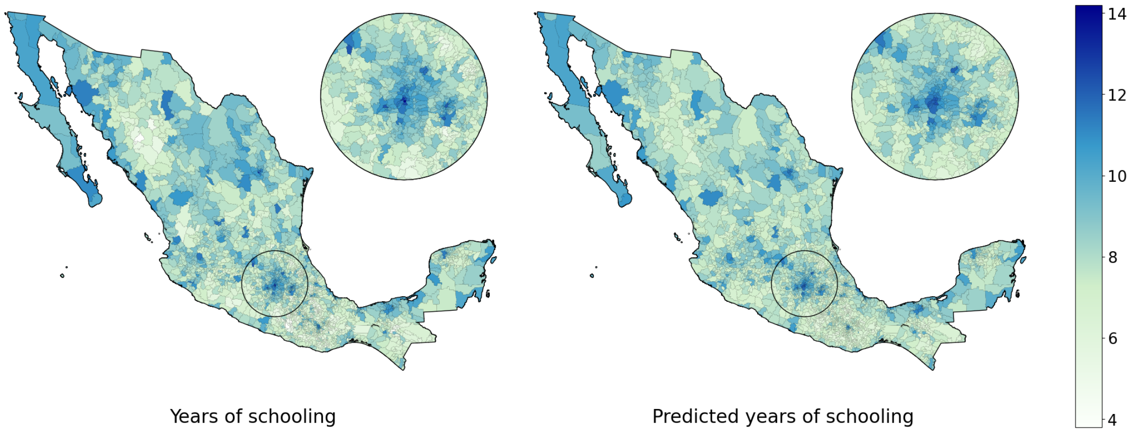
Model performance results for different educational outcomes in Mexican municipalities (blue) and US counties (red). All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the R^2 across validation folds for each outcome and country. The table on the right presents the R^2 based on out-of-sample predictions for the full data sets (stacked across folds). Population-weighted R^2 are presented in parentheses.

power than the simple elastic net model (i.e., a regularized linear model). For all outcomes, the ensemble of all models outperforms the best-performing individual model, highlighting the benefits of stacking.

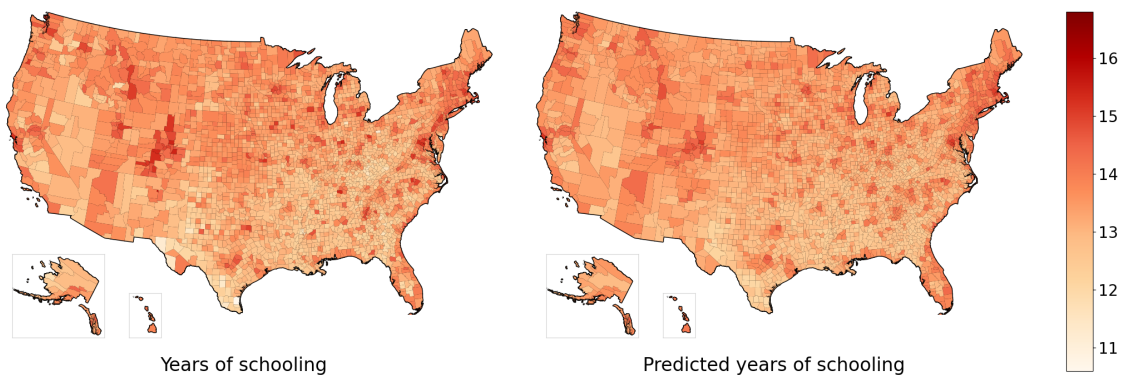
As Figures 5a and 5b show, our model produces the attenuated predictions that are typical for continuous outcomes (Ratledge et al., 2022), meaning that, on average, estimates are too high in low-education and too low in high-education areas.[15] This pattern also becomes apparent when comparing maps of true and predicted years of schooling (see Figures 4a and 4b). While spatial patterns look very similar for the two measures, they are slightly less fine-grained in the prediction maps.

[15]The regression line in Figure 5 and Appendix Figure A.2 does not take population weights into account. The fact that there are many sparsely populated areas at the lower, and few, but very populous areas at the higher end of the education distribution, creates the illusion that the line does not fit the data.

(a) Predictions for Mexico



(b) Predictions for the United States



(c) Prediction Error

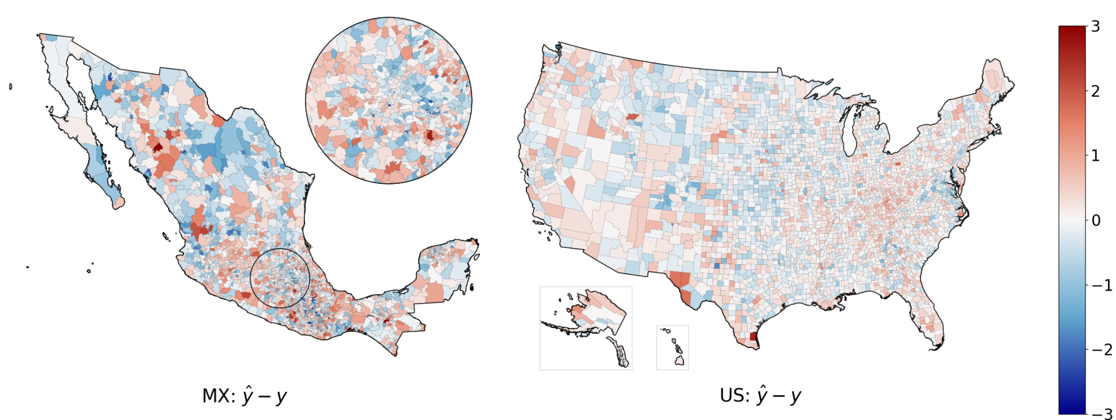


Figure 4: Maps of true vs. predicted years of schooling

Predicted values for all municipalities and counties are obtained by combining out-of-sample predictions from all folds. In Figure 4c, red indicates overprediction and blue underprediction of true values.

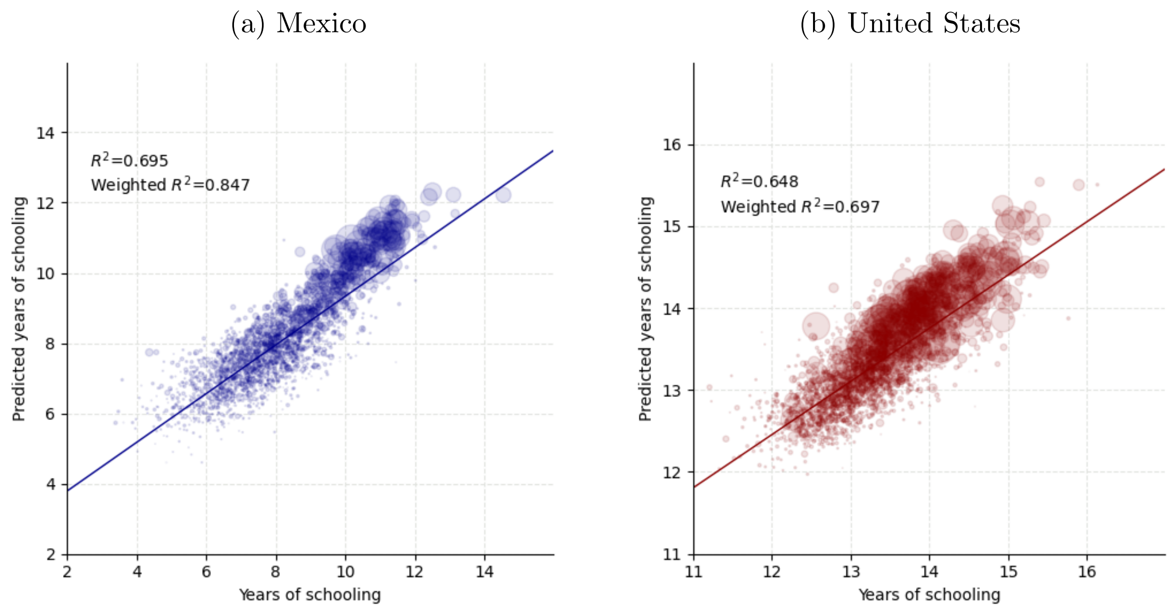


Figure 5: True vs. predicted years of schooling

Predicted values for all municipalities and counties are obtained by combining out-of-sample predictions from all folds. Bubble size is proportional to the population in each unit. R^2 and population weighted R^2 shown. The line indicating the best linear fit is not population-weighted.

3.2 Robustness and Generalizability

To assess the *reliability* of our predictions, we quantify uncertainty at two levels. First, the boxplots in Figure 3 show variation in model performance across different cross-validation folds, indicating relatively stable performance with modest variation for most outcomes. Second, we assess prediction uncertainty at the unit level by re-running the full cross-validation procedure with different random fold assignments (see Appendix Section A.5 for details). We find that prediction uncertainty is generally higher in Mexican municipalities than in US counties, likely reflecting the larger underlying variance in years of schooling in Mexico. Within both countries, uncertainty is substantially higher in municipalities or counties with smaller populations, fewer Twitter users, and greater reliance on spatial imputation (Figures A.8 and A.9). These patterns closely mirror those of prediction errors (Figure A.10), and both uncertainty and errors exhibit spatial clustering (Figures 4c and A.8).

We also assess the *spatial generalizability* of our approach. To this end, we conduct leave-one-state-out cross-validation (e.g., Breen et al., 2025; Chi et al., 2022), where models are trained excluding all municipalities or counties from one state at a time and then used to predict outcomes in the held-out state. This tests whether the model can generalize not only to unseen municipalities or counties, but to entirely new geographic areas with potentially different socioeconomic contexts. In line with expectations and previous research, this more stringent test of spatial transferability yields lower performance estimates than our main five-fold cross-validation (Figures A.5 and A.6). Yet, performance remains strong, with unweighted R^2 for years of schooling of 0.57 in Mexico and 0.59 in the United States, and population-weighted R^2 of 0.79 and 0.69, respectively.

Finally, we assess *temporal generalizability*. Given the substantial changes in Twitter's data accessibility and user composition since our data collection period (Özturan et al., 2025; Nutakki et al., 2025; Robertson, 2023), we examine whether our 2021 Twitter data remain predictive of more recent educational outcomes. For the United States, we use updated data from the most recent American Community Survey (2019-2023 compared to 2017-2021) to re-evaluate our models. Comparable updated data are not yet available for Mexico (most recent census: 2020, which we already use in our main analysis). Performance is very similar across all outcomes with declines of 0-2 percentage points compared to our main estimates (Figure A.7). For years of schooling, we report an R^2 of 0.64 and a population-weighted R^2 of 0.69. This stability suggests that social media data can provide reliable predictions even when outcome data extend beyond the data collection period, likely reflecting the slow-moving nature of educational attainment.

3.3 Feature Importance

As our model is based on a limited number of interpretable inputs (see Sections C and D in the Appendix), we can explore how important various types of features are to the success of our approach. Figure 6 shows how different groups of features perform on their own. A model using only population data serves as a benchmark and achieves an R^2 of 0.48 for Mexico and 0.34 for the United States. Simple Twitter penetration data, that is, user and tweet densities or counts, already outperforms the population model, with R^2 values of 0.57 for Mexico and 0.36 for the United States. Particularly in Mexico, knowing where people tweet is thus more informative about human capital concentration than knowing where people live.

The performance of usage statistics, i.e., features such as the average tweet length or the number of followers, is high in both countries, accounting for 55 to 58 percent of the variance in educational outcomes. The same is true for topic variables, which reach an R^2 around 0.5 in both countries. Error and network statistics, on the other hand, seem to be much more strongly related to education in Mexico (R^2 of 0.55 for errors and 0.51 for networks) than in the United States (R^2 of 0.42 for errors and 0.34 for networks). Finally, sentiment features are the only group of variables that fails to surpass the benchmark

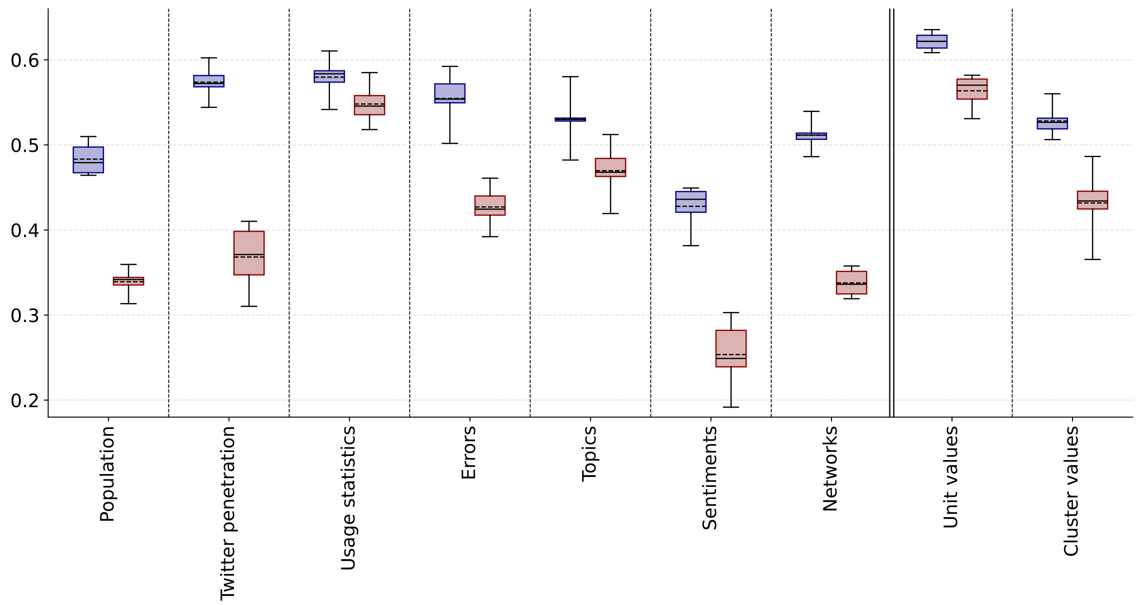


Figure 6: Performance of feature subgroups

Performance of feature subgroups for Mexico (blue) and the United States (red): Population (2x4 features, i.e., 4 at the unit level and 4 at the cluster level), Twitter penetration (2x4 features), usage statistics (2x11 features), spelling mistakes (MX: 2x23 features, US: 2x16 features), topics (2x19 features), sentiment (2x4 features), and networks (2x4 features), as well as all features at the unit level (i.e., municipality or county) and all features at the cluster level (i.e., including spatial neighbors). All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the R^2 across validation folds for each outcome and country. The outcome is years of schooling in all models.

model. We can also evaluate how our model benefited from including cluster-level features (see Figure 6). When limiting ourselves to unit-level features, we report R^2 values of 0.63 (MX) and 0.56 (US), as opposed to 0.70 (MX) and 0.65 (US) for the full model.[16] Thus, exploiting information from spatial neighbors is critical to the predictive power of our models. Overall, the performance of no single group of features comes close to that of the overall model, suggesting that the different inputs are complementary.

