Highly Disaggregated Land Unavailability*

Chandler Lutz
Securities And Exchange Commission

Ben Sand
York University

Land Unavailability Data:
https://github.com/ChandlerLutz/LandUnavailabilityData

April 2, 2022

Abstract

We combine high-resolution satellite imagery with modern machine learning techniques to construct novel datasets that capture the geographic determinants of U.S. housing supply. This Land Unavailability (LU) measure is a markedly more accurate house price predictor than the popular proxy of Saiz (2010). LU is also uncorrelated with housing demand proxies, supporting its use as an instrument for house prices. We apply LU to fundamental housing finance problems to provide substantially more precise housing wealth elasticity estimates; novel empirical tests of the supply-side speculation theory; and new evidence on the relationship between house prices and self-employment during COVID-19.

JEL Classification: R30, R31, R20;

Keywords: Land Unavailability, House Price Prediction, Buildable Land, Housing Market, Real Estate

*Lutz: lutzc@sec.gov. https://chandlerlutz.github.io/. Sand: bmsand@yorku.ca. https://ben-sand.github.io/. The Securities and Exchange Commission disclaims responsibility for any private publication or statement of any SEC employee or Commissioner. This article expresses the authors' views and does not necessarily reflect those of the Commission, the Commissioners, or other members of the staff.

1 Introduction

As housing is the largest financial asset for most households, economists often estimate the impact of house price fluctuations on broader outcomes. To do so, they typically employ an instrumental variable (IV) approach since reverse causality obfuscates the relationship between house prices and the variable of interest. The most popular such instrument, the Saiz (2010) elasticity IV, stems from supply-side constraints and combines land unavailable due to geographic building restrictions with local housing market regulations.[1]

Yet recently, Davidoff (2016) criticized the Saiz elasticity IV due to its potential correlation with local housing demand proxies, perhaps invalidating it as an instrument for house prices. While Guren et al. (2021, henceforth GMNS), for example, use a rich panel data structure to circumvent Davidoff's criticisms, they also find that the Saiz elasticity IV lacks predictive power. The low predictive power of the Saiz IV in the first stage of a two-stage least squares (2SLS) framework leads to imprecise estimates in the second stage, increasing uncertainty in the causal relationship between house price changes and other economic outcomes. In particular, GMNS note that the Saiz IV uses the land unavailable due to geographic constraints within a 50k km radius of each city's central city centroid. Yet cities vary drastically in size and shape, making the blunt Saiz method inappropriate for many cities. Indeed, a homogeneous approach across the heterogeneous cities, for example, means that Saiz elasticity has relatively large coverage when cities are geographically small (e.g., the northeast) but relatively small coverage when cities are large (e.g., the southwest). Researchers also face other constraints when using the Saiz IV. The Saiz dataset contains a small number of cities, just 270, compared to the Freddie Mac house price dataset used by GMNS that has 380 cities or the Zillow data with over 800 cities. A smaller number of cross-sectional observations increases finite sample bias in 2SLS estimates. Moreover, the

[1]Several recent papers have employed the Saiz elasticity IV or land unavailability in their study of housing markets and the broader economy. For papers on the financial crisis, see Mian and Sufi (2009, 2011); Mian et al. (2013); Mian and Sufi (2014). In entrepreneurship and firm formation, see Adelino et al. (2015). Other and related applications include Chaney et al. (2012); Aladangady (2017); Baum-Snow and Han (2020); Tan et al. (2020); Büchler et al. (2021); Conklin et al. (2021); Gabriel et al. (2021); Gamber et al. (2021); Gupta et al. (2021).

Saiz dataset uses 1999 MSA/NECMA definitions, whereas most current housing datasets use modern CBSA aggregations, perhaps creating measurement error as researchers must port the Saiz proxy to recent delineations.

To overcome these criticisms, we construct new estimates of land unavailability (LU), with complete coverage of the contiguous United States at multiple levels of disaggregation down to zip codes. Our intent is to create plausibly exogenous, highly predictive proxies that researchers can use as instruments for house prices.

