

TECHNICAL APPENDIX: CREATING THE DATASET OF LAND-USE REFORMS

Part A: Selecting the News Articles

To select the news articles that would represent reforms, we first needed to select the newspapers from which we would search, the time period, and the search terms we would use to identify the subset of articles to tag for the analysis.

Selecting the newspapers and time period

The team met multiple times and solicited advice from an expert panel to select these parameters. Using data provided by Access World News, we assigned newspapers to metro areas based on their city name. To do this, we first matched city names to 2014 Census Places by cleaning the Census and newspaper place names, and matching based on exact match of city and state, where possible. Where this was not possible, we used a fuzzy matching method by applying a standard Jaro-Winkler algorithm to compare string similarity among the remaining unmatched places from the Access World News data where the match score exceeded 0.9705. Using this method, we were able to match 2,759 of 3357 places from the Access World news database, for an error rate of 17.8%.

We combined the Access World News data with Census data and HPI index data from Zillow and calculated several at the metro level to help us to choose the top 40 metro areas to include in the analysis. We chose 40 metro areas because it was the maximum number of news sources, we could acquire from Access World News as part of our project budget.

We observed that news coverage became more and more sparse over time, such that there was much less news coverage before the year 2000, and much more after 2005. Therefore, to be conservative, but also allow for a longer time period – two real estate cycles – we would conduct the analysis from 2000 to the present and analyze cities based on their coverage since the year 2000 and 2005.

To choose the 40 metro areas for the analysis, we decided to choose metro areas with relatively better news coverage and higher population growth – in other words, areas that might be facing pressure to expand housing supply with extensive local news coverage. We selected all newspapers in the Access World News dataset that we could identify from the place matching method, located in these 40 metro areas – a total of 2,147 newspapers. The final 40 metro areas chosen are as follows, with the pertinent – though not exhaustive – set of statistics attached.

Table A1: Selected CBSAs

CBSA	CBSA Name	Average News Coverage Days Rank	Population Change, 1990 to 2000	Population Change, 2000 to 2010
16980	Chicago-Naperville-Elgin, IL-IN-WI (Metro)	1	11%	4%
35620	New York-Newark-Jersey City, NY-NJ-PA (Metro)	2	9%	3%
37980	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD (Metro)	5	5%	5%
33460	Minneapolis-St. Paul-Bloomington, MN-WI (Metro)	7	17%	10%
31080	Los Angeles-Long Beach-Anaheim, CA (Metro)	4	10%	4%
47900	Washington-Arlington-Alexandria, DC-VA-MD-WV (Metro)	15	16%	17%
14460	Boston-Cambridge-Newton, MA-NH (Metro)	3	6%	4%
12060	Atlanta-Sandy Springs-Roswell, GA (Metro)	9	38%	24%
33100	Miami-Fort Lauderdale-West Palm Beach, FL (Metro)	8	23%	11%
41860	San Francisco-Oakland-Hayward, CA (Metro)	6	12%	5%
17140	Cincinnati, OH-KY-IN (Metro)	14	9%	6%
42660	Seattle-Tacoma-Bellevue, WA (Metro)	13	19%	13%
18140	Columbus, OH (Metro)	10	15%	14%
19100	Dallas-Fort Worth-Arlington, TX (Metro)	24	29%	23%
12580	Baltimore-Columbia-Towson, MD (Metro)	12	7%	6%
26420	Houston-The Woodlands-Sugar Land, TX (Metro)	22	25%	26%
40140	Riverside-San Bernardino-Ontario, CA (Metro)	17	26%	30%
28140	Kansas City, MO-KS (Metro)	27	12%	11%
45300	Tampa-St. Petersburg-Clearwater, FL (Metro)	36	16%	16%
38060	Phoenix-Mesa-Scottsdale, AZ (Metro)	26	45%	29%
48620	Wichita, KS (Metro)	21	12%	9%
16740	Charlotte-Concord-Gastonia, NC-SC (Metro)	33	28%	29%
38900	Portland-Vancouver-Hillsboro, OR-WA (Metro)	39	27%	15%
35840	North Port-Sarasota-Bradenton, FL (Metro)	18	21%	19%
36740	Orlando-Kissimmee-Sanford, FL (Metro)	45	34%	30%
10900	Allentown-Bethlehem-Easton, PA-NJ (Metro)	52	8%	11%
28940	Knoxville, TN (Metro)	23	15%	12%
40900	Sacramento--Roseville--Arden-Arcade, CA (Metro)	43	21%	20%
46060	Tucson, AZ (Metro)	34	27%	16%
49340	Worcester, MA-CT (Metro)	32	6%	7%
12420	Austin-Round Rock, TX (Metro)	49	48%	37%
19740	Denver-Aurora-Lakewood, CO (Metro)	62	31%	17%
27260	Jacksonville, FL (Metro)	40	22%	20%
29820	Las Vegas-Henderson-Paradise, NV (Metro)	25	86%	42%
24340	Grand Rapids-Wyoming, MI (Metro)	51	18%	6%
13820	Birmingham-Hoover, AL (Metro)	76	10%	7%
15180	Brownsville-Harlingen, TX (Metro)	56	29%	21%
25420	Harrisburg-Carlisle, PA (Metro)	66	7%	8%