To better understand these complementarities, we use SHAP (SHapley Additive ex-Planations) to estimate marginal contributions (Lundberg and S.-I. Lee, 2017). This approach is based on cooperative game theory (Shapley, 1953) and allows to compute importance scores for different features and feature groups (see Appendix A.2). When looking at the contributions of individual Twitter features, the user density emerges as the single important predictor in the majority of models (i.e., models for different countries and educational outcomes). It is the most important feature in five models (US: 4, MX: 1) and the second most important in the remaining (MX: 3). The importance

[16]This provides a lower bound for the true benefit of exploiting spatial information as cluster-level features are also used to impute missing values and extreme outliers.

of other features varies more strongly between countries, but topics such as sports, simple usage statistics including the tweet length or the account age, as well as network closeness centrality (only US), tend to be highly predictive too. When the SHAP values are aggregated into feature groups similar to Figure 6, we observe that the SHAP-based feature importance ranking closely resembles the ranking based on group-specific model performance.[17]

3.4 Performance Heterogeneity

We now explore how our model is affected by the limited number of tweets in sparsely populated areas (Figure 7). In line with expectations, performance is substantially higher when limiting the evaluation to municipalities or counties with more tweets or users. This relationship is even more pronounced when looking at different population thresholds. Particularly in Mexico, model performance increases drastically if we exclude smaller municipalities, where both input and output data is likely to be more noisy. This is consistent with finding that, in both countries, the population-weighted R^2 is substantially higher than the unweighted R^2 for all outcomes.

It is also informative to look at performance by the amount of data we use to make the predictions. For our main analyses, we streamed Twitter data for two months, and used millions of tweets to construct municipality or county-level indicators. To see if similar results can be achieved with a shorter data collection period, we re-run the entire feature engineering and model training procedure on different subsets of our data. As Figure 8 shows, drastically shortening the data collection period only marginally reduces performance. This is especially true for Mexico, where one day of tweets already yields an R^2 of more than 0.65. In the US, on the other hand, it takes about week of Twitter data to account for 60 percent of the variation in county-level education outcomes. As the curves for both countries flatten out almost completely after a few weeks, extending the data collection period beyond two months is likely to yield only negligible additional performance gains.

[17] Aggregations are the sum of absolute SHAP values (see Figure A.3). Feature importance rankings show high stability across cross-validation folds. Spearman rank correlations are $\rho = 0.69$ for Mexico and $\rho = 0.81$ for United States, and the most important feature overall ranks first in all folds for both countries (see Appendix Section A.2 for details).

[18] Standard errors (shaded area) are computed using $\sqrt{\frac{4r^2(1-r^2)^2(n-k-1)^2}{(n^2-1)(n+3)}}$, where n is the sample size and k is the number of features (Cohen et al., 2013).

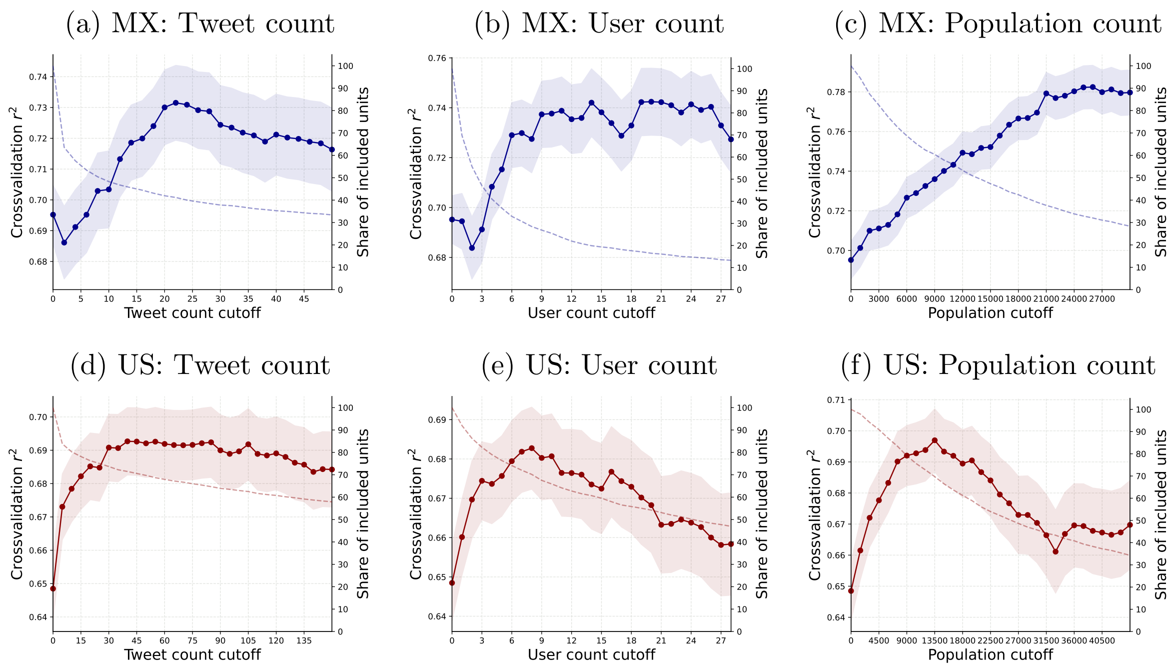


Figure 7: Performance heterogeneity

Performance heterogeneity by user, tweet, and population count for Mexican municipalities (blue) and US counties (red). The *solid line* shows the R^2 (including standard errors) for units (municipalities or counties) above different tweet, user, or population count cutoffs.[18] The *dashed line* shows the proportion of units included at each cutoff.

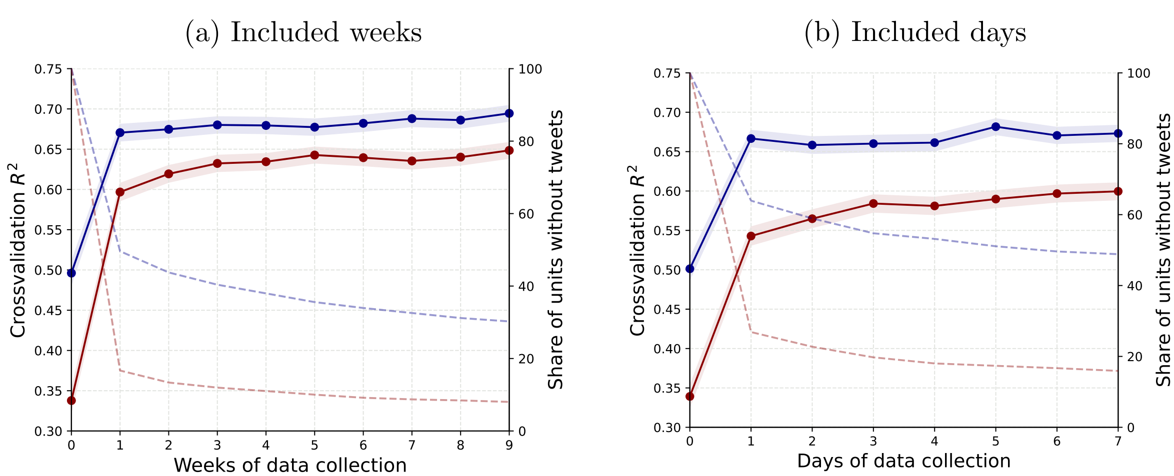


Figure 8: Performance by data collection period

Model performance for shorter data collection periods for Mexican municipalities (blue) and US counties (red). The value for 0 weeks/days corresponds to an R^2 of our baseline model using population data only. Standard errors are computed using the same formula as reported in Figure 7.

3.5 Downstream Performance

In addition to being directly useful to better understand local patterns in development outcomes and target interventions accordingly, predicted measures may also serve to study relationships with other variables. Using wealth data for Mexico and income data for the United States (see Appendix Table D.1), we thus explore how our Twitter-based indicator performs in downstream regression tasks. The fact that machine-learning-derived indicators are noisy measures gives rise to several potential biases that can compromise such applications. If *edu* is the true distribution of the indicator we predicted as $\widehat{edu}$ (e.g., years of schooling), and *econ* is another variable whose relationship to *edu* we wish to study (e.g., wealth), three types of measurement error may occur (see simulations in Appendix Figure B.1):

1. Attenuation bias: A random measurement error in $\widehat{edu}$ will dilute the correlation between *edu* and *econ*. This results in an attenuation bias when regressing *econ* on $\widehat{edu}$, but not in the opposite specification, and decreases precision in both cases (see, e.g., Fuller, 1987).

2. Berkson-type error: A bias that has only recently gained attention (see Ratledge et al., 2022) arises when measurement errors are correlated with *edu*. The typical behavior of machine learning models is to overpredict for low values and underpredict for high values, a pattern that is very apparent in our application, where the correlation between the prediction error (i.e., $\widehat{edu}$ - *edu*) and *edu* amounts to about -0.6. This does not affect the correlation between *edu* and *econ*, but it distorts coefficients in downstream regressions. Specifically, it leads to a downward bias when $\widehat{edu}$ is used as the outcome variable, and to an upward bias when it acts as the explanatory variable.

3. Correlated learning: If the features used to predict $\widehat{edu}$ contain wealth or income-related information, our model might exploit the correlation between *econ* and *edu* to make better predictions. Given that some of our features, such as Twitter penetration, usage and network features, are likely related to both education and broader socioeconomic conditions, this is to be expected in our setting. Indeed, our feature matrix is almost as predictive of economic outcomes ($R^2 = 0.64$ for wealth in Mexico and $R^2 = 0.62$ for income in the US) as of education.[19] This creates an artificially

[19] This is substantially higher than a model using education only (years of schooling) for the prediction (MX: 0.57, US: 0.50), suggesting that our feature matrix indeed contains wealth and income-related information that is independent of education levels. Estimates are based on re-running the same machine learning procedure we use to predict education for wealth and income.

strong correlation between $\widehat{edu}$ and $econ$. When using $\widehat{edu}$ as the dependent variable, this only leads to overoptimistic standard errors. If $\widehat{edu}$ is the independent variable (and edu and $econ$ are positively correlated), it additionally induces an upward bias for the point estimate.