We extend the popular Saiz proxy in several directions. First, our data use more accurate satellite imagery that is now available from the United States Geographical Survey (USGS). We also recognize that cities are heterogeneous and that any specific LU proxy is likely measured with error. Thus, in constructing LU, we exploit modern machine learning techniques to combine land unavailability estimates computed at multiple levels of disaggregation. The result is a data-driven LU dataset that accounts for city-level idiosyncrasies and has strong predictive power for house prices.

Our overarching methodology spans housing datasets, requiring only minor modifications for the desired output. Broadly, output is either (1) a panel dataset of LU-based house price predictions where we combine various LU proxies using machine learning algorithms to predict a panel of house prices; or (2) a cross-sectional dataset with a single land unavailability estimate for each geographic unit (e.g., one LU estimate for each CBSA), where the optimal geographic polygon for which to calculate land unavailability is chosen by cross-validation. Either the panel data or the cross-sectional data can be used as an instrument for house prices, depending on the empirical setup. The LU panel dataset is useful when the econometric setup also employs panel data (e.g., GMNS and equation 1 below), while the cross-sectional LU data is apt when the econometric equation of interest is in long differenced form (e.g., Adelino et al. (2015) and equation 3 below).

We first examine the correlation between LU and proxies for housing demand, as using land unavailability as an IV depends on its exogeneity relative to demand factors. As noted above, recent research has suggested that the Saiz elasticity IV is not exogenous, potentially

invalidating its use for causal inference (Davidoff, 2016). Criticisms of deploying land unavailability as an IV contend that households' demand with respect to unobserved factors related to amenities, the economics of agglomeration associated with higher education, and labor demand shocks has been increasing. These factors, thus, cannot be accounted for through standard region fixed effects. Positive correlation between LU and the foregoing housing demand factors would imply that LU is correlated with unobserved demand changes and subsequently is not a valid instrument for house prices. Yet previous attempts to assess the exogeneity of land unavailability suffer from an intrinsic sample selection issue: The only previously available land unavailability proxy from Saiz (2010) used MSAs, where MSA instantiation requires a population of 50,000 or more. MSAs thus do not provide complete coverage of the United States and are biased towards historical delineations of development. Using complete coverage of the United States and local geographic data, we find that LU is not positively correlated with amenities; the portion of people who are college educated or foreign born; or annual Bartik (1991) labor demand shocks. Moreover, in common 2SLS setups, when using the components of LU (land unavailability due to steep sloped terrain, water, or wetlands) separately as instruments, we find in overidentification tests that we cannot reject the null hypothesis that the over-identifying restrictions are valid. Altogether, these results support the use of LU as a candidate instrument for house prices.

Next, we employ our comprehensive LU dataset in the study of three fundamental housing market problems to emphasize various advantages of our highly disaggregated land unavailability measure. First, we replicate Guren et al. (2021) to show that our data has superior geographic coverage compared to Saiz elasticity and, in combination with machine learning methods, yields notably more precise 2SLS estimates when the first stage is a regression of house price changes on LU. Second, we construct a novel index for buildable land that allows for new empirical tests of theoretical housing market hypotheses that were not previously possible with other data. Finally, we demonstrate that our approach generalizes to different geographic aggregations by exploiting variation at both the county and CBSA levels in our study of the impact of house price growth on self-employment during COVID-19.

Our first application centers on re-estimating housing wealth elasticities, within the empirical setup of Guren et al. (2021), via the equation

$$\Delta y_{i,r,t} = \psi_i + \xi_{r,t} + \beta \Delta p_{i,r,t} + \Gamma X_{i,r,t} + \epsilon_{i,r,t} \tag{1}$$

$\Delta y_{i,r,t}$ is the log annual change in quarterly retail employment per capita (a consumption proxy in year-over-year first-difference form) for CBSA i in Census division r at time t. Similarly, $\Delta p_{i,r,t}$ is the log annual change in quarterly house prices for CBSA i. The coefficient of interest, β, measures the housing wealth elasticity. For reference, GMNS theoretically compute a housing wealth elasticity of 0.09, meaning a 10 percent increase in house prices leads to a 0.9 percent increase in consumption. ψ_i, $\xi_{r,t}$, and $X_{i,r,t}$ represent CBSA fixed effects, Census region $\times$ time fixed effects, and other controls, such as industry shares, respectively.[2]