31540	Madison, WI (Metro)	67	16%	13%
41940	San Jose-Sunnyvale-Santa Clara, CA (Metro)	38	13%	6%

Selecting the search terms

After selecting the newspapers, we needed to select the search terms that would help us narrow down the articles from to make the manual tagging more fruitful and reduce computational time. The process of choosing search terms involved consultation with the research team and a group of external experts in land-use research and policy. The research team created a list of important local regulations that might affect housing production and received feedback from the advisory group to expand that list to the full set of options. Then, the research team each ranked land-use policies in order of importance from 1 (most important) to 5 (least important). All land-use policies marked as priority 1 by the research team were included in the analysis. We decided not to include all potential policies because additional policies beyond priority 1 would require additional manual tagging of data (described in detail in Part B), which would have added an unacceptable level of additional cost to our budget.

After choosing the 21 land-use policies to study, the team constructed and edited search terms in consultation with the external advisory group. After multiple iterations of testing the results using the Access World News online interface, the team determined that the following search terms restricted the results to a reasonable set without excluding articles that referenced true land-use reforms. Note that we combined all policies into a single, large search term, and added an “AND” statement restricting the articles to those that also included legislative language indicating some sort of administrative or legislative action.

```
(
  (
    ("zoning" OR "land use") AND ("single family homes" OR "single family
    dwellings" OR "single family residences")
  ) OR
  "minimum lot size" OR ("lot size" AND "planning") OR
  (
    "zoning" AND ("height limit" OR "height regulation" OR "height restriction")
  ) OR
  "floor area ratio" OR "floor space ratio" OR "floor space index" OR
```

"site density" OR "site ratio" OR "bulk control" OR "volume control" OR
 "urban growth boundary" OR "urban growth zone" OR
 ("zoning" AND "manufactured housing") OR ("zoning" AND "mobile homes") OR
 (
 ("building code" OR "building codes") NEAR7 ("passed" OR "passes" OR "enacts"
 OR "enacted" OR "legislation" OR "adopts" OR "adopted" OR "approved" OR
 "approves" OR "legalizes" OR "legalized" OR "voted")
) OR
 "land assembly" OR ("zoning" AND ("subdivision rule" OR "subdivision rules" OR
 "subdivision ordinance")) OR
 "accessory dwelling units" OR ADU OR "accessory apartments" OR "secondary
 dwellings" OR "secondary units" OR "granny flat" OR ("zoning" AND ("minimum unit
 size" OR "tiny homes" OR "tiny house")) OR
 "upzoning" OR "downzoning" OR "upzone" OR "downzone" OR
 "minimum parking requirements" OR (minimum NEAR parking) OR "parking
 requirement" OR "parking requirements" OR ("zoning" AND "transit oriented
 development") OR ("zoning" AND "TOD") OR
 ("zoning" AND ("building moratorium" OR "development moratorium")) OR
 (
 ("zoning" OR "planning department") AND ("minimum setback" OR "minimum
 setbacks")
) OR
 (
 "occupancy requirement" AND ("planning" OR "zoning")
) OR
 "inclusionary zoning" OR "inclusionary housing" OR "mandatory affordable housing" OR
 "affordable housing requirement" OR
 ("new zoning" AND "planning" AND "mixed use") OR
 "entitlement permitting" OR ("zoning" AND "permit" AND ("streamlining" OR
 "consolidation" OR "entitlement")) OR
 “rent control” OR “rent controlled” OR “decontrol”

)

AND ("passed" OR "passes" OR "enacts" OR "enacted" OR "legislation" OR "adopts" OR "adopted" OR "approved" OR "approves" OR "legalizes" OR "legalized" OR "voted")

We combined each individual search term we selected in a master search term, which produced a total of 76,410 results from January 1, 2000 to January 13, 2019.