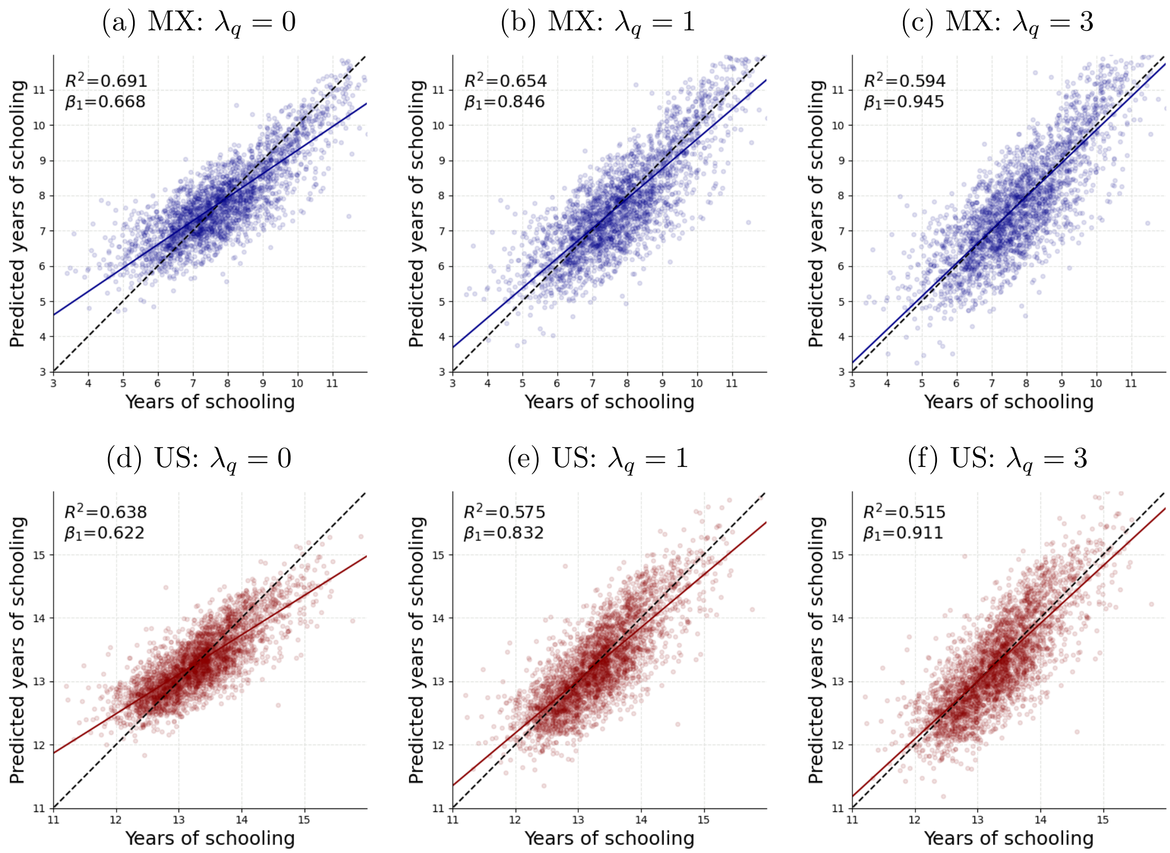


Figure 9: Effects of Berkson-type error correction

True vs. predicted values with correction of the Berkson-type error for Mexican municipalities (blue) and US counties (red). For the correction, we apply an adjusted loss function in the final ridge regression model that performs the stacking. Following Ratledge et al. (2022), we add an additional penalty term to the standard loss function of the ridge regression, which consists of the mean squared error (MSE) plus an L_2 penalty. The adjusted loss function is thus $MSE + \lambda_l L_2 + \lambda_q Q_{bias}$, where λ_q is the strength of the additional penalty and a hyperparameter that can be tuned. Q_{bias} is the maximum of the squared quintile-specific biases, equal to $\max_j(\mathbb{E}[\hat{y}_i - y_i | y_i \in Q_j]^2)$, where $Q_j \in \{Q_1, \dots, Q_5\}$, and $\hat{y}_i$ is the predicted y for observation i. The figure shows the effect of three λ_q parameters on the prediction bias. Solid lines indicate the best linear fit of each model, while dashed black lines represent the expected fit without bias ($\beta_1 = 1$).

With these considerations in mind, we now compare the downstream correlations (Appendix Figure B.2) and regression results (Table 2) of $\widehat{edu}$ and $econ$ with the true correlations captured by edu. As Figure B.2 in the Appendix shows, the predicted education indicator consistently understates true correlations, suggesting that the attenuation bias dominates over a potential bias due to correlated learning. Table 2 further shows that the slope of the regression coefficients is considerably underestimated for all outcomes when using $\widehat{edu}$ as the dependent variable of the regression and slightly overestimated in the reverse specification, a pattern that is consistent with a Berkson-type error. Hence, it appears that the correlation estimates are mainly affected by attenuation, while biases in regression coefficients are largely driven by a Berkson-type error.

Table 2: Downstream regression results

	Mexico				United States			
	Years of Schooling	Post-Basic	Secondary	Primary	Years of Schooling	Bachelor	College	High School
β_t: $edu \sim econ$	0.740 (0.014)	0.661 (0.015)	0.703 (0.014)	0.728 (0.014)	0.692 (0.013)	0.707 (0.013)	0.655 (0.013)	0.487 (0.016)
β_p: $\widehat{edu} \sim econ$	0.549 (0.013)	0.499 (0.014)	0.526 (0.012)	0.516 (0.012)	0.496 (0.011)	0.526 (0.012)	0.470 (0.011)	0.320 (0.011)
β_c: $\widehat{edu}_c \sim econ$	0.748 (0.017)	0.651 (0.018)	0.744 (0.018)	0.687 (0.016)	0.699 (0.016)	0.727 (0.017)	0.661 (0.018)	0.362 (0.013)
$\beta_t - \beta_p$	-0.191 (0.012)	-0.161 (0.011)	-0.177 (0.012)	-0.212 (0.014)	-0.196 (0.011)	-0.181 (0.012)	-0.185 (0.011)	-0.167 (0.012)
$\beta_t - \beta_c$	0.008 (0.014)	-0.010 (0.013)	0.041 (0.015)	-0.042 (0.016)	0.007 (0.014)	0.021 (0.016)	0.005 (0.017)	-0.125 (0.014)
β_t: $econ \sim edu$	0.740 (0.014)	0.661 (0.015)	0.703 (0.014)	0.728 (0.014)	0.692 (0.013)	0.707 (0.013)	0.656 (0.013)	0.488 (0.016)
β_p: $econ \sim \widehat{edu}$	0.794 (0.018)	0.717 (0.019)	0.826 (0.019)	0.863 (0.019)	0.765 (0.017)	0.738 (0.017)	0.767 (0.018)	0.640 (0.023)
β_c: $econ \sim \widehat{edu}_c$	0.577 (0.013)	0.539 (0.015)	0.564 (0.013)	0.646 (0.015)	0.535 (0.012)	0.520 (0.012)	0.443 (0.012)	0.515 (0.019)
$\beta_t - \beta_p$	0.054 (0.012)	0.056 (0.012)	0.123 (0.013)	0.135 (0.014)	0.072 (0.015)	0.031 (0.014)	0.111 (0.014)	0.152 (0.025)
$\beta_t - \beta_c$	-0.162 (0.010)	-0.122 (0.011)	-0.139 (0.011)	-0.083 (0.012)	-0.157 (0.013)	-0.187 (0.012)	-0.212 (0.012)	0.028 (0.024)
N	2,457	2,457	2,457	2,457	3,140	3,140	3,140	3,140

The predictions for different educational outcomes, referred to as edu, are denoted as $\widehat{edu}$, and $econ$ is wealth for Mexico and income for the United States. For $\widehat{edu}_c$, we apply a Berkson error correction with $\lambda_q = 3$ for years of schooling and $\lambda_q = 15$ for all other outcomes (i.e., all percentages). Results are reported in standard deviations ($\widehat{edu}$ and $\widehat{edu}_c$ are standardized using the distribution of edu). $\beta_t - \beta_p$ is the original bias and $\beta_t - \beta_c$ is the bias using the predictions based on the adapted loss function. Education is the dependent variable in the top panel and the independent variable in the bottom panel. Standard errors in parentheses.

While the attenuation bias and correlated learning cannot be avoided in most applications, it is possible to refine our model in a way that minimizes the Berkson error. Following Ratledge et al. (2022), we add a further penalty term for a quintile-specific bias to the loss function of our final stacking model. If the weight given to this penalty is sufficiently high, the tendency to understate high values and overstate low values effectively disappears (see Figure 9), albeit at the cost of reduced overall performance, with a decrease in the R^2 by about 10 percentage points. Using this new set of predictions (see Table 2), the bias in the top panel ($\widehat{edu} \sim econ$) becomes negligible for most outcomes.[20] In the bottom panel ($econ \sim \widehat{edu}$) the direction of the bias is reversed as the attenuation bias begins to dominate. This suggests that when appropriately modeled, predicted indicators can yield accurate estimates in downstream regression tasks, provided they serve as the outcome rather than the treatment variable. Fortunately, the former scenario is more common, as it enables the evaluation of interventions or policy changes.

4 Conclusion

Our results show that human capital can be accurately inferred from Twitter data using machine learning. We are able to account for 70 percent of the variation in years of schooling in Mexico and 65 percent in the United States. This is substantially higher than the performance reported in previous attempts to predict human capital, and comparable to the effectiveness of satellite data in predicting wealth. As only a few days of Twitter data are needed to achieve a good performance and the natural language processing tools we use for feature preparation support many different languages, our approach has substantial potential for broader application. In addition, despite the lower Twitter penetration, our model tends to perform better for Mexico than for the United States, suggesting that the method is also relevant to less affluent regions with lower levels of social media usage.

In addition to being directly useful for understanding spatial patterns and targeting interventions, predicted indicators also have the potential to advance scientific research by providing inputs for downstream inference tasks. This paper highlights that such applications are not without caveats. Our data and simulations show that estimates in downstream regression tasks tend to be subject to several biases. We further demonstrate that these biases can be corrected using an adapted loss function (see Ratledge et al., 2022) if the predicted indicator serves as the dependent variable. If carefully tuned, machine-learning-derived indicators can thus become a valuable data source to study effects on outcomes for which ground truth data are unavailable. However, more research is needed to better understand the empirical relevance of each of the biases, and experiment with the most effective ways to address them.

[20] The bias becomes insignificant for 5 out of 8 outcomes. The correction appears to be particularly effective for outcomes that have a higher initial R^2. In the last model (high school), which is also the one with the lowest initial R^2, the penalized loss function achieves only a limited slope correction with $\lambda_q = 15$ (not shown) and the regression is thus unable to recover the true effect.

These contributions notwithstanding, three important limitations should be acknowledged. *First*, we report performance of our approach for a specific platform at a specific point in time, namely Twitter in 2021. Twitter (now X) data has since changed substantially, with tighter restrictions and higher costs for large-scale data collection. In parallel, the platform has undergone shifts in governance, user composition, and posting behavior (Özturan et al., 2025; Nutakki et al., 2025; Robertson, 2023). While the data we use remain highly predictive of today's education levels, real-time prediction or tracking changes over time would require recalibrating the model to the current Twitter environment or to a different platform. *Second*, our evaluation is limited to two countries. Although the United States and Mexico differ substantially in economic development, language, and Twitter penetration, we cannot make precise predictions regarding transferability to other countries, especially those with very low platform adoption. The strong performance in Mexico, which has relatively low Twitter penetration compared to other countries in the Americas, suggests our approach may be applicable beyond high-penetration contexts, but further research is needed to confirm this. *Third*, within countries, Twitter data is less informative at the lower end of the education distribution and in less populated areas with lower Twitter penetration. This likely reflects that Twitter use is concentrated among the highly educated, making the platform less suited for distinguishing between low and medium education levels. Including data from other platforms with less selective usage patterns may be a promising avenue for future research.

Looking forward, recent advances in large language models present both interesting possibilities and challenges for our approach. On the one hand, LLMs could enable more sophisticated evaluation of tweet content, including advanced error detection, zero-shot topic classification, and text embeddings as features, potentially increasing predictive performance and facilitating multi-language integration.[21] On the other hand, as social media users increasingly adopt AI writing assistants, education signals from indicators such as spelling errors may become diluted. Nevertheless, most feature groups, such as penetration, usage statistics, network structures, and topics, should remain informative, and differential AI adoption across demographic groups (Alikhani, Harris, and Patnaik, 2025; OECD, 2024) may create new predictive signals.

Overall, our results demonstrate that social media data can provide accurate spatially granular estimates of educational attainment. As social media use continues to expand globally, this approach offers an increasingly promising complement to traditional survey methods for measuring human capital development.

[21] Note, however, that costs under current commercial API pricing remain prohibitively high for processing millions of tweets.

Data Availability

The datasets generated and analyzed in this study are available in the Harvard Dataverse repository at https://doi.org/10.7910/DVN/CWS2BJ.

Author Contributions

Martina Jakob conceptualized the project and drafted the manuscript. Sebastian Heinrich and Martina Jakob contributed equally to data acquisition, preparation, analysis, and visualizations. Both authors revised and approved the submitted manuscript.

Funding

This research received no external funding.

References

Aiken, Emily, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock (2022). "Machine learning and phone data can improve targeting of humanitarian aid". *Nature* 603.7903, 864–870.

Akrami, Haleh, Anand A Joshi, Sergül Aydöre, and Richard M Leahy (2022). "Deep quantile regression for uncertainty estimation in unsupervised and supervised lesion detection". *The Journal of Machine Learning for Biomedical Imaging* 1, 008.

Alikhani, Malihe, Ben Harris, and Sanjay Patnaik (Nov. 25, 2025). *How are Americans using AI? Evidence from a nationwide survey*. URL: https://www.brookings.edu/articles/how-are-americans-using-ai-evidence-from-a-nationwide-survey/ (visited on 01/17/2026).

Angelopoulos, Anastasios N and Stephen Bates (2021). "A gentle introduction to conformal prediction and distribution-free uncertainty quantification". *arXiv preprint arXiv:2107.07511*.

Barro, Robert J. and Jong Wha Lee (2013). "A new data set of educational attainment in the world, 1950–2010". *Journal of Development Economics* 104, 184–198.

Becker, Gary S. (1964). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*. Tech. rep. National Bureau of Economic Research.

Blumenstock, Joshua E., Gabriel Cadamuro, and Robert On (2015). "Predicting poverty and wealth from mobile phone metadata". *Science* 350.6264, 1073–1076.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). "Twitter mood predicts the stock market". *Journal of Computational Science* 2.1, 1–8.

Breen, Casey F, Masoomali Fatehkia, Jiani Yan, Xinyi Zhao, Douglas R Leasure, Ingmar Weber, and Ridhi Kashyap (2025). "Mapping subnational gender gaps in internet and mobile adoption using social media data". *Proceedings of the National Academy of Sciences* 122.42, e2416624122.

Burke, Marshall, Anne Driscoll, David B. Lobell, and Stefano Ermon (2021). "Using satellite imagery to understand and promote sustainable development". *Science* 371.6535, eabe8628.

Burke, Marshall and David B. Lobell (2017). "Satellite-based assessment of yield variation and its determinants in smallholder African systems". *Proceedings of the National Academy of Sciences* 114.9, 2189–2194.