Table 1 previews the estimation results from equation 1 and presents 2SLS estimates of housing wealth elasticities from 1978 - 2017. Column 1 replicates GMNS and uses Saiz elasticity as an instrument for house prices.[3] Notice first that the Saiz data contain just 270 CBSAs versus the 380 available in the broader GMNS dataset. Moreover, the Saiz elasticity IV has only moderate predictive power for house prices, with a first stage F-statistic of 14.37 and a first stage partial R^2 (e.g., the predictive power of just the instrument) of 0.02. The second stage housing wealth elasticity estimate is statistically insignificant with a relatively large standard error. The instrument in column 2 is LU calculated within a 50k km radius of each CBSA's principal city centroid, following the Saiz method. Note that in column 2 we retain the number of CBSAs available in the original Saiz dataset. Relative to the Saiz elasticity IV in column 1, which combines land unavailability and local housing market regulations, the IV in column 2 only employs land unavailability. The LU IV in column 2 also uses 2015 CBSA definitions to match the GMNS dataset, whereas Saiz elasticity uses 1999 delineations. The 2SLS estimate increases moving from column 1 to column 2, but its precision changes little. Column 3 expands the number of CBSAs to include all those in the

[2]See Guren et al. (2021) for a full list of controls.

[3]To use Saiz elasticity (a cross-sectional dataset) within a panel setup, GMNS multiple the Saiz elasticity proxy for each city by national level house prices.

contiguous U.S., relative to the 270 CBSAs in the original Saiz data, leading to a 40 percent increase in the available cross-sectional units. Yet increasing the size of the dataset (moving from column 2 to column 3) yields similar first stage predictive power and 2SLS estimates when the excluded instrument is land unavailability computed using the Saiz method.

Column 4 implements our preferred machine learning (ML) based, LU instrument. The LU-ML instrument combines land unavailability estimates across several levels of disaggregation and thus accounts for CBSA-level idiosyncrasies in the predictive relationship between land unavailability and house prices. The results using this LU-ML IV are noteworthy. To start, first stage predictive power rises substantially: The first stage partial R^2 increases by a factor of 5 compared to the Saiz elasticity IV (column 1) or a factor of 10 relative to LU within a 50k km circle of each CBSA's principal city centroid (Saiz method; columns 2 and 3). This jump in first stage fit leads to markedly more precise second stage estimates, with the 2SLS standard error falling from 0.54 in column 1 (Saiz elasticity) to 0.018 in column 4 (LU-ML). In fact, the precision of the 2SLS coefficient in column 4 matches the precision of GMNS's preferred estimates that use a more elaborate instrument. The coefficient in column 4 is also similar to the GMNS theoretical estimates. Together, these results highlight the benefits of the LU-ML instrument that uses machine learning techniques to combine land unavailability estimates computed at multiple levels of disaggregation.

Last, column 5 uses the components of LU (land unavailability due to steep sloped terrain, water, and wetlands) as instruments. As there is heterogeneity across cities, we employ cross-validation techniques to choose the optimal polygon shape for which to compute the components of land unavailability for each CBSA. The aim of this additional exercise is to examine the output from an overidentification test. The overidentification results in column 5 indicate that we cannot reject the null that coefficient estimates from the LU components are statistically different, supporting the use of the LU components as instruments for house prices.