Part B: Manual Labeling

To produce a useful dataset of reforms across 40 metro areas and more than 76,000 news articles within the time and budget constraints of this project, we rely on a machine learning algorithm to automatically tag the algorithms, rather than a manual, expert driven process.

However, the machine needs a “training set” of expert labeled data from which to learn to properly tag the full set of news articles. To accomplish this task, we trained a team of four manual taggers with a background in housing and/or land-use policy analysis to tag as many news articles as possible within the constraints of this project.

In total, the expert taggers recorded information for 568 news articles that they randomly selected from the Access World News online search tool as the full manually tagged dataset. For each of the land-use reforms referenced in Part A, the taggers used the search term for that specific reform type, entered it into the Access World News search tool, and tagged at least 20 articles, to ensure each reform type received sufficient manual tags to train the algorithm. The taggers recorded the following information for each news article, with specific instructions as referenced below.

Table A2: Tagging Variables

Term	Definition
jurisdiction_of_reform	Jurisdiction in which the reform took place
state_of_reform	State in which the reform took place
geo_extent	Geographic extent to which the reform applies
Month_takes_effect	Which month it will take effect (if noted) -- marked by 1 for Jan, 2 for Feb, etc.
Year_takes_effect	Year reform will take effect
Is_Reform	Does the article indicate an actual reform? Yes= 1, No = 0, Future vote =2
Is_Major	Does this article mark a truly major change in zoning or land use? If you're not sure, the answer is No. Yes=1, No=0
change_operator_geographic	Did the reform change the geographic extent to which a code applies? 0 is shrink, 1 is same, 2 is expanded
change_operator	Did the reform change the degree of regulation over design, process, or entitlement? 0 is more restrictive, 1 is same, 2 is less restrictive
Tag	Which code the reform targets May be multiple - multiple reforms in one article should be multiple rows
Notes_on_words	Notes on particular words or phrases that indicate reform, tags, etc.

Table A3: Instructions for Tagging

Tags	+ Restriction (0)	- Restriction (2)	+ geography	- geography
single-family housing	lower lot coverage %, more sight line restrictions, occupant #s, business ban, traffic impact, yard requirement, lot depth requirements	higher lot coverage allowed, no sight lines, higher occupant #s, business allowed, traffic ok; eliminated, struck down less greenspace required, no design review, higher density limit, larger building width	Increase, expand, enlarge, create	Reduce, shrink, eliminate
multi-family housing	more greenspace required, design review longer, lower density limit, smaller building width	design review, higher density limit, larger building width	Increase, expand, enlarge, create	Reduce, shrink, eliminate
lot sizes	minimum allowed size increased	minimum allowed size decreased	Increase, expand, enlarge, create	Reduce, shrink, eliminate
height limit	lower height limit	higher height limit	Increase, expand, enlarge, create	Reduce, shrink, eliminate
density or Floor-Area-Ratio (FAR)	lower density or FAR	higher density or FAR	Increase, expand, enlarge, create	Reduce, shrink, eliminate
mobile/manufactured homes	older homes not allowed on certain lots, higher lot size requirement, restrict to parks non-conforming use for removeable home, ban	allow older homes on land, lower lot sizes, allow removable home, eliminate restriction to parks, lift ban	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Ministerial approvals, by-right conditionalities, non-conformities			Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate

subdivision or land assembly conditions	Limiting purposes of subdivision/land assembly, limiting number of subdivisions of one plot or number of plots combinable into single lot, occupancy restrictions on divided/assembled plots, fees to government or former plot owners	Remove restrictions on subdivision/land assembly purposes, allow multiple subdivision or lot combinations, no occupancy restrictions, no fees or ways one owner can block deal	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Building code	increase restrictions on interior finishes, fire and smoke protection, egress, accessibility, exterior walls, energy efficiency, roofing, soil, foundations, electrical, plumbing, public right of way	remove restrictions on interior finishes, fire and smoke protection, egress, accessibility, exterior walls, energy efficiency, roofing, soil, foundations, electrical, plumbing, public right of way	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
minimum dwelling unit sizes	raise the minimum sq. ft of floor space per occupant	Lower the minimum sq. ft of floor space per occupant	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
off-street parking requirements	Raise number of parking spaces/size of space required per unit	lower number of parking spaces/size of spaces required per unit	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Accessory Dwelling Units/ADUs/Granny Flats	raise minimum lot sizes or setbacks, higher occupancy requirements on accessory dwelling units.	lower minimum lot sizes or setbacks, lower occupancy requirements for ADUs	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Mixed residential and non-residential development	reduced types of uses allowed in a zone (residential and commercial to only commercial), increased specificity on types of allowed uses in mixed use zones	Lowered restrictions on allowed use (industrial, residential, and commercial uses allowed), lower specificity in	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate

allowed-use sub-
types

Growth boundaries or greenbelts	Growth boundary moved inwards, made permanent, allowed uses specified more tightly	Growth boundary expanded outwards, allowed uses inside and outside not dictated, zoning restrictions outside of boundary eliminated Quota eliminated, moratorium removed/abolished, restricted in application to fewer buildings, allows more buildings	Expanded, enlarged, annexed land, pushed out	Shrunk, reduced
Development moratorium, growth cap, quotas	Quota lowered, moratorium extended in time, expanded in what it applies to		Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
minimum setbacks	minimum setback increased (# ft)	minimum setback decreased (# ft lower) looser	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Occupancy requirement	More people required to live permanently on property (in owner's family); Relationships of acceptable occupants defined; Cap # of people allowed to live per sq. foot	requirements on # people required to live permanently on property, allow more people per sq. ft of housing space	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
UDO/streamlined land use and development entitlement permitting	Fewer procedures included in UDO, taken out of unified development ordinance; new processes or reviews added prior to entitlement or permitting	More processes added into UDO, fewer unique permits or approvals to acquire, fewer hearings	Increase, expand, enlarge, create	Increase, expand, enlarge, create

		Voluntary; higher # of units trigger AH req; lower fees or AH units required per unit of market development; lower AH equivalency requirements (offsite, less expensive materials); higher AMI levels		
Inclusionary zoning/mandatory affordable housing ratio	Made mandatory; lower # of units triggers affordable housing requirements; more AH required per unit; higher fees per unit; higher AH quality equivalency requirements (e.g. onsite, same materials); lower AMI requirements		Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Density bonuses or incentives	Increase or enact density bonus, such as increased height, FAR, parking, etc. incentives	Decrease or abolish density bonus or incentives	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Discretion/standardized criteria in permitting	Increase discretion of city in permitting	Decrease discretion of city in permitting Removing rights to appeal, shortening process or adding required deadlines, removing veto powers	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Appeals process	Adding additional rights to appeal, lengthening process, allowing additional veto powers		Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Number of hearings	Increase the number of hearings	Decrease the number of hearings Decrease costs to landlords, fewer requirements on landlords, reduce time window, abolish or limit existing rent control, increase allowed rent changes	Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate
Rent Control	Increase costs to landlords, more requirements on landlords, extend time window of rent control, enact new rent control requirements, reduce allowed rent changes		Increase, expand, enlarge, create	Reduce, restrict, shrink, eliminate

Each key variable for tagging was chosen and refined by the research team, with feedback from the advisory group of external experts. Note that only articles tagged as “is_reform” of 1 or 2 were

tagged with additional information, otherwise the tagger moved to the next article. Tagger quality was reviewed by one primary tagger, but only a random subset of 100 were verified and corrected if errors or inconsistencies were found.

Part C: Analysis Methods

Analysis Environment

The following analysis was completed solely on a remote Access World News server named the “walled garden” environment. As part of the research team’s agreement with Access World News, no news content was to leave the walled garden environment. The Access World News team provided pre-installed Python and Docker utilities we used to complete the analysis.