Chang, Serina, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec (2021). "Mobility network models of COVID-19 explain inequities and inform reopening". *Nature* 589.7840, 82–87.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, Drew Johnston, Martin Koenen, Eduardo Laguna-Muggenburg, Florian Mudekereza, Tom Rutter, Nicolaj Thor, Wilbur Townsend, Ruby Zhang, Michael C. Bailey, Pablo Barberá, Monica Bhole, and Nils Wernerfelt (2022). "Social capital I: measurement and associations with economic mobility". *Nature* 608, 108–121.

Chi, Guanghua, Han Fang, Sourav Chatterjee, and Joshua E Blumenstock (2022). "Microestimates of wealth for all low-and middle-income countries". *Proceedings of the National Academy of Sciences* 119.3, e2113658119.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Curtis, Brenda, Salvatore Giorgi, Anneke E. K. Buffone, Lyle H. Ungar, Robert D. Ashford, Jessie Hemmons, Dan Summers, Casey Hamilton, and H. Andrew Schwartz (2018). "Can Twitter be used to predict county excessive alcohol consumption rates?" *PloS ONE* 13.4, e0194290.

Fuller, Wayne A. (1987). *Measurement error models*. Wiley Series in Probability and Mathematical Statistics. John Wiley Sons, Inc.

Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei (2017). "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States". *Proceedings of the National Academy of Sciences* 114.50, 13108–13113.

Gómez, Juan Carlos, Luis Miguel López Santamaría, Mario Alberto Ibarra Manzano, and Dora Luz Almanza Ojeda (2021). "Predicción automática del nivel educativo en usuarios de Twitter en México". *Realidad, datos y espacio Revista internacional de estadística y geografía* 12.1.

Head, Andrew, Mélanie Manguin, Nhat Tran, and Joshua E. Blumenstock (2017). "Can human development be measured with satellite imagery?" *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. ICTD '17. Lahore, Pakistan: Association for Computing Machinery.

Huang, Xiao, Zhenlong Li, Yuqin Jiang, Xiaoming Li, and Dwayne Porter (2020). "Twitter reveals human mobility dynamics during the COVID-19 pandemic". *PloS ONE* 15.11, e0241957.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon (2016). "Combining satellite imagery and machine learning to predict poverty". *Science* 353.6301, 790–794.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch (July 2018). "Marian: Fast neural machine translation in C++". *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, 116–121.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). "How censorship in China allows government criticism but silences collective expression". *American Political Science Review* 107.2, 326–343.

Kuffer, Monika, Karin Pfeffer, and Richard Sliuzas (2016). "Slums from space—15 years of slum mapping using remote sensing". *Remote Sensing* 8.6, 455.

Lobell, David B. (2013). "The use of satellite data for crop yield gap analysis". *Field Crops Research* 143, 56–64.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 4765–4774.

Maddox, Wesley J, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson (2019). "A simple baseline for bayesian uncertainty in deep learning". *Advances in Neural Information Processing Systems* 32.

Mboga, Nicholus, Claudio Persello, John Ray Bergado, and Alfred Stein (2017). "Detection of informal settlements from VHR images using convolutional neural networks". *Remote Sensing* 9.11, 1106.

Mellon, Jonathan and Christopher Prosser (2017). "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users". *Research & Politics* 4.3, 2053168017720008.

Mincer, Jacob (1974). *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen Carley (2013). "Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose". *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 7. 1, 400–408.

Nutakki, Brahmani, Rosa M Navarrete, Giuseppe Carteny, and Ingmar Weber (2025). "Is there anything Left?: A Global Analysis on Changes in Engagement with Political Content on Twitter in the Musk Era". *Journal of Quantitative Description: Digital Media* 5.

OECD (2024). *ICT Access and Use by Education*. Organisation for Economic Co-operation and Development.

Özturan, Burak, Alexi Quintana-Mathé, Nir Grinberg, Katherine Ognyanova, and David Lazer (2025). "Declining information quality under new platform governance". *Harvard Kennedy School Misinformation Review*.

Poushter, Jacob, Caldwell Bishop, and Hanyu Chwe (2018). *Social media use continues to rise in developing countries but plateaus across developed ones*. Tech. rep. Pew Research Center.

Ratledge, Nathan, Gabe Cadamuro, Brandon de la Cuesta, Matthieu Stigler, and Marshall Burke (2022). "Using machine learning to assess the livelihood impact of electricity access". *Nature* 611.7936, 491–495.

Robertson, Craig T. (2023). *What our research says about news audiences on Twitter (now X)*. Reuters Institute for the Study of Journalism, University of Oxford. Accessed January 15, 2026.

Shapley, Lloyd S (1953). "A Value for n-Person Games". *Contributions to the Theory of Games II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton: Princeton University Press, 307–317.

Sheehan, Evan, Chenlin Meng, Matthew Tan, Burak Uzkent, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon (2019). "Predicting Economic Development Using Geolocated Wikipedia Articles". *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2698–2706.

Smirnov, Ivan (2020). "Estimating educational outcomes from students' short texts on social media". *EPJ Data Science* 9.1, 1–11.

Stahel, Werner A. (2000). *Statistische Datenanalyse: Eine Einfuehrung fuer Naturwissenschaftler*. Vieweg+Teubner Verlag Wiesbaden.

Statista (2022). *Number of social media users worldwide from 2017 to 2027*. URL: `https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/` (visited on 02/2023).

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem (2015). "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data". *PLoS ONE* 10.2, e0107042.

Sun, Jie, Liping Di, Ziheng Sun, Yonglin Shen, and Zulong Lai (2019). "County-level soybean yield prediction using deep CNN-LSTM model". *Sensors* 19.20, 4363.

Ushio, Asahi and Jose Camacho-Collados (2022). "TweetNLP: Cutting-edge natural language processing for social media". *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.* Abu Dhabi, U.A.E.: Association for Computational Linguistics.

Wardrop, Nicola A., Warren C. Jochem, Tomas J. Bird, Heather R. Chamberlain, Donna J. Clarke, David Kerr, Linus Bengtsson, Sabrina Juran, Vincent Seaman, and Andrew J. Tatem (2018). "Spatially disaggregated population estimates in the absence of national population and housing census data". *Proceedings of the National Academy of Sciences of the United States of America* 115, 3529–3537.

Wesolowski, Amy, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdisalan M. Noor, Robert W. Snow, and Caroline O. Buckee (2012). "Quantifying the impact of human mobility on malaria". *Science* 338.6104, 267–270.

Yeh, Christopher, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke (2020). "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". *Nature Communications* 11.1, 2583.

Yin, Junjun, Yizhao Gao, and Guangqing Chi (2022). "An evaluation of geo-located Twitter data for measuring human migration". *International Journal of Geographical Information Science* 36.9, 1830–1852.

Supplementary Information

A Model Performance

A.1 Main Results

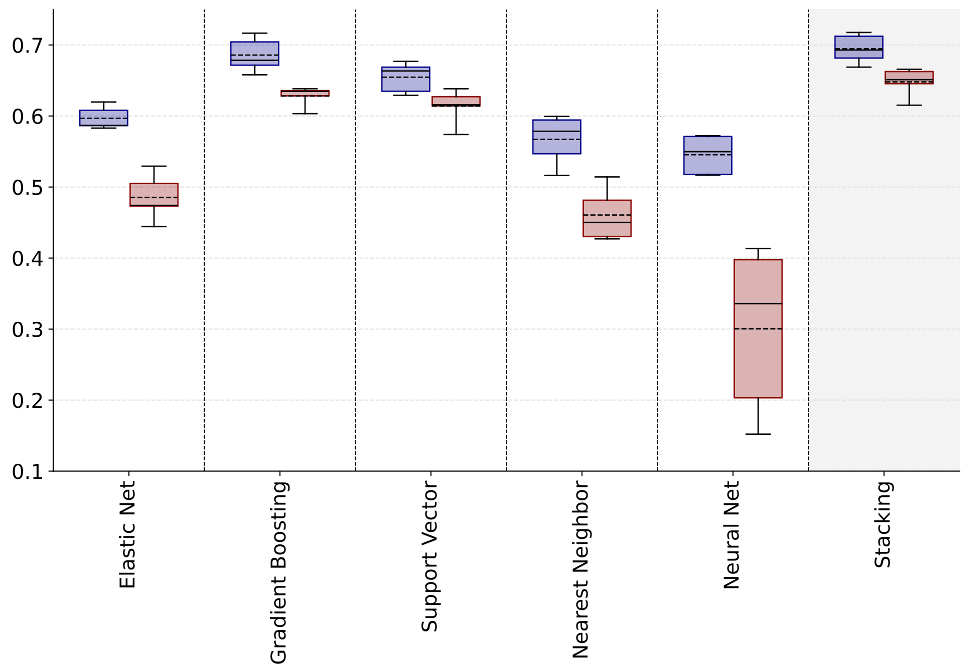


Figure A.1: Performance of individual models

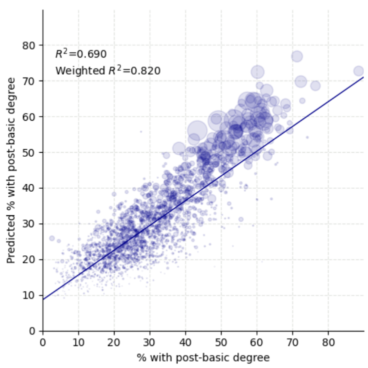
Performance of individual models considered in the final stacking model for years of schooling in Mexico (blue) and the United States (red). All models are evaluated through five-fold cross-validation. Boxplots show the median (solid line), mean (dotted line), the 20th & 80th percentile (box limits), as well as the minimum & maximum (whiskers) for the R^2 across validation folds for each outcome and country.

Table A.1: Performance of individual models

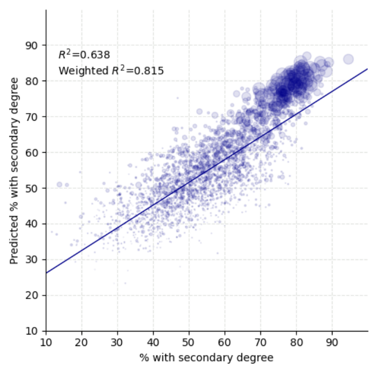
	Mexico				United States			
	Years of Schooling	Post Basic Education	Secondary Education	Primary Education	Years of Schooling	Bachelor Degree	Some College	Only High School
Elastic Net	0.597 (2.8%)	0.621 (-5.6%)	0.548 (-15.3%)	0.495 (-17.7%)	0.485 (0.2%)	0.522 (-1.4%)	0.461 (4.6%)	0.350 (-4.0%)
Gradient Boosting	0.686 (56.9%)	0.689 (61.6%)	0.638 (62.1%)	0.600 (60.7%)	0.628 (56.5%)	0.674 (52.7%)	0.603 (55.6%)	0.467 (53.0%)
Support Vector Machine	0.655 (29.1%)	0.602 (6.2%)	0.544 (7.7%)	0.459 (-0.1%)	0.614 (41.0%)	0.669 (37.7%)	0.576 (37.1%)	0.457 (42.7%)
Nearest Neighbour Matching	0.567 (0.6%)	0.576 (-0.2%)	0.523 (7.5%)	0.490 (11.5%)	0.461 (3.3%)	0.504 (0.2%)	0.425 (3.3%)	0.359 (15.2%)
Multi-layer Perceptron	0.545 (13.2%)	0.654 (40.3%)	0.590 (41.4%)	0.557 (48.1%)	0.300 (6.2%)	0.627 (16.9%)	0.424 (6.7%)	-0.523 (3.2%)
Stacking	0.695	0.689	0.638	0.607	0.648	0.697	0.620	0.500

Mean R^2 and stacking weights (in parentheses) across folds for different models and outcomes. Note that R^2 values for the final stacking model reported as our main results are computed using the combined out-of-sample predictions of all folds rather than as the mean across folds, and may thus slightly differ.