In our second application, we use satellite imagery to construct a new dataset that precisely measures the amount of buildable land in 2001 within a geographic polygon. Buildable

land is the amount of land available for development after removing existing development, steep sloped terrain, water, wetlands, and parks. In a sense, buildable land is the complement to LU but also accounts for previous construction and parks. We then use this dataset to examine one of the largest puzzles in housing finance: Why did traditionally elastic housing markets, like Las Vegas, experience substantial 2000s house price growth even though these markets had ample room for housing construction? In particular, we empirically test the land supply-side speculation theory of Nathanson and Zwick (2018). This theory posits that homebuilders during the 2000s boom viewed traditionally elastic housing markets with intermediate amounts of available land (e.g., Las Vegas and Phoenix) as potentially inelastic in the long run. Homebuilders then proceeded to bid up the prices of land in these intermediate markets, and, as land is a crucial input for home construction, prices increased as well. While Nathanson and Zwick (2018) do provide anecdotal evidence in support of their theory, they were unable to perform formal statistical tests as no comprehensive buildable land dataset was previously available. In this paper, we undertake such tests and find that housing markets with intermediate amounts of buildable land experienced larger house price booms during the 2000s, relative to those with smaller or larger quantities of buildable land, congruent with the supply-side speculation theory.

Finally, our last application examines the relationship between self-employment and house price growth during the COVID-19 pandemic. This analysis, while preliminary, assesses the COVID-19 era external validity of Adelino et al. (2015). Adelino et al. (2015) study the impact of house price growth on self-employment during the 2000s housing boom using Saiz elasticity as an instrument for house prices. In our updated analysis, we use LU compiled from multiple levels of disaggregation via machine learning techniques as well as a markedly larger cross-sectional dataset that comprises over 2,000 counties or 800 CBSAs. As in our housing wealth elasticity application, LU has substantially stronger first stage predictive power than Saiz elasticity, highlighting the benefits of our LU measure. Indeed, the first stage F-statistic, a proxy for IV predictive power, when using LU as an instrument is three times as large as that for the Saiz elasticity IV used by Adelino et al. We also find that

the results of Adelino et al. extend to the COVID-19 pandemic (through 2021Q2), using data aggregated to either counties or CBSAs: Local house price appreciation increases the number of small firms (those with less than 5 employees) but has no affect on the number of larger firms.

2 Data

The United States Geographical Survey (USGS) provides the two main datasets that we use to measure slope, water, and wetlands land unavailability.[4] The first is the USGS National Elevation Dataset (NED) 3DEP 1 arc-second Digital Elevation Model (DEM). The 1 arc-second DEM data provide continuous coverage of the United States at approximately a resolution of 30 meters. The original Saiz dataset uses lower resolution 3 arc-second DEM data with a resolution of approximately 90 meters. The DEM data allow us to calculate slope files and hence the percentage of land unavailable due to a steep slope. Our second main dataset is the USGS 2011 Land Cover dataset. These data use LandSat imagery to classify land use in the U.S. The relevant categories for land unavailability are water (oceans, lakes, rivers, etc.) and wetlands. LandSat imagery is also used in the construction of our buildable land dataset. Finally, we incorporate other geospatial data such as shapefiles for various geographies from the U.S. Census Bureau and satellite imagery from Google Maps.

2.1 Other Data

Our data also include several key housing and economic variables. House prices are from Freddie Mac or Zillow. From the 2000 U.S. Census at the zip code level, we retain the percentage of people with a college education, the percentage of foreign born, and housing density. Data also span a zip code amenities index that aggregates information on access to restaurants and bars, retail shopping, public transit, and other amenities. From the BLS Quarterly Census of Employment and Wages (QCEW), we compute Bartik (1991) labor demand shocks and firm counts by employee size. The Missouri Data Bridge provides a geographic correspondence engine to crosswalk data across geographies. Last, data for our

[4] https://www.usgs.gov/the-national-map-data-delivery

housing wealth elasticity application are from Guren et al. (2021).

3 Land Unavailability Overview

The groundbreaking work of Saiz (2010) provides the foundation for this paper as it was the first to use satellite imagery and GIS methods to compute proxies of land unavailability. Saiz starts by using the USGS 90 meter DEM data to calculate the percentage of land unavailable due to a steep slope. Specifically, he notes that land with a slope above 15 percent faces architectural impediments to construction. The second dataset that Saiz employs is the 1992 Land Cover dataset. This latter dataset, combined with digital contour maps, measures the percentage of unavailable land due to oceans, lakes, rivers, etc. Combining these two sources, Saiz computes the share of land unavailable for construction (due steep sloped terrain, water, and wetlands) within a 50k kilometer (km) radius around the centroid of each MSA's first central city using 1999 MSA/NECMA definitions.