CoreNLP Analysis

The research team used [Stanford CoreNLP](#) (version 3.9.1) a natural language software tool, to analyze and parse the news text. The Stanford CoreNLP server was provided by Access World News technical staff as a Docker container within the walled garden environment. In the Python programming language, we read in each news article from the search results and sent the text to the CoreNLP server. CoreNLP analyzed each article word by word, and returned the base forms of the words, their parts of speech, and named entity recognition type – whether the words are names of companies, people, date, etc. We cleaned the text of the articles based on the CoreNLP results, which included removing punctuation, labeling words by parts of speech, and labeled words that are numbers and dates.

TF-IDF Feature Extraction

Term Frequency-Inverse Document Frequency (TFIDF) is used to extract important keywords from a collection of articles. After parsing the text with CoreNLP, we used TFIDF to generate an initial list of keywords. We calculate TFIDF analysis for keyword-part of speech combinations, and for parts of speech alone. For the keyword-part of speech combination, we use TFIDF ngrams of length 2 and 4, while for the parts of speech, we use TFIDF ngrams of lengths 2, 3, and 4. We chose these lengths because keyword-part of speech combinations of these lengths represents the equivalent of unigrams and bigrams for keywords alone. We chose ngrams of length greater than one for parts of speech to include certain combinations of the way articles talk about reforms that might differ by class categories.

A keyword-part of speech combination represents one continuous string of text, where each word is followed by its part of speech (e.g., “building nns moratorium nns”). Using keyword-part

of speech combinations instead of keywords alone helps us separate similar words with different functions. All keywords used represent lemmas calculated by CoreNLP, not the original keyword, to standardize keywords with the same meaning.

Using these inputs, we calculated the difference in the TFIDF statistic for each word between one class label and the remainder of the class labels to choose top keywords as independent variables for our machine learning model, as follows.

For example, when choosing keywords to help us predict whether an article was referencing a land-use reform, we conducted TFIDF analysis separately for articles that are labelled as reform and not reform to create two lists of keywords. We then analyzed each keyword from the two lists and counted the number of articles that include the keyword in each group of articles – reform versus not reform – dividing that frequency by the total number of articles in each group. Next, if the word appeared in both groups of the articles, we ranked the words by frequency difference – from highest to lowest. We took the top 15 percent of the keywords that have the highest frequency differences, combined with the words that only appear in one group of the articles – reform vs not reform – and included them as independent variables in the machine learning model. Note that we chose 15 percent using expert judgment given it provided a large number of features such that we could differentiate the news articles and not too many where we might be most likely to overfit.. For each independent variable (is_reform, reform_type, etc.), we generated a different list of top keywords to use as independent variables in our model. In addition to keywords, we repeat this same method for parts of speech for each predicted variable to generate a separate list of parts of speech. The independent variables for parts of speech and keywords are then combined and used to predict each dependent variable in our chain of models.

Create Independent Variables for Machine Learning Model

Our goal is to use a machine learning model to manually tag articles. In our case, the machine learning model consists of the tag as the dependent variable – for example, is_reform = 1 or is_reform = 0 – and hundreds of independent variables, which we generate from several sources, as follows. The number of independent variables varies for each dependent variable predicted in our chain of models.

- 1) According to the top keyword-part of speech combinations and parts of speech we generated from TFIDF, we created independent variables indicating how frequent the word appears in the article – the number of times the word appears divided by the length of the news article in words;

2) We use the list of manually selected keywords from the search query related to specific housing reforms – e.g., single family home, minimum lot size, subdivision rules, etc. – and create dummy variables that take the value of 1 if they appear in the news article and 0 otherwise. This technique helps us to predict the tag of the reform;

3) Dummy variables indicating if certain named entities, which include organization names, dates, location names, and ordinal numbers, appear in the article;

4) Using the part of speech variables from CoreNLP, we calculate the percentage of words in the article that are past tense verbs, present tense verbs, or future tense verbs.