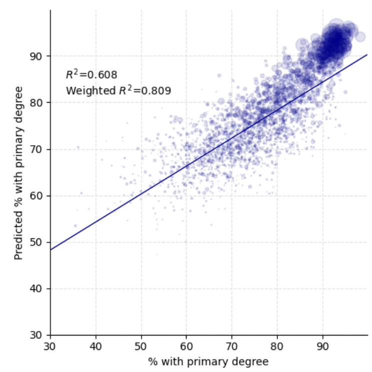
(a) MX: Post-basic degree



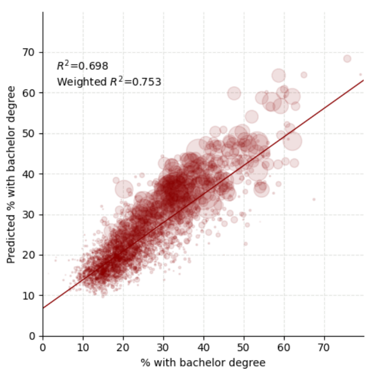
(b) MX: Secondary degree



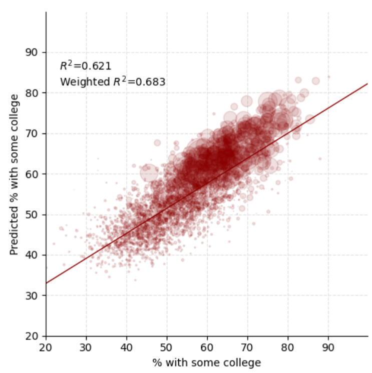
(c) MX: Primary degree



(d) US: Bachelor



(e) US: Some College



(f) US: High school

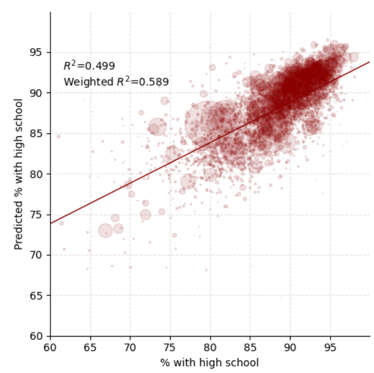


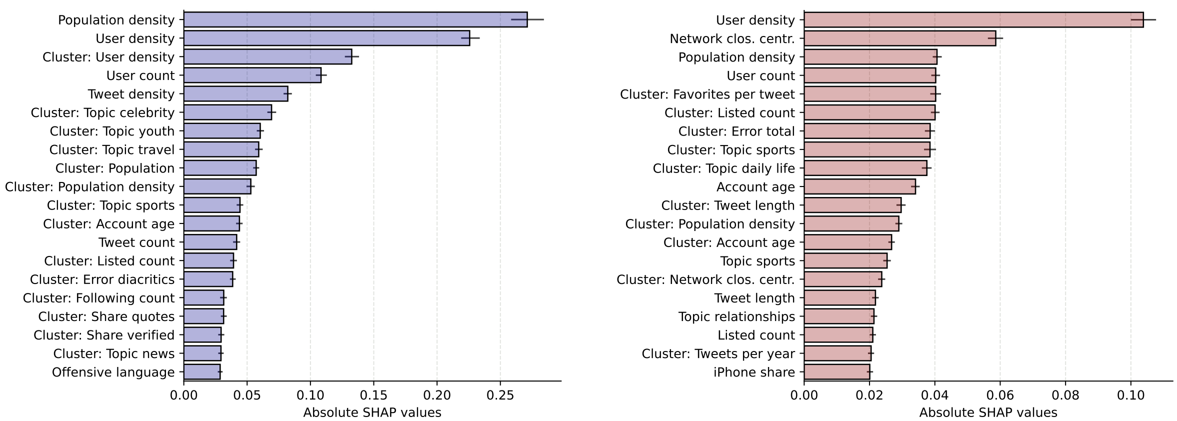
Figure A.2: True vs. predicted values for secondary outcomes

Predicted values for all Mexican municipalities (blue) and US counties (red) are obtained by combining out-of-sample predictions from all folds. Bubble size is proportional to the population in each unit. R^2 and population weighted R^2 shown. Line indicating best linear fit is not population weighted.

A.2 Feature Importance

We compute SHAP values from the final stacking regressor, which combines predictions from five base models (elastic net, gradient boosting, support vector regression, nearest neighbor regression, and neural network) and is thus representative of our overall approach. SHAP values are calculated for each municipality or county based on out-of-sample predictions from cross-validation. We present results for the most important individual features (Figure A.3a) as well as per group aggregations (Figure A.3b) based on sums of absolute SHAP contributions. Note that while for signed SHAP values, the

(a) Most important features



(b) Aggregation by feature subgroups

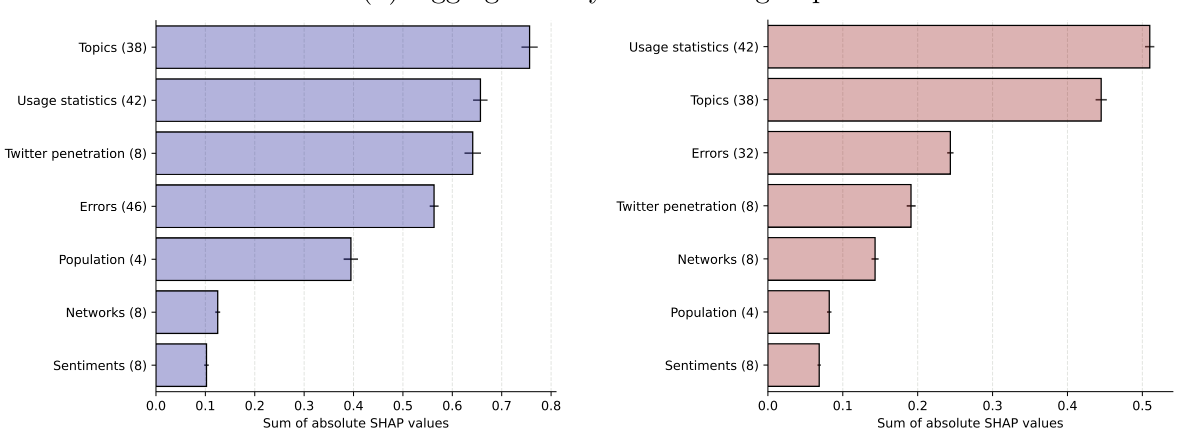


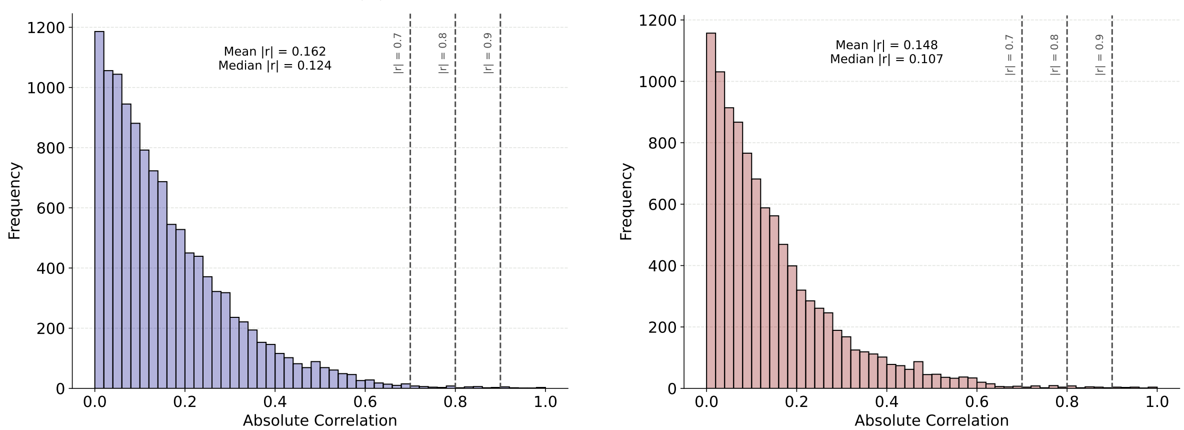
Figure A.3: SHAP value feature importance for stacking regressor

Most important features and feature subgroups in the stacking regressor for Mexico (blue) and the United States (red). The displayed feature importance is based on SHAP (SHapley Additive exPlanations) values, an approach based on cooperative game theory (Lundberg and S.-I. Lee, 2017). In Figure A.3a, the 20 most important features are shown for each country, ranked according to the absolute SHAP value. Figure A.3b presents the sum of absolute SHAP values for 7 different feature groups. Note that due to the high correlation between some features, estimates should be interpreted with caution.

sum of feature contributions equals the difference between the individual prediction and the average prediction across all observations, this “additivity property” does not hold for absolute values (Figure A.3a) or sums of absolute values (Figure A.3b). Our presented measures should thus be interpreted as heuristic summaries of predictive contribution rather than formal decompositions.

While SHAP is specifically designed to assess contributions of correlated individual features, results could still be unstable under very high multicollinearity. We assess robustness in two steps. First, we examine pairwise feature correlations in our feature matrices, finding that only a very limited number of feature pairs are highly correlated

(a) Distribution of pairwise correlations



(b) High-correlation pairs by feature group

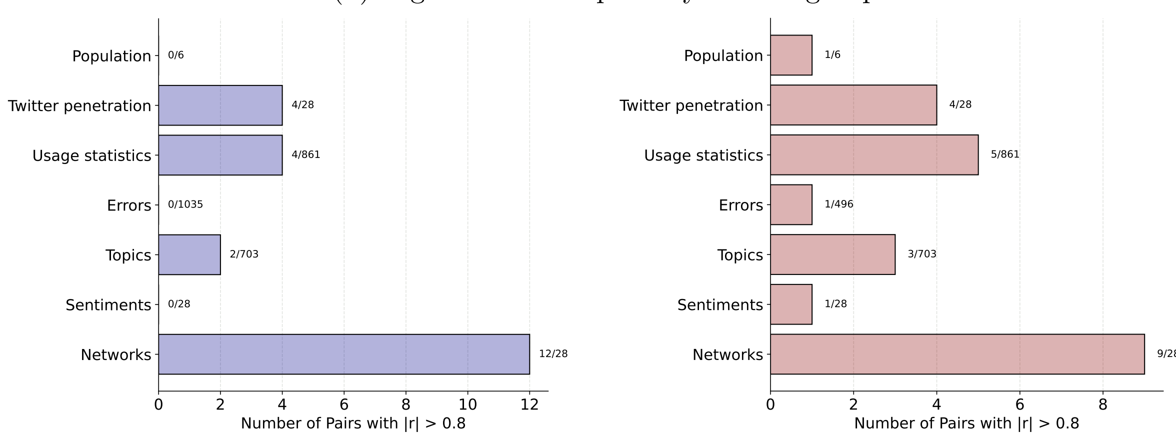


Figure A.4: Feature correlation analysis

Pairwise correlations between features for Mexico (blue) and the United States (red). Figure A.4a shows the distribution of absolute correlation coefficients across all feature pairs. Vertical dashed lines indicate thresholds at $|r| = 0.7$, 0.8, and 0.9. Figure A.4b displays the number of highly correlated feature pairs ($|r| > 0.8$) within each feature group, shown as fractions of total possible pairs. High correlations occur predominantly within network groups, while all other groups show few or no high-correlation pairs in both countries.

(see Figure A.4), with mean absolute correlations of $|r| = 0.16$ for Mexico and $|r| = 0.15$ for the United States. Second, to directly assess whether feature importance rankings are stable across different model specifications and data samples, we examine the consistency of SHAP values across cross-validation folds. Each fold represents a different subsample of municipalities/counties and a distinct model configuration with fold-specific hyperparameters and weights. Feature importance rankings prove remarkably stable, with mean Spearman rank correlations of $\rho = 0.69$ for Mexico and $\rho = 0.81$ for the United States. In both countries, the most important feature overall also consistently ranks first across all five folds. Among the top five features, 92% rank within the top five in each individual fold for Mexico and 68% for the United States, while all of them rank within the top ten in both countries.

A.3 Leave-One-State-Out Cross-Validation

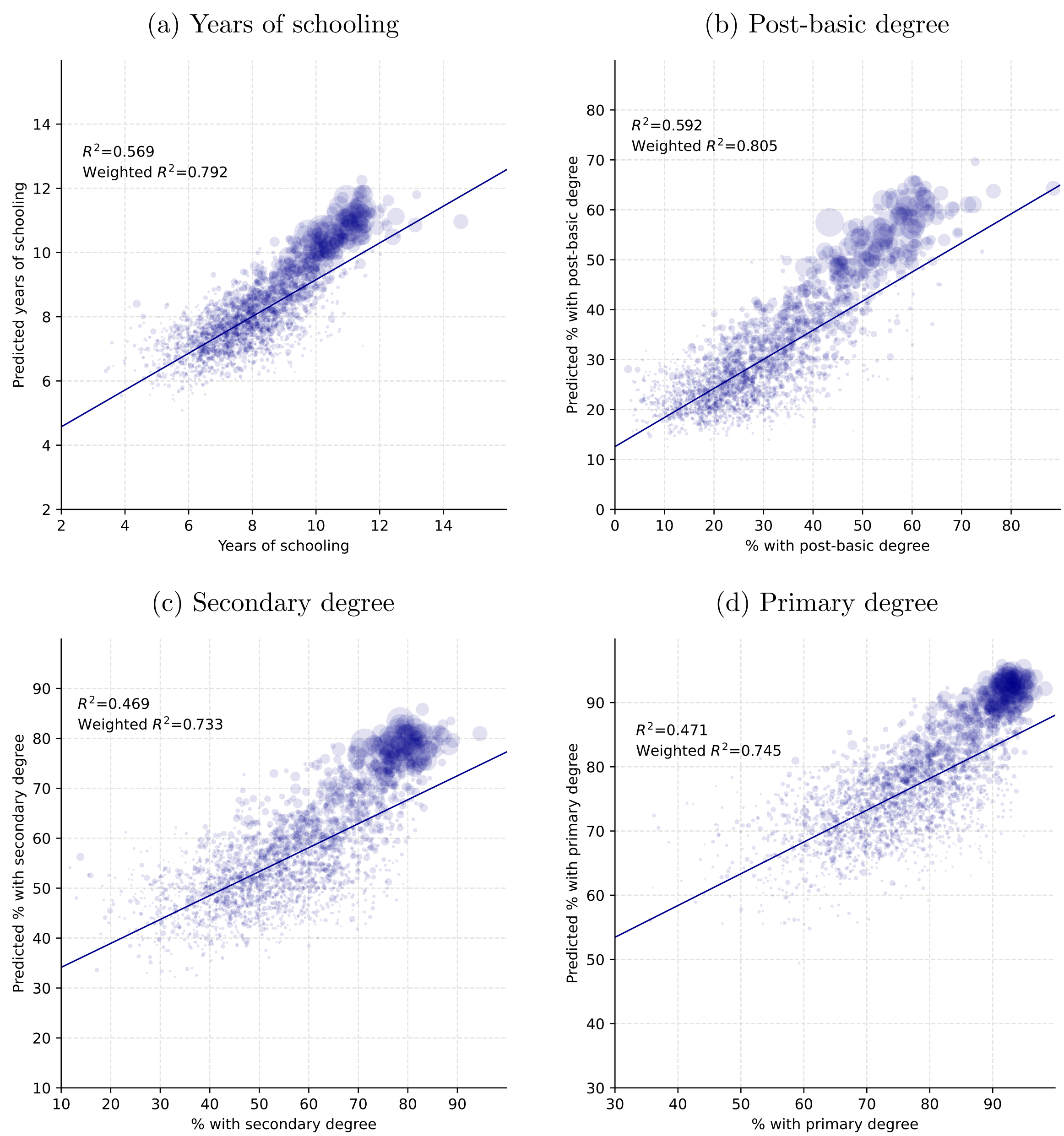


Figure A.5: Leave-one-state-out cross-validation results for Mexico

Predicted values from leave-one-state-out cross-validation where models are trained excluding all municipalities from one state at a time and evaluated on the municipalities of the held-out state. Bubble size is proportional to population. R^2 and population weighted R^2 shown. Line indicating best linear fit is not population weighted.

