Do local planners outperform a statistical model in selecting site inventories?

Salim Damerdji

February 2023

Abstract

Under California's housing element law, cities demonstrate they have zoned for enough housing by identifying where housing will get built. With a dataset of over 150,000 parcels of land in San Francisco for the fourth and fifth RHNA cycles, I compare how the City of SF fared against a logistic regression model in predicting where housing will be built, and I find that [X]. Additionally, using debiased machine learning, I find that, after controlling for observed confounders, a site's inclusion in the site inventory [is / is not] associated with a higher likelihood of housing being built on the site.

1 Introduction

Every eight years, cities in California update their housing elements to meet new targets for housing production. A centerpiece of every housing plan is an inventory of sites where the city claims it can accommodate enough housing to meet its housing targets. Traditionally, site inventories are selected by local planners using a combination of local knowledge and data-inspired heuristics. More recently, the state's large cities - namely, San Francisco and Los Angeles - have shifted towards using statistical models to predict housing growth.

The first question this paper seeks to answer is whether a statistical model outperforms local planners in classifying which sites are likely to be developed into housing. This binary classification problem involves imbalanced classes, as only a fraction of a percent of parcels are developed into new housing in the span of eight years, the length of a planning period. I compare a logistic regression model to the decisions made by local planners. I find that [X].

The second question this paper seeks to answer is whether the sites selected by planners are associated with a higher rates of housing development after controlling for observed confounders like the parcel's zoning, assessed land value, the age of the property, the neighborhood, and so on. If planners' selection of sites reflects variables not contained in the dataset, then the treatment - viz., inclusion in the site inventory - is not exogenous given observables, and so a

causal interpretation is not warranted. Nevertheless, we can identify the statistical significance of a site's inclusion in the site inventory in a partially linear model produced by Chernozhukov et al.'s double machine learning method, using a gradient boosted classifier for the propensity score model and a gradient boosted regression model for predicting the outcome [Che+18]. After expanding categorical variables with one-hot encoding, the dimensionality of the feature vector is well-over three hundred, without taking into account plausible interactions terms. As a result, using ML to estimate the non-linear portion of this partially linear model reduces variance in estimation given the high-dimensional nuisance parameters.

In short, this paper will ask first whether local planners outperform statistical models; and, if not, then are their decisions associated with *any* statistically significant difference in outcomes when controlling for observed confounders?

2 Data Sources

This project merges several data sources provided by the City and County of San Francisco.

The BlueSky dataset tracks roughly 150,000 parcels from 2001 to 2016 and includes information on the existing building envelope, the potential buildable envelope given the parcel's zoning designation, historical status, and residential status. Buildings with historical status are harder to build on due to California's Environmental Quality Act; buildings with tenants are harder to build on due to San Francisco's robust tenant protections; and the delta between the existing building envelope and the potential building envelope is correlated with the returns of redevelopment. Thus, this dataset provides several variables that a priori are predictive of where housing will get built. This dataset was, in fact, used by San Francisco in its 6th cycle housing element to identify sites to select for its site inventory.

This dataset is joined with data from the county tax assessor in 2007 and 2015, the starts of RHNA4 and RHNA5 respectively. The tax assessor's data includes information on the age of the property, the construction type, the property's square footage, the basement area, lot area, lot shape, the ownership status, the prior sale date of the land, the assessed improvement value, the assessed land value, the number of bedrooms, baths, stories, and units, and more.

Additionally, using San Francisco's Department of Building Inspection's dataset of permits, for each parcel, I know in the preceding eight years how many times the Department of Building Inspection recieved a permit to build, teardown, improve, or alter something on the parcel. A priori, one would think that permits to improve a parcel are negative indicators that the owner is interested in tearing down the property to rebuild. Conversely, it's reasonable that demolition permits are lead indicators for future development on the parcel.

Finally, I use the Association of Bay Area Government's dataset on San Francisco's 4th and 5th cycle site inventories. These inventories identify where

the city claims it can realistically accommodate its housing targets. View as the positive class predictions of a binary classifier, we can compare these predictions against the predictions of a statistical model.

3 Exploratory Data Analysis

[Should I provide summary(df)? Could be overkill.]

3.1 Tax Data

A lot of these variables are right-skewed and log transforms are helpful.
Mention outliers, e.g. 555 California Street.

3.2 Zoning Data

3.3 Pipeline Data

[Insert corrplot for each permit type vs development.]

4 Methods

4.1 Logistic Regression

4.2 Double ML

5 Results

6 Conclusion

References

[Che+18] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.