Machine Learning Model

We randomly divided the 568 news articles into a training set and test set – 80 percent of the observations into training set and 20 percent into the test set. Tests were performed on the training set using 10-fold Cross Validation, which is a procedure by which we split the dataset into 10 random pieces, (datasets), each with the same number of rows, and generate accuracy scores by iterating through each piece. The first time around, we treated the first piece as a "test" dataset with which we tested the accuracy of our model, and the other 9 pieces as our "training" dataset, from which the machine learning model would learn the patterns in the data and predict the results for the "test" piece. We then repeated this procedure for the remaining 9 pieces, ran this process 10 times in total, and calculated the average accuracy across all 10 runs to test out of sample accuracy and avoid overfitting. Ultimately, because the sample size was small, we use 10-fold cross-validation on all 568 articles to generate final accuracy scores. Before doing so, we ensured all tests were run without training on any out of sample data to ensure we preserved out of sample accuracy, we. For binary predictions – where the dependent variable takes the value of 0 or 1 – we use a random forest algorithm with 2000 trees and default scikit-learn parameters from version 0.20. Random forest, a commonly used machine learning algorithm for classification, consists of a large number of decision trees. Each decision tree operates in a relatively unrelated manner with each other using different subsets of the data and provides its own prediction result. The random forest model then consolidates all the trees together to produce a final prediction. as an average across all individual trees' predictions. For multiclass predictions – where the dependent variable can take more than 2 values – we use the same random forest parameters wrapped with the One-vs-the-rest classifier with default scikit-learn parameters from version 0.20. We predicted each outcome in the following order:

Variable predicted	Number of independent variables	Number of dependent variable unique values
Is reform or not	934	2
Reform type	1077	2
Major reform	1088	2
Change operator	1012	3
Change operator geographic	1807	3
Tags	1918	23
Geographic extent	1425	3

After manually tagging additional articles in our full dataset, the team determined that there was a large false positive rate among articles identified as major, passed, reforms and larger inaccuracies in the reform change operator variable. As a result, the research team tagged additional articles and using these articles combined with true reforms from the initial tagging exercise, trained a second machine learning model to predict major, passed reforms. The change operator was trained solely using major, passed reforms from the manual tags. Note that each model was tuned to reduce false negatives, which had the effect of a small increase in false positives and a very small decrease in overall accuracy.

Variable predicted	Number of independent variables	Number of dependent variable unique values
Is reform, AND passed, AND major	972	2
Change operator (on only real, passed, major reforms)	1243	2

Rule-Based Tagging Model

For certain variables that involved a large set of potential tags and were therefore infeasible to extract with machine learning techniques, we used a rule-based tagging algorithm: the estimated city and county of reform, and the estimated date of reform. To choose the estimated date of reform, we chose the most commonly occurring word or phrase marked as a “Date” by CoreNLP’s named entity recognition system.

We used a more complex process to estimate the city and county of reform. We first compile a master list of city and county names by metro area, state, and the United States as a whole. If we can locate the publication location of the news article within a given metro area, which we can for the vast majority of publications (see the following section for more information), we select the city and county mentioned in the article with the greatest frequency that is also a city or county within the metro area of the publication. If the publication cannot be located within a given metro area but can be for a given state, we skip the metro area and instead select the city and county mentioned in the article with the greatest frequency that is also a city or county within the state of the publication. If the publication cannot be geographically located, we choose the city and county mentioned in the article with the greatest frequency that is also a valid city or county in the United States.

In future work, we would like to apply more complex and accurate techniques to better capture the location and date of reform from the article, such as using a combination of machine learning and rule based techniques to better identify the date relative to the date of the publication in the news article when terms like “last week” or “next month” are used.

Enriching with Geographic Information

For each article, we receive the publication name, city, and state from Access World News. To better link the news data to existing data sources at defined Census geographies, we add the following geographic information to the Access World News data based on the location of publication: Census Place, Census Place ID, Census County ID, Core Based Statistical Area (CBSA) ID, CBSA Name, Census Place name, County name, Census State ID, State name. We also add the following geographic identifiers for the estimated location of reform, discussed in the previous section: Census Place ID and Census County ID.

To enrich the publication location data, we collect source data on 2014 Census Place definitions, 2015 CBSA definitions, and 2014 county definitions from the [Missouri Census Data Center Geocorr 14 tool](#) (referred to as “Geocorr data” below). We clean the Access World News publication city name by converting city names to lowercase, simplifying publications with multiple cities to the first city, and removing any non-standard characters. We also filter the Geocorr data by keeping the places that have an [allocation factor](#) that is higher than 0.5, and standardize the place names by removing the “city/town/cdp” text at the end of place names. We

then merged the publication locations with the Geocorr data using the state and city names, and only keep the publications that we are able to locate within a metropolitan area.