As an example of the Saiz approach, in figure 1 we plot Google satellite imagery for the Los Angeles-Long Beach MSA, using the 1999 delineations of Saiz (2010). Here, the blue outlined area represents the polygon boundary for the Los Angeles-Long Beach MSA. The orange polygons signify the central cities (Los Angeles, Long Beach, Pasadena, and Lancaster). The red dots are the centroids of each central city polygon, and the yellow circle has a 50k km radius around the first central city centroid (in this case, the Los Angeles central city). The 50k km circle around the first central city centroid is the area used by Saiz to calculate land unavailability. Clearly, the location of the first central city centroid determines the geography used in the Saiz calculation: The 50k km radius circle captures Los Angeles proper but does not cover the central city around the Lancaster and Palmdale areas, two cities with a combined 2000 population of over 230,000, or eastern Los Angeles around Pomona. In the greater LA area, these are the exact regions where much of the new construction occurs. Moreover, Anaheim and Irvine, two cities in south LA often included as a part of LA in the modern CBSA definitions used by many housing datasets, are also left out of Saiz circle. Last, the Saiz circle does not cover the disjointed polygons representing the Catalina islands.

Generally, the Saiz circles undercover MSAs that span large geographic areas but cover more land area for comparatively smaller polygons (noting again that every Saiz circle has a 50k km radius around the central city centroid). Larger MSAs are typically in the southwest, while the MSAs in the northeast are usually smaller. Thus, the coverage of the Saiz circles is correlated with geography. Cities are also heterogeneous along other dimensions, such as the shape of their polygons or their location relative to other cities. Therefore, as noted by GMNS, Saiz applies a blunt, homogeneous approach across heterogeneous cities. This homogeneity reduces the predictive power of land unavailability for house prices.

Therefore, in this paper, we construct new measurements of the percentage of undevelopable land in a geographic area, where the levels of spatial aggregation span cities, counties, commuting zones, zip codes, and various polygons related to these entities. Then, noting that any individual land unavailability proxy is likely measured with error and there are idiosyncrasies across cities, we employ machine learning techniques to combine land unavailability estimates computed at multiple levels of disaggregation.

First, we employ higher resolution satellite imagery from the USGS than that used in the original Saiz dataset (see section 2). Then for each census geographic delineation, we compute multiple land unavailability proxies at several levels of disaggregation, as the optimal polygon with which to calculate land unavailability likely varies across geographies.

An example of our approach for the Los Angeles-Long Beach-Anaheim CBSA, using the 2015 delineations employed by GMNS, is in figure 2. In panel A, as a baseline, we draw the Los Angeles CBSA polygon (blue lines) on top of Google satellite imagery for the greater LA and surrounding areas. Compared to the 1999 MSA definition used by Saiz (figure 1), we can see that the 2015 CBSA polygon includes the southern LA cities of Anaheim and Irvine, as noted above. Differences in geographic delineations over time thus present notable challenges for researchers. Indeed, porting land unavailability across delineations may bias the relationship between land unavailability and house prices. Hence, an immediate benefit of our dataset is that we compute LU for multiple geographic definitions ranging from the original 1999 Saiz MSAs/NECMAs through 2020 definitions. Matching land unavailability

to the delineations used in other datasets eliminates a vector of uncertainty for researchers studying housing markets.

Returning to figure 2, panels B to D show the polygons that our methodology exploits in the measurement of land unavailability. We consider various buffered polygons around the first principal city (panel B), buffers around the CBSA polygon (panel C), and circles around the principal city centroid (panel D). More specifically, in panel B, we calculate land unavailability within the first principal city polygon (by population size; orange polygon), in this case corresponding to the Los Angeles principal city. Then, we buffer this polygon by 5 percent (inner yellow polygon) and calculate land unavailability. From there, we sequentially increase the buffer size by 5 percentage points until the buffer reaches 25 percent (outer yellow polygon) of the original principal city polygon. The result is 6 separate land unavailability measurements (corresponding to buffers ranging from 0 percent to 25 percent) around the first principal city polygon. These proxies directly capture land unavailability at the population-weighted center of a CBSA.