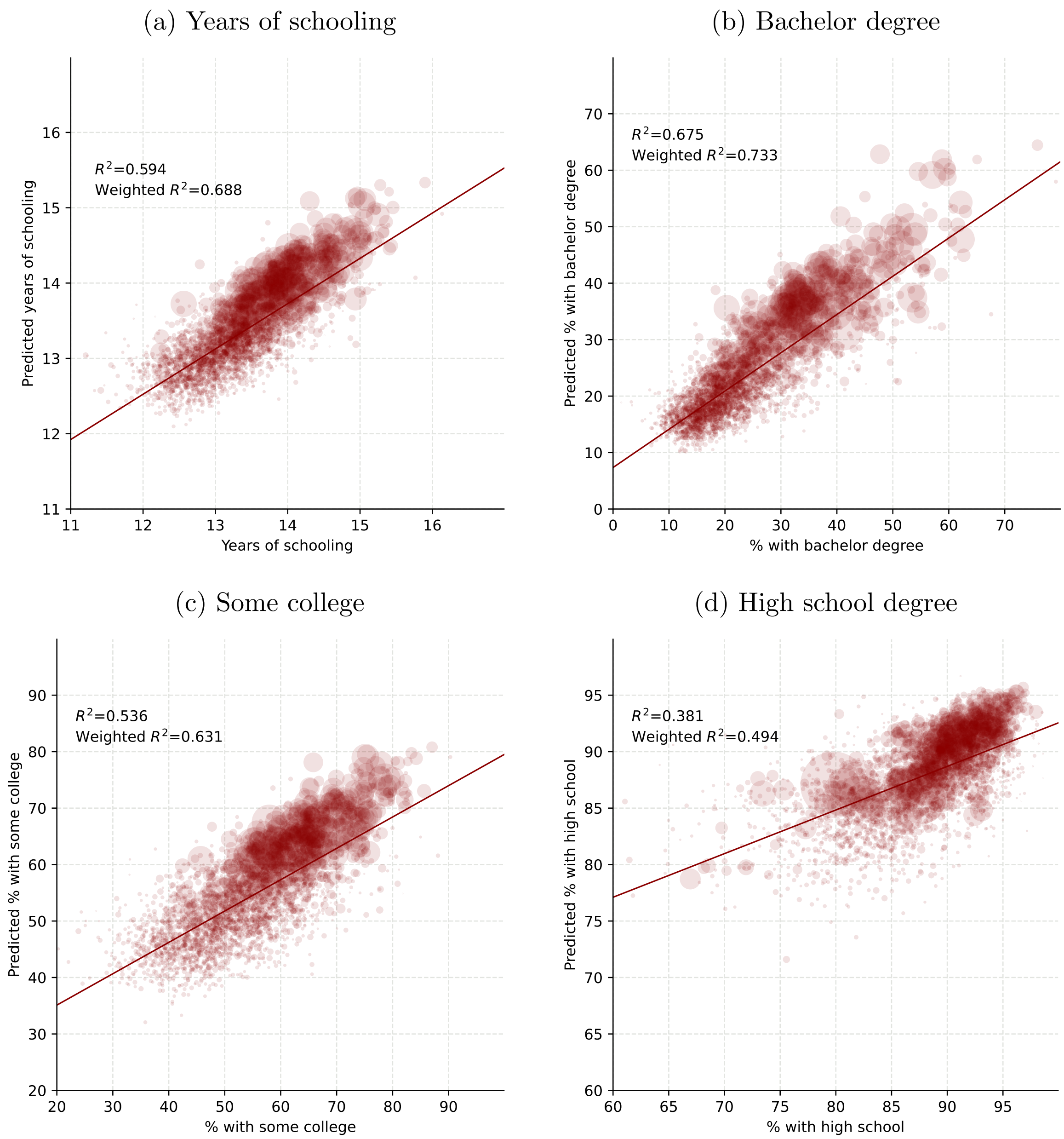


Figure A.6: Leave-one-state-out cross-validation results for the United States

Predicted values from leave-one-state-out cross-validation where models are trained excluding all counties from one state at a time and evaluated on the counties of the held-out state. Bubble size is proportional to population. R^2 and population weighted R^2 shown. Line indicating best linear fit is not population weighted.

A.4 Model Predictions for Future Outcomes

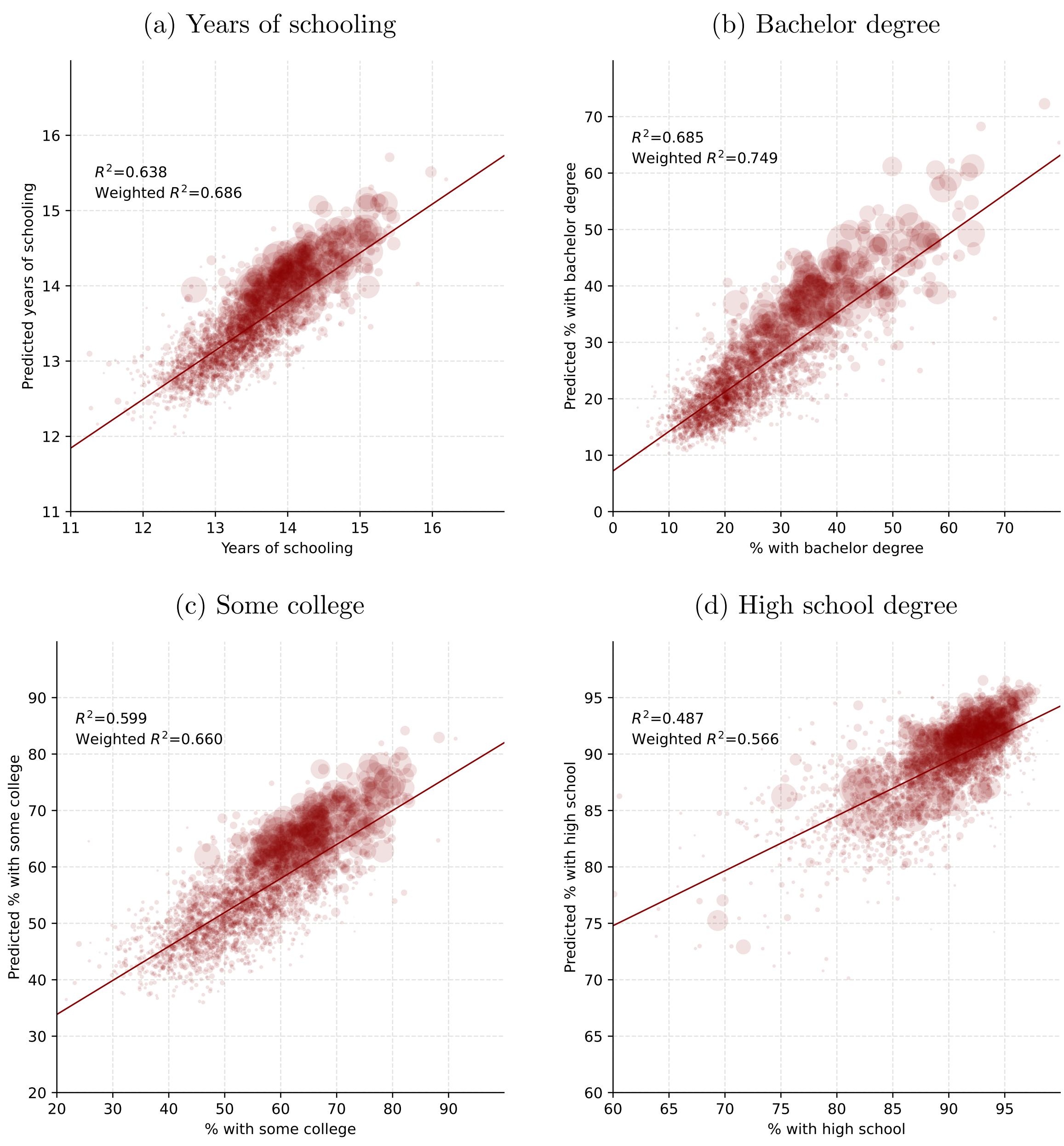


Figure A.7: Future education outcome results for the United States

Predictions vs. true values based on a model retrained and evaluated using more recent American Community Survey data (2019–2023) compared to the main analysis (2017–2021). Bubble size is proportional to population. R^2 and population weighted R^2 shown. Line indicating best linear fit is not population weighted.

A.5 Prediction Uncertainty

To assess the reliability of our predictions, we quantify prediction variability by repeatedly re-running the full cross-validation procedure with different random fold assignments. Specifically, we re-estimate the model 20 times using different random seeds for fold construction and compute, for each municipality or county, the standard deviation of the resulting out-of-sample predictions. This allows us to characterize prediction uncertainty at the unit level and examine how it correlates with observable characteristics.[22]

Uncertainty is generally higher in Mexican municipalities than in US counties (Figure A.8). This likely reflects the higher variance in years of schooling in Mexico (SD = 1.49 for Mexico vs. SD = 0.66 in the US). Within both countries, variability is substantially higher in less populous municipalities or counties, in areas with fewer Twitter users and tweets, and where the model relies more heavily on spatial imputation (Figure A.9). In Mexico, prediction variability additionally increases at lower levels of educational attainment, a pattern that is much less pronounced in the United States.

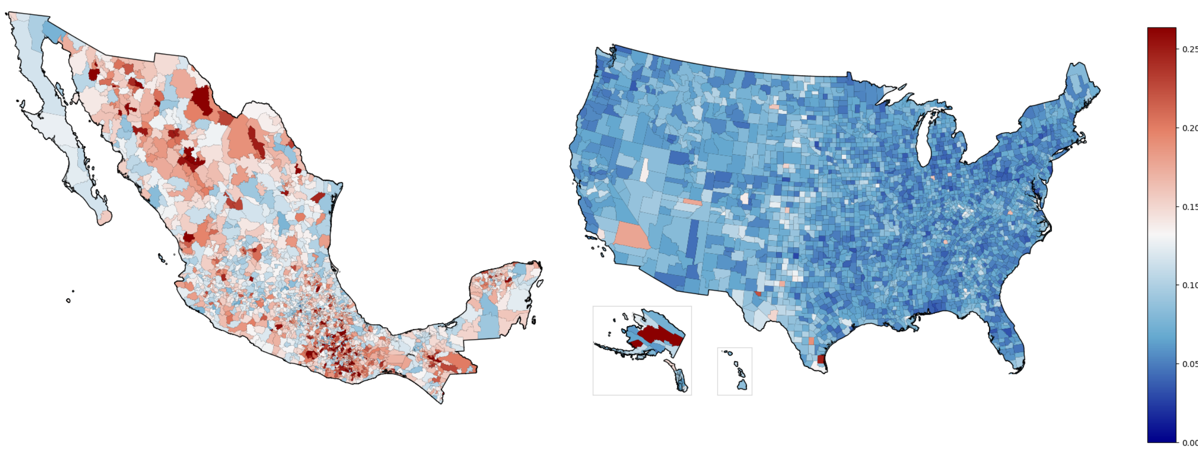


Figure A.8: Prediction uncertainty for Mexican municipalities and US counties
Prediction uncertainty for years of schooling, measured as the standard deviation of predicted values across repeated cross-validation re-runs, for each municipality in Mexico (left) and county in the United States (right).

[22]This simple approach is consistent with our overall training setup and provides an easy-to-interpret measure of prediction stability, but is computationally intensive due to repeated model re-estimation. More recent approaches to predictive uncertainty estimation include Bayesian machine learning methods, quantile-based models, and conformal prediction (e.g., Maddox et al., 2019; Akrami et al., 2022; Angelopoulos and Bates, 2021).