For the estimated location of reform, we simply merge the standardized city and county name with their respective IDs from the Geocorr file.

Deduplication

Different news outlets may use exactly or somewhat similar text to describe the same issue. On manual inspection of our dataset, we found that the majority of cases of duplication we encountered were re-prints of print articles in an online edition, with some re-prints of similar articles across newspapers.

To signal this similarity to researchers, we provide a deduplication ID that takes the same value for all articles we believe are substantially similar. We use the following procedure to assign deduplication IDs:

First, we put the articles that have the same publication city and whose publication dates are within 6 months into the same group. Then, we compare the articles within the same groups using the “[SequenceMatcher](#)” method in the Python programming language (an variation on the Ratcliff-Obershelp algorithm), which returns us a similarity score between 0 to 1. If the similarity score between two articles is higher than 0.5, these two articles are assigned the same deduplication ID.

Part D: Results

Based on our model, 47.9 percent of the news articles are labelled as housing reform related.

Model Accuracy

The accuracy for each model is listed in the table below, relative to the baseline accuracy, which results from choosing the most common class as the predicted value for all rows:

Model Type	Base Line Accuracy ¹	Accuracy
Reform or not	59.51%	78.51%
Reform type	59.76%	82.54%
Major reform	68.80%	76.83%
Change operator	48.24%	74.71%
Change operator geographic	87.39%	88.84%
Tags	11.61%	62.89%
Geo extent	84.58%	91.02%
Is Reform, Passed, AND Major	55.79%	76.44%
Change operator, Reform, Passed, AND Major only	59.24%	77.08%

To ensure accuracy of major reforms out of sample, we manually coded an additional 50 articles, 25 marked as a passed, major reform, by the variables “Reform or not”, “Reform type”, and “Major reform”, and by the variable “Is Reform, Passed, AND Major”, and 25 that were not marked as such. Of the 25 marked as passed, major reforms, 11 (44%) were true passed, major reforms. Of the 25 not marked as such, only 2 (8%) were true major reforms. In other words, this small sample, the algorithm missed 2/13 (15%) of true reforms.

Confusion Matrix

For each variable predicted, we provide the prediction accuracy on the test set for the combination of all 10 of the 10-folds predicted out of sample. The total number of articles in the models predicting the reform characteristics (e.g., reform type, major reform) may be smaller than the number of articles that are predicted as reform, because the researchers who manually labelled

¹ The percentage of the most frequent value of a variable.

the dependent variables in the training and test sets were not certain about the characteristics of a small number of reforms. The number of articles in the model predicting the geographic extent of the article is larger than the number of articles predicted as reform because we predicted the geographic extent for all articles in the training and test set.

Is reform:

		Predicted value		
		Not reform	Is reform	Total
True value	Not reform	150	80	230
	Is reform	42	296	338
	Total	192	376	568

Reform Type:

		Predicted value		
		Current reform	Future vote	Total
True value	Current reform	182	20	202
	Future vote	39	97	136
	Total	221	117	338

Major Reform:

		Predicted value		
		Major reform	Not major reform	Total
True value	Major reform	58	59	117
	Not major reform	28	230	258
	Total	86	289	375

Change Operator:

		Predicted value			
		0: more restrictive	1: same	2: less restrictive	Total
True value	0: more restrictive	142	0	22	164
	1: same	8	0	20	28
	2: less restrictive	36	0	112	148
	Total	186	0	154	340

Change operator geographic:

		Predicted value			
		0: shrink	1: same	2: expanded	Total
True value	0: shrink	0	10	1	11
	1: same	0	305	0	305
	2: expanded	1	27	5	33
	Total	1	342	6	349

Geographic Extent:

		Predicted value			
		City	County	State	Total
	City	391	4	0	395

True value	County	29	34	0	63
	State	9	0	0	9
	Total	430	37	0	467

Tags:

Since we have 23 different tags, we present the confusion matrix for each value separately below to avoid visual confusion. Note that in each table, the column names are the predicted values, and the row name represents the true value.