We apply the same approach to the overall CBSA polygon (figure 2, panel C), yielding 5 additional land unavailability estimates, corresponding to the CBSA polygon with buffers ranging from 0 to 20 percent. The advantage of buffering the CBSA polygon is that it covers the whole CBSA and accounts for idiosyncrasies in the polygon's shape. Finally, in panel D, we expand Saiz's original approach by calculating land unavailability within multiple circles around the first principal city centroid, instead of just within a 50k km radius circle. In the case of Los Angeles in panel D, larger circles may be appropriate as they cover eastern LA, including San Bernardino and Riverside, and southern Los Angeles. Altogether in panel D, we calculate land unavailability within 9 distinct circles, with radii around the first principal city centroid ranging from 20k to 100k km.

Overall, figure 2 shows how we calculate land unavailability measurements for 20 different polygons associated with each CBSA. However, the best land unavailability predictors for house prices likely vary across CBSAs due to differences in polygon shapes, surrounding cities, or the direction of new construction.

Thus, the second step of our LU approach uses the various land unavailability proxies and machine learning (ML) techniques to predict house prices. Our ML framework spans housing datasets and only needs minor modifications depending on the empirical setup. Output from our approach is either (1) a panel dataset of LU-based predictors; or (2) a cross-sectional dataset with a single land unavailability estimate for each geographic unit, where the optimal geographic polygon for which to calculate land unavailability is chosen by cross-validation. We use the LU panel dataset in our housing wealth elasticity application (GMNS and equation 1). The cross-sectional LU data are useful when the econometric equation of interest is in long differenced form (e.g., Adelino et al. (2015) and equation 3 below).

Here, we describe our methodology within the GMNS empirical setup that uses a panel of Freddie Mac house prices, but our framework is similar for other housing datasets and empirical approaches. See appendix B for more details.

In the first stage of a 2SLS research design, GMNS use the Saiz elasticity IV to predict a city-time panel of Freddie Mac house prices (with 2015 CBSA delineations, those described for Los Angeles in figure 2). They multiply the Saiz elasticity cross-sectional data by national level house prices to create a panel. Their regressions control for CBSA fixed effects, among other variables. Our approach builds on this setup.

First, as in GMNS, we multiply each of our 20 land unavailability proxies by national house prices to create a city-time panel. We then residualize the Freddie Mac CBSA house price indices and each land unavailability proxy with respect to CBSA fixed effects, again following GMNS. To gauge predictive performance, we set up training and test sets. Since the Freddie Mac dataset is a panel beginning in 1978, the training datasets correspond to 10 year rolling windows inclusive of all CBSAs. The out-of-sample test sets are specific to each CBSA and comprise all observations not used in the training set. We then apply multiple machine learning algorithms to each training set and calculate the corresponding out-of-sample root mean-squared error (RMSE) for each CBSA using the test sets, akin to out-of-sample cross-validation (CV). The final LU instruments correspond to full sample

house price predictions for each CBSA, using the algorithm that produces the lowest average out-of-sample RMSE for that CBSA.

For example, using the residualized data and OLS as the candidate algorithm, we first regress a panel of Freddie Mac house prices on a specific Land Unavailability proxy for data from 1978-1988. The given land unavailability proxy might correspond to the first principal city polygon (e.g., the orange polygon in figure 2, panel B). The coefficients from the training regression (slope and intercept) are then combined with the test data to predict house prices for each CBSA from 1988 and compute out-of-sample RMSEs by CBSA. We iteratively repeat this process for all subsequent 10 year training windows, all land unavailability proxies, and all considered algorithms. We include all land unavailability proxies for more sophisticated ML algorithms, such as XGBoost, random forest, or lasso, rather than building separate models that use our land unavailability measurements individually. The end result for just the Los Angeles CBSA, for example, is a matrix of out-of-sample RMSEs with test set windows along the rows and the various algorithms in the columns. The chosen algorithm for the Los Angeles CBSA has the lowest average RMSE across all test sets (e.g., the mean by column). With the preferred algorithm in hand, we then generate the LU house price instrument for just Los Angeles by running the LU proxies through the chosen algorithm and retrieving the fitted values, representing the house price predictions. This entire process is repeated for every CBSA in the dataset to build a CBSA-time panel of instruments for Freddie Mac house prices.