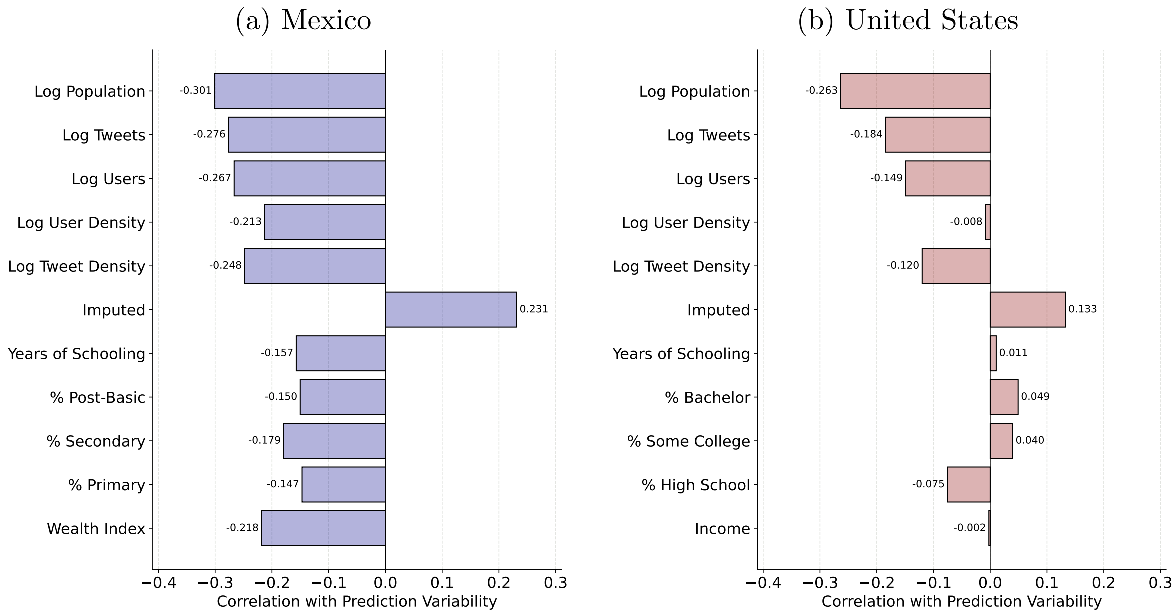


Figure A.9: Correlates of prediction variability

Unconditional correlation between prediction uncertainty for years of schooling, measured as the standard deviation of predicted values across repeated cross-validation re-runs, and selected characteristics of Mexican municipalities and US counties.

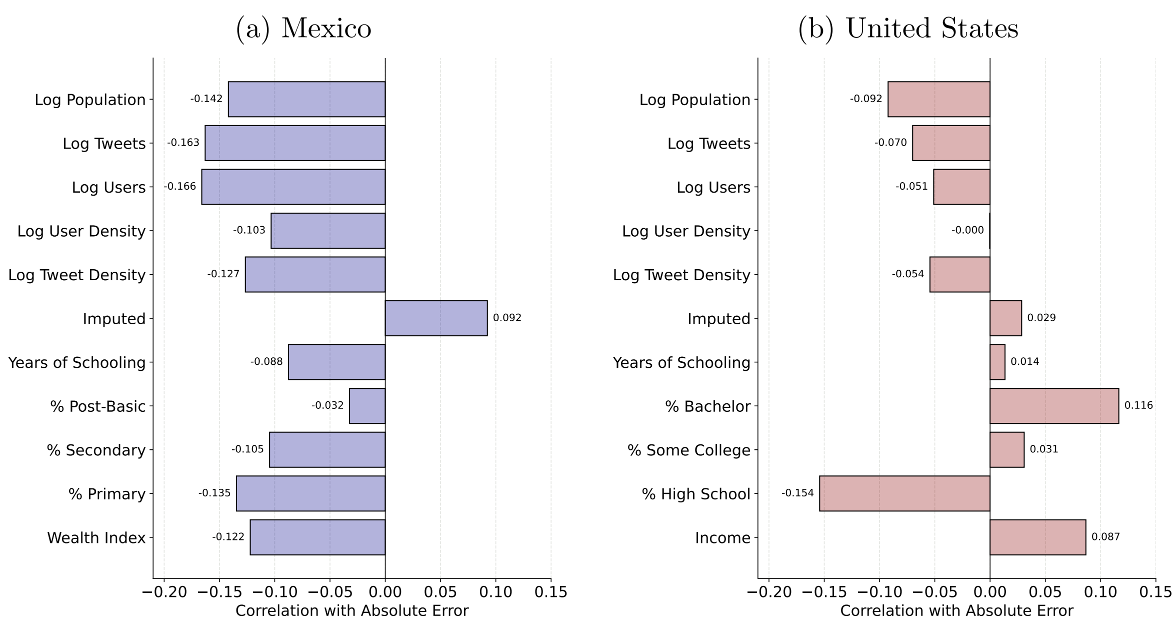


Figure A.10: Correlates of prediction error

Unconditional correlation between absolute prediction errors for years of schooling and selected characteristics of Mexican municipalities and US counties.

Relatedly, we examine correlates of absolute prediction errors. Unsurprisingly, results are similar to uncertainty, suggesting that areas with higher prediction variability also tend to have larger prediction errors (see Figure A.10). Given that spatial clustering is a key component of our data preparation pipeline, we also disaggregate performance by imputation status. As expected, we find larger mean average prediction errors (MAE) for areas that rely more heavily on imputed features. However, increases in MAE compared to non-imputed areas are only 10 percent for the US and 20 percent for Mexico, suggesting that predictions remain reasonably accurate even in these sparse-data contexts (Figure A.11).

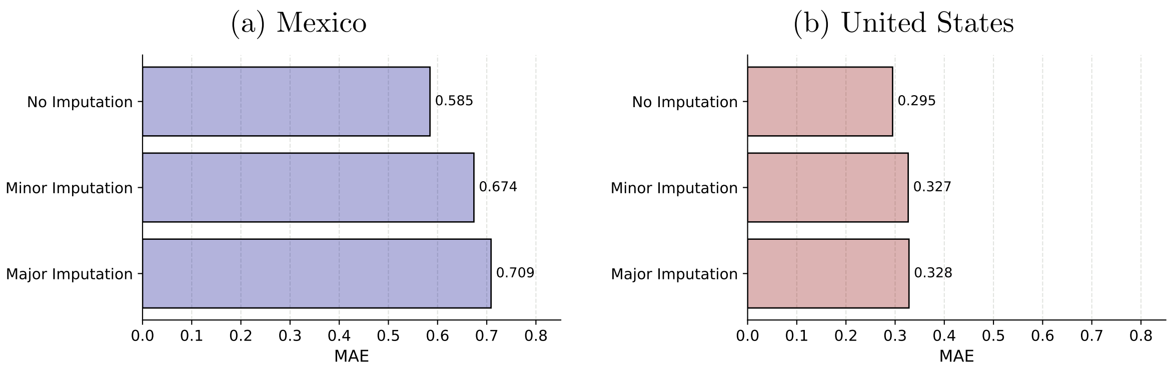


Figure A.11: Model performance by imputation level

Mean Absolute Error (MAE) for predictions of years of schooling in Mexican municipalities (blue) and US counties (red) grouped by imputation level. "No Imputation": units with at least 5 tweets (complete observed data); "Minor Imputation": units with 1-4 tweets (outlier imputation based on spatial neighbors); "Major Imputation": units with 0 tweets (full imputation of all tweet content features based on spatial neighbors). Lower MAE indicates better predictive performance.

B Bias Correction

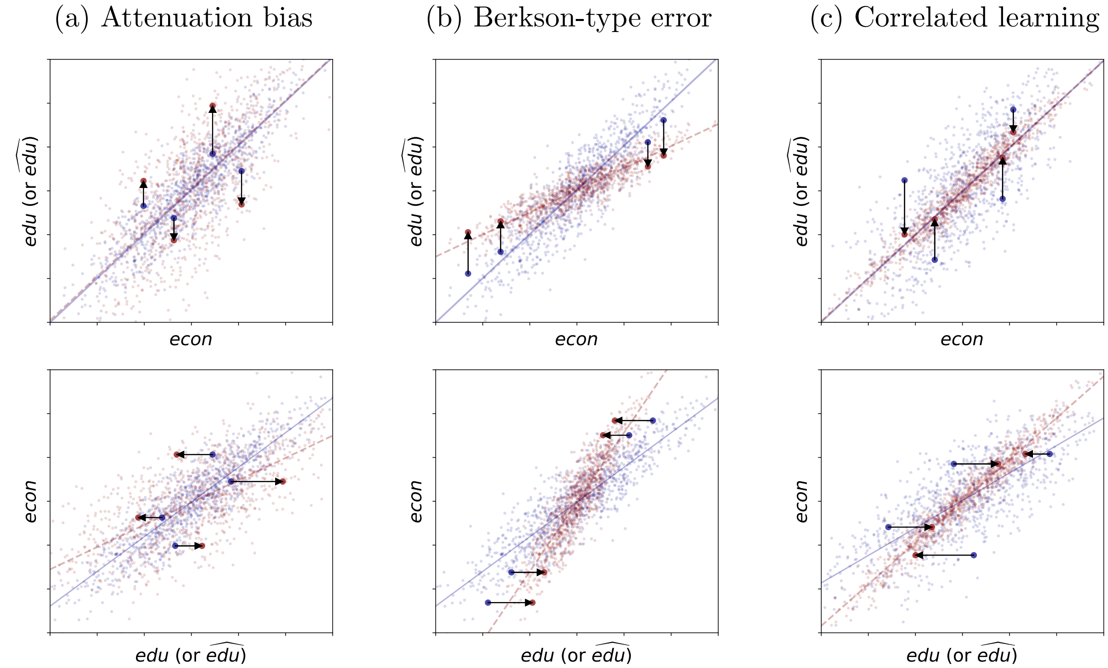


Figure B.1: Simulation of different types of biases in downstream regression tasks

Scatter plots and best linear fit for *edu* (blue) and $\widehat{edu}$ (red) with different types of measurement errors. Arrows indicate the movement of typical points as a result of each measurement error. In the upper row, *edu* (or $\widehat{edu}$) is the outcome of the regression, while it features as the explanatory variable in the lower row.

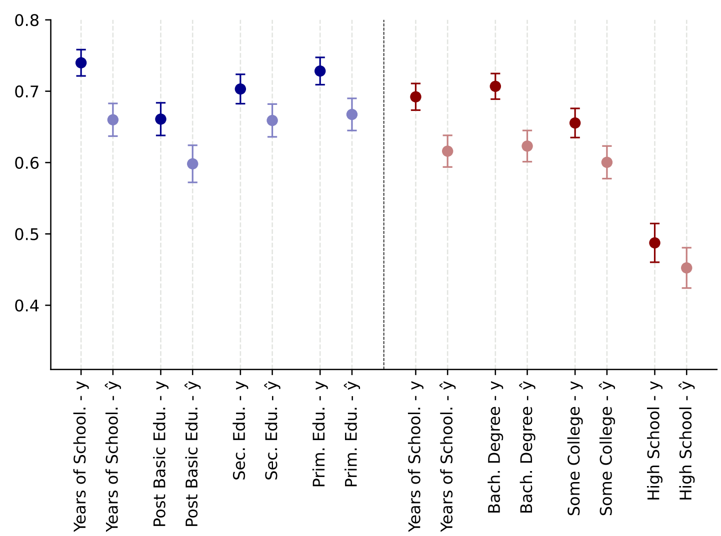


Figure B.2: Correlation of observed and predicted education with wealth or income

Correlations between true and predicted educational outcomes and wealth in Mexico (blue) as well as income in the United States (red). 95% confidence intervals shown.

C Feature Statistics

Table C.1: Survey statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Years of Schooling	MX	7.83	1.49	3.40	7.72	14.55
	US	13.30	0.66	9.37	13.28	16.13
Post Basic Education	MX	0.28	0.13	0.03	0.26	0.89
Bachelor Degree	US	22.61	9.71	0.00	20.22	79.14
Secondary Education	MX	0.54	0.14	0.12	0.54	0.95
Some College	US	53.67	10.72	7.41	53.61	90.31
Primary Education	MX	0.76	0.11	0.36	0.76	0.98
High School	US	87.60	6.04	21.85	88.83	98.61
Population	MX	51,173.11	147,322.51	81.00	13,552.00	1,922,523.00
	US	105,661.95	333,146.18	57.00	25,790.00	9,829,544.00
Wealth Index	MX	0.68	0.12	0.07	0.70	0.94
Income	US	57,455.86	14,582.81	22,901.00	55,143.50	160,305.00

Table C.2: Twitter penetration and usage statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Tweet count	MX	1,093.25	6,363.61	0.00	8.00	119,126.00
	US	7,195.60	42,124.22	0.00	271.00	1,472,677.00
User count	MX	50.11	269.79	0.00	2.00	5,891.00
	US	299.54	1,548.14	0.00	24.00	52,602.00
Share weekdays	MX	0.70	0.22	0.00	0.72	1.00
	US	0.70	0.15	0.00	0.71	1.00
Share workhours	MX	0.29	0.21	0.00	0.28	1.00
	US	0.31	0.15	0.00	0.31	1.00
Follower count	MX	251.37	1,210.64	0.00	111.83	36,807.50
	US	304.11	754.33	0.00	229.50	24,799.80
Following count	MX	358.47	480.14	0.00	261.71	7,603.16
	US	415.74	556.26	1.00	359.50	25,202.00
Tweet count	US	2,659.14	6,237.16	1.00	1,927.00	183,023.00
	MX	2,405.62	5,565.75	1.00	961.45	79,700.00
User mobility	MX	1.61	0.75	1.00	1.50	10.00
	US	1.76	0.80	1.00	1.71	32.00
iPhone share	US	0.62	0.23	0.00	0.67	1.00
	MX	0.28	0.27	0.00	0.25	1.00
Instagram share	US	0.21	0.24	0.00	0.12	1.00
	MX	0.14	0.23	0.00	0.06	1.00
Favorites per tweet	US	1.73	8.74	0.00	1.35	463.46
	MX	3.76	23.95	0.00	1.25	892.38
Tweets per year	US	495.19	1,703.97	0.39	325.13	52,472.55
	MX	841.93	6,183.31	0.82	260.06	210,975.91
Account age	MX	5.67	2.67	-0.03	5.88	13.56
	US	7.06	1.69	-0.02	7.10	14.65