		Predicted value		
		Accessory Dwelling Units/ADUs/Granny Flats	Occupancy requirement	Total
True Value	Accessory Dwelling Units/ADUs/Granny Flats	36	1	37

	Accessory Dwelling Units/ADUs/Granny Flats	Total
Appeals process	1	1

	Accessory Dwelling Units/ADUs/Granny Flats	Building code	Development moratorium, growth cap, quotas	density or Floor-Area-Ratio (FAR)	height limit	lot sizes	mixed residential and non-residential development	mobile/manufactured homes	Growth boundaries or greenbelts	Inclusionary zoning/mandatory affordable housing ratio	Total
Building code	2	5	2	6	3	2	6	1	1	1	29

	Development moratorium, growth cap, quotas	density or Floor-Area-Ratio (FAR)	mobile/manufactured homes	Total
Development moratorium, growth cap, quotas	23	3	1	27

	density or Floor-Area-Ratio (FAR)	Total
Discretion/standardized criteria in permitting	1	1

	Growth boundaries or greenbelts	Total
Growth boundaries or greenbelts	12	12

	Inclusionary zoning/mandatory affordable housing ratio	density or Floor-Area-Ratio (FAR)	mixed residential and non-residential development	Total
Inclusionary zoning/mandatory affordable housing ratio	22	2	1	25

	Accessory Dwelling Units/ADUs/Granny Flats	density or Floor-Area-Ratio (FAR)	mixed residential and non-residential development	Total
Occupancy requirement	2	1	2	5

	Development moratorium, growth cap, quotas	Rent Control	Total
Rent Control	1	9	10

	density or Floor-Area-Ratio (FAR)	mixed residential and non-residential development	Total
UDO/streamlined land use and development entitlement permitting	2	2	4

	Accessory Dwelling Units/ADUs/Granny Flats	Total
accessory Dwelling Units/ADUs/Granny Flats	1	1

	Accessory Dwelling Units/ADUs/Granny Flats	Discretion/standardized criteria in permitting	density or Floor-Area-Ratio (FAR)	mixed residential and non-residential development	height limit	Total
density or Floor-Area-Ratio (FAR)	1	1	36	1	2	41

	Development moratorium, growth cap, quotas	density or Floor-Area-Ratio (FAR)	height limit	mixed residential and non-residential development	Total
height limit	2	4	25	1	32

	Accessory Dwelling Units/ADUs/Granny Flats	Development moratorium, growth cap, quotas	Growth boundaries or greenbelts	density or Floor-Area-Ratio (FAR)	height limit	lot sizes	mixed residential and non-residential development	Total
lot sizes	2	1	1	2	1	7	5	19

	Accessory Dwelling Units/ADUs/Granny Flats	density or Floor-Area-Ratio (FAR)	minimum dwelling unit sizes	Total
minimum dwelling unit sizes	1	1	16	18

	Development moratorium, growth cap, quotas	density or Floor-Area-Ratio (FAR)	height limit	lot sizes	mixed residential and non-residential development	Total
minimum setbacks	2	1	4	1	2	10

	Accessory Dwelling Units/ADUs/Granny Flats	Development moratorium, growth cap, quotas	Inclusionary zoning/mandatory affordable housing ratio	density or Floor-Area-Ratio (FAR)	height limit	lot sizes	mixed residential and non-residential development	off-street parking requirements	subdivision or land assembly conditions	Total
mixed residential and non-residential development	1	3	1	4	4	1	12	1	3	30

	minimum dwelling unit sizes	mixed residential and non-residential development	mobile/manufactured homes	Total
mobile/manufactured homes	1	1	13	15

	Building code	Development moratorium, growth cap, quotas	height limit	minimum setbacks	mixed residential and non-residential development	Total
multi-family housing	1	1	1	1	1	5

	Accessory Dwelling Units/ADUs/Granny Flats	density or Floor-Area-Ratio (FAR)	mixed residential and non-residential development	off-street parking requirements	Total
off-street parking requirements	1	1	1	5	8

	Building code	Development moratorium, growth cap, quotas	Total
permitting fees	1	2	3

	Development moratorium, growth cap, quotas	density or Floor-Area-Ratio (FAR)	lot sizes	mixed residential and non-residential development	subdivision or land assembly conditions	Total