4 Correlations Between LU and Housing Demand Proxies

The use of land unavailability as an instrument relies on its exogeneity relative to other proxies for housing demand. Specifically, if higher land unavailability is exogenous and predicts higher house price growth, then land unavailability should not be positively correlated with factors of housing demand. In the literature, there has been debate on this issue. Mian and Sufi (2011, 2014) claim that land unavailability is exogenous, while Davidoff (2016) contends that land unavailability is positively correlated with housing demand factors.[5] Our study

[5]As noted above, GMNS use a panel data structure to circumvent Davidoff's criticisms.

differs from previous attempts to assess the exogeneity of land unavailability as we use a more highly disaggregated dataset with complete coverage of the contiguous United States.

In contrast, previous studies that aim to assess the exogeneity of land unavailability employ data at the MSA level. Yet MSAs only cover a fraction of U.S. land area. This limited coverage biases any correlations between land unavailability and housing demand factors towards regions with higher levels of historical development (e.g., the northeastern U.S.). Indeed, for a city to be classified as an MSA using the 1999 Saiz delineations, it must have a population of at least 50,000. As land unavailability increases the cost of home building and construction, MSAs are less likely to be located in areas with high land unavailability, all else equal. To see this, consider figure 3 that plots Google satellite imagery for the U.S. and coverage for Saiz elasticity. The red circles have a 50k km radius around each MSA first central city centroid as in the Saiz dataset. The figure clearly shows a strong negative correlation between the instances of MSAs and land unavailability due to rugged terrain, for example. This pattern is visible in the Rocky Mountain region, where five MSAs sit at the base of the Rockies in Colorado. Yet even in the populated pacific states (California, Washington, and Oregon), there is a negative relationship between MSA instantiation and terrain slope. Indeed, the northern ascent of California MSAs is limited by the Mendicino and Shasta National Forests, while Seattle lies between Olympic National Park and Wenatchee National Forest. Thus, judging the exogeneity of land unavailability using only MSAs will lead to biased results.

We hence examine the correlations between land unavailability and demand proxies with near complete national coverage. The proxies of demand that we consider include a zip code amenities index compiled from a large internet aggregator of such information, the college share in 2000, the foreign share in 2000, and housing density in 2000. We consider these variables, as well as land unavailability computed within a 5 percent buffered polygon of each geographic unit, at the zip code and three-digit zip code levels. The output of regressions of the various housing demand factors on land unavailability (LU) is in table 2. Panel A shows the results using zip code data, while panel B measures all variables at the

three-digit zip code level. Robust standard errors are clustered at the two-digit zip code level. As suggested by Davidoff (2016), in order for land unavailability to fail the exclusion restriction, it must be positively correlated with other housing demand indicators.[6] Instead, our results show that LU is not positively correlated with housing demand proxies. At the zip code level, LU is negatively correlated with the amenities index, foreign share, and housing density but uncorrelated with college share. These results are not surprising as increased land unavailability makes the construction of housing and amenities more expensive. At the three-digit zip code level, LU is negatively correlated with amenities but uncorrelated with the other housing demand variables.