Variable	Country	Mean	SD	Min	Median	Max
Account age	MX	5.67	2.67	-0.03	5.88	13.56
	US	7.06	1.69	-0.02	7.10	14.65
Listed count	MX	2.86	7.79	0.00	1.00	151.40
	US	12.39	30.79	0.00	6.67	711.60
Followers per following	MX	0.64	2.39	0.00	0.38	68.93
	US	0.71	1.60	0.00	0.60	61.70
Share quotes	MX	0.07	0.11	0.00	0.04	1.00
	US	0.09	0.07	0.00	0.10	1.00
Share replies	MX	0.24	0.22	0.00	0.23	1.00
	US	0.22	0.13	0.00	0.24	1.00
Share verified	MX	0.00	0.03	0.00	0.00	1.00
	US	0.01	0.03	0.00	0.00	1.00
Tweet length	MX	69.94	29.48	4.00	67.53	274.00
	US	80.86	21.74	6.00	79.01	275.00
Hashtags per tweet	MX	0.30	0.51	0.00	0.17	8.00
	US	0.38	0.52	0.00	0.28	8.00
Mentions per tweet	US	0.53	0.31	0.00	0.56	4.44
	MX	0.45	0.45	0.00	0.41	7.00
Urls per tweet	US	0.38	0.52	0.00	0.28	8.00
	MX	0.30	0.51	0.00	0.17	8.00
Emojis per tweet	MX	0.81	0.70	0.00	0.72	6.67
	US	0.55	0.54	0.00	0.50	15.00

Table C.3: Error statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Error typography	MX	7.71	9.03	0.00	6.57	170.57
	US	2.55	2.47	0.00	2.18	30.61
Error grammar	MX	0.17	0.54	0.00	0.00	9.80
	US	0.55	0.82	0.00	0.43	15.87
Error confusions	MX	0.14	0.51	0.00	0.00	11.11
	US	0.10	0.26	0.00	0.04	7.44
Error casing	MX	1.29	3.49	0.00	0.18	55.56
	US	1.73	2.74	0.00	1.29	60.61
Error misc	MX	0.12	1.22	0.00	0.00	47.15
	US	0.19	0.44	0.00	0.13	14.71
Error style	US	0.43	0.92	0.00	0.30	23.81
	MX	0.02	0.24	0.00	0.00	7.19
Error repetitions style	US	0.00	0.00	0.00	0.00	0.01
	MX	0.00	0.00	0.00	0.00	0.08
Error semantics	US	0.01	0.11	0.00	0.00	4.57
	MX	0.01	0.06	0.00	0.00	1.94
Error variants	US	0.01	0.06	0.00	0.00	2.51
	MX	0.07	0.72	0.00	0.00	25.64
Error punctuation	US	1.02	1.39	0.00	0.91	32.26
	MX	0.66	1.83	0.00	0.24	35.51
Error typos	US	7.19	5.94	0.00	6.78	181.82
	MX	12.47	13.82	0.00	10.16	285.71
Error total	MX	25.28	17.43	0.00	22.87	285.71
	US	13.87	7.68	0.00	13.16	181.82

Variable	Country	Mean	SD	Min.	Median	Max.
Error expressions	MX	0.02	0.15	0.00	0.00	2.82
Error redundancy	MX	0.00	0.01	0.00	0.00	0.23
Error prepositions	MX	0.00	0.00	0.00	0.00	0.04
Error verb agreement	MX	0.02	0.27	0.00	0.00	10.42
Error misspelling	MX	1.18	4.04	0.00	0.50	125.00
Error proper nouns	MX	0.00	0.01	0.00	0.00	0.34
Error diacritics	MX	1.07	1.60	0.00	0.68	16.67
Error context	MX	0.00	0.01	0.00	0.00	0.38
Error repetitions	MX	0.01	0.14	0.00	0.00	5.38
Error norm change	MX	0.07	0.28	0.00	0.00	5.56
Error noun agreement	MX	0.26	0.80	0.00	0.02	14.93
Error collocations	US	0.01	0.20	0.00	0.00	8.20
Error nonstandard	US	0.00	0.02	0.00	0.00	0.49
Error redundancy	US	0.05	0.42	0.00	0.01	18.87
Errror compounding	US	0.02	0.13	0.00	0.00	5.32

Table C.4: Topic statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Topic arts & culture	MX	3.87	2.28	0.30	3.39	24.89
	US	2.97	1.54	0.27	2.68	29.31
Topic business	MX	3.00	3.15	0.26	2.33	46.23
	US	2.79	2.10	0.24	2.55	36.69
Topic celebrity	MX	6.83	5.06	0.31	6.33	49.20
	US	6.10	2.95	0.37	6.33	42.41
Topic daily life	MX	26.03	8.28	1.22	25.48	59.21
	US	21.30	5.91	1.26	20.74	60.28
Topic family	MX	4.03	3.09	0.32	3.52	32.72
	US	3.99	1.89	0.30	3.74	33.24
Topic fashion	MX	1.33	1.47	0.20	1.00	18.35
	US	1.67	1.49	0.19	1.49	38.24
Topic films	MX	4.12	4.23	0.36	3.38	64.47
	US	4.25	3.46	0.26	4.04	67.58
Topic fitness & health	US	2.42	1.61	0.31	2.33	37.45
	MX	2.38	2.36	0.22	1.88	31.72
Topic food & dining	US	3.12	3.35	0.16	2.72	48.61
	MX	2.29	3.29	0.15	1.42	47.21
Topic gaming	MX	1.47	1.55	0.17	1.14	31.18
	US	2.24	1.31	0.29	2.07	19.91

Variable	Country	Mean	SD	Min	Median	Max
Topic educational	MX	1.45	1.65	0.15	1.10	24.93
	US	1.80	1.65	0.16	1.52	30.81
Topic music	MX	4.67	5.57	0.13	3.74	69.11
	US	4.01	3.14	0.12	3.85	66.19
Topic news	MX	13.00	9.50	0.34	11.69	83.39
	US	12.29	5.85	0.40	12.00	80.56
Topic hobbies	MX	7.31	3.21	0.49	6.92	34.64
	US	5.12	2.06	0.34	4.82	26.46
Topic relationships	MX	6.27	3.70	0.29	5.78	27.32
	US	4.76	1.98	0.30	4.58	21.19
Topic science	MX	1.87	2.79	0.23	1.22	54.54
	US	1.69	1.71	0.29	1.49	46.50
Topic sports	MX	5.87	6.44	0.31	4.54	65.84
	US	15.14	9.31	0.12	14.36	85.78
Topic travel	MX	3.28	3.37	0.24	2.31	30.65
	US	2.95	2.37	0.23	2.19	30.52
Topic youth	MX	0.93	1.26	0.18	0.68	22.34
	US	1.40	1.34	0.18	1.15	23.65

Table C.5: Sentiment statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Sentiment negative	MX	0.16	0.12	0.00	0.16	0.95
	US	0.16	0.08	0.00	0.17	0.91
Sentiment positive	MX	0.38	0.18	0.01	0.37	0.99
	US	0.50	0.13	0.01	0.48	0.99
Hate speech	MX	0.04	0.03	0.01	0.04	0.42
	US	0.05	0.02	0.01	0.04	0.33
Offensive language	MX	0.15	0.07	0.03	0.15	0.89
	US	0.16	0.06	0.03	0.16	0.83

Table C.6: Network statistics by country

Variable	Country	Mean	SD	Min	Median	Max
Network in degree	MX	0.14	0.87	0.00	0.00	15.17
	US	0.36	2.17	0.00	0.01	67.43
Network out degree	MX	0.14	0.79	0.00	0.00	14.65
	US	0.36	1.91	0.00	0.01	56.15
Network clos. centr.	MX	0.16	0.18	0.00	0.00	0.55
	US	0.34	0.16	0.00	0.40	0.68
Network pagerank	MX	0.00	0.00	0.00	0.00	0.06
	US	0.00	0.00	0.00	0.00	0.05

D Feature Descriptions

Table D.1: Survey indicator description

Label	Description
Years of Schooling	Average years of schooling in municipality (MX) or county (US) according to census. We approximate years of schooling for the US by attainment statistics (see main text)
Post Basic Education	Share of population with post basic education
Secondary Education	Share of population with secondary education
Primary Education	Share of population with primary education
Wealth Index	Index based on share of households that have 13 wealth related items according to the Mexican census, sum across standardized items
Bachelor Degree	Share of county level population with some college level education
Some College	Share of population with a bachelor degree
High School	Share of population with high school education
Income	Income statistics provided by US census
Population	Population counts according to census

Table D.2: Network indicator description

Label	Description
Network in degree	Number outgoing references measured by mentions and quotes (log scale)
Network out degree	Number incoming references measured by mentions and quotes (log scale)
Network clos. centr.	Pagerank for municipalities (MX) or counties (US) according to respective network based on mentions and quotes (log scale)
Network pagerank	Closeness centrality for municipalities (MX) or counties (US) according to respective network based on mentions and quotes (log scale)

Table D.3: Twitter penetration and usage indicator description

Label	Description
Tweet count	Number of tweets
User count	Number of users
Share weekdays	Share of tweets created during weekdays (Monday-Friday)
Share workhours	Share of tweets created during workhours (Monday-Friday, 8:00am-4:00pm))
Follower count	Median number of followers per user (log scale)
Following count	Median number of friends per user (log scale)
Tweet count	Median number of tweets per user (log scale)
User mobility	Average number of municipalities (MX) or counties (US) users tweet from (log scale)
iPhone share	Share of tweets sent from an iPhone
Instagram share	Share of tweets sent via Instagram (log scale)
Favorites per tweet	Number of likes per tweet, median (log scale)
Tweets per year	Median number of tweets per year (log scale)
Account age	Age of average account

Table D.4: Twitter penetration and usage indicator description

Label	Description
Account age	Age of average account
Listed count	Average number of public lists user is a member of (log scale)
Followers per following	Number of followers divided by number of accounts a user follows, median (log scale)
Share quotes	Share of tweets that are quotes (log scale)
Share replies	Share of tweets that are replies (log scale)
Share verified	Share of verified users (log scale)
Tweet length	Average number of characters per tweet (log scale)
Hashtags per tweet	Average number of hashtags per tweet (log scale)
Mentions per tweet	Average number of mentions per tweet (log scale)
Urls per tweet	Average number of urls per tweet (log scale)
Emojis per tweet	Number of emoji per tweet (log scale)

Table D.5: Error indicator description (both countries)

Label	Description
Error total	Number of errors per character (log scale)
Error casing	Casing error (log scale)
Error confusions	Word confusions (log scale)
Error grammar	Grammar error (log scale)
Error variants	Errors regarding American and British English (log scale)
Error misc	Miscellaneous error (log scale)
Error punctuation	Punctuation error (log scale)
Error repetitions style	Style error related to repetitions (log scale)
Error semantics	Semantic error (log scale)
Error style	Style error (log scale)
Error typography	Typography error (log scale)
Error typos	Typo (log scale)

Table D.6: Error indicator description (country-specific)

Label	Description
Error noun agreement	Noun verb agreement error (log scale)
Error verb agreement	Verb subject agreement error (log scale)
Error norm change	Deviation from linguistic norms (log scale)
Error collocations	Collocation error (log scale)
Error compounding	Compounding error (log scale)
Error context	Context dependent error (log scale)
Error diacritics	Errors regarding accents (diacritic marks, log scale)
Error expressions	Incorrect expression (log scale)
Error misspelling	Misspelling (log scale)
Error nonstandard	Error related to non-standard English (log scale)
Error prepositions	Error related to prepositions (log scale)
Error proper nouns	Error related to proper nouns (log scale)
Error redundancy	Redundancy in text (log scale)
Error redundancy	Redundancy in text (log scale)
Error repetitions	Repetition in text (log scale)

Table D.7: Topic indicator description

Label	Description
Topic arts & culture	Share of tweets classified into the arts & culture topic (log scale)
Topic business	Share of tweets classified into the business & entrepreneurs topic (log scale)
Topic celebrity	Share of tweets classified into the celebrity & pop culture topic (log scale)
Topic daily life	Share of tweets classified into the diaries & daily life topic (log scale)
Topic family	Share of tweets classified into the family topic (log scale)
Topic fashion	Share of tweets classified into the fashion & style topic (log scale)
Topic films	Share of tweets classified into the films, tv & video topic (log scale)
Topic fitness & health	Share of tweets classified into the fitness & health topic (log scale)
Topic food & dining	Share of tweets classified into the food & dining topic (log scale)
Topic gaming	Share of tweets classified into the gaming topic (log scale)

Table D.8: Topic indicator description

Label	Description
Topic educational	Share of tweets classified into the learning & educational topic (log scale)
Topic music	Share of tweets classified into the music topic (log scale)
Topic news	Share of tweets classified into the news & social concern topic (log scale)
Topic hobbies	Share of tweets classified into the other hobbies topic (log scale)
Topic relationships	Share of tweets classified into the relationships topic (log scale)
Topic science	Share of tweets classified into the science & technology topic (log scale)
Topic sports	Share of tweets classified into the sports topic (log scale)
Topic travel	Share of tweets classified into the travel & adventure topic (log scale)
Topic youth	Share of tweets classified into the youth & student life topic (log scale)

Table D.9: Sentiment indicator description

Label	Description
Sentiment negative	Average share of tweets with negative sentiment in contrast to positive and neutral
Sentiment positive	Average share of tweets with positive sentiment in contrast to negative and neutral
Hate speech	Score indicating hate speech, average (log scale)
Offensive language	Score indicating offensive language, average (log scale)