Another key determinant of housing demand is changes to labor demand within a city. We follow the labor literature and Davidoff (2016) and employ Bartik (1991) Labor demand shocks at the county level. We assess the correlation between LU and labor demand shocks through the following specification, estimated separately for each year from 2001 to 2019:

$$Bartik_i = \alpha + \beta \cdot LU_i + \epsilon_i \tag{2}$$

$Bartik_i$ represents the annual BLS QCEW Bartik shock for county i and LU is land unavailability for county i computed using a 5 percent buffer around each county polygon. The results are in figure 4. Error bars correspond to ± 2 robust standard errors clustered at the state level. Clearly, the Bartik labor demand shocks are largely uncorrelated with LU since 2000.

Overall, the results in this section indicate that LU is uncorrelated with key housing demand factors. Thus, LU likely does not violate the IV exclusion restriction as a candidate instrument for house prices.

5 LU and Housing Wealth Elasticities

In our first application of the LU data, we estimate housing wealth elasticities within the framework of GMNS (equation 1). Our aim is to compare 2SLS estimates where the excluded instrument is Saiz elasticity, LU, or GMNS's preferred sensitivity instrument. To build their

[6] Indeed, Davidoff notes that as households' demand for amenities, the economics of agglomeration, etc. has been increasing over time, they cannot be accounted for using standard region fixed effects.

sensitivity instrument, GMNS first regress CBSA house price growth (year-over-year log first-differences) on the house price growth of the corresponding census regions, with city-specific coefficients. The sensitivity instrument for each CBSA is then the estimated city-specific coefficient from this regression multiplied by that CBSA's census region house price growth.

We replicate GMNS's primary results in table 3. GMNS estimate equation 1 for three separate periods 1978-2017 (panel A), 1990-2017 (panel B), and 2000-2017 (panel C). OLS results are in column 1, while columns 2 and 3 show results that use the sensitivity instrument and the Saiz elasticity IV, respectively.

The OLS results for the full sample (1978-2017) correspond to an elasticity estimate of 0.083, meaning a 10 percent increase in house prices is associated with a 0.83 percent gain in retail employment. Note that theoretically that GMNS find an elasticity estimate of 0.09. Column 2 presents GMNS's preferred estimates that use the sensitivity instrument. The coefficient in panel A for the 1978-2017 period in column 2 is 0.058 with a standard error of 0.017. Compare this to the results in column 3, panel A that use the Saiz elasticity IV.[7] In column 3, the standard error is relatively large at 0.048. The large standard error stems from the low predictive power of the Saiz IV in the first stage. Indeed, the first stage F-statistic in column 3 that uses the Saiz IV is just 19.67 versus 249.08 for the sensitivity instrument (column 2). Panels B and C likewise report estimates for the 1990-2017 and 2000-2017 time periods.

The bottom panel of table 3 provides various notes on each regression specification. In particular, the Saiz IV has a limited sample size. The number of CBSAs used in column 3 when Saiz elasticity is the IV is just 270, compared to 380 for the overall dataset. Also, GMNS use census region $\times$ year-quarter fixed effects in columns 1 and 2 but only year-quarter fixed effects in column 3. Employing census region $\times$ year-quarter fixed effects when Saiz elasticity is the excluded instrument increases the second stage standard error further (see table 4, column 2).

Table 4 extends the GMNS empirical framework to specifications where LU serves as an

[7] To use Saiz elasticity (a cross-sectional dataset) within a panel setup, GMNS multiple the Saiz elasticity proxy for each city by national level house prices.

instrument for house prices. To ease comparison to the GMNS results, column 1 replicates the GMNS findings that use the Saiz elasticity IV (e.g., column 3 of table 3). Column 2 of table 4 again employs the Saiz elasticity IV but uses census region $\times$ time fixed effects instead of just time fixed effects (e.g., to match the GMNS OLS and sensitivity IV estimates, columns 1 and 2 of table 3). Including the census region $\times$ time fixed effects reduces the coefficient estimates slightly, while the standard errors increase.

In column 3, the instrument is LU within a 50k km circle around each CBSA's principal city centroid, following the Saiz method. The first difference between Saiz elasticity and LU within a 50k km circle is that LU uses the 2015 CBSA definitions, in line with GMNS's Freddie Mac house price dataset, whereas Saiz uses 1999 MSA/NECMA definitions. The second difference between the two instruments is that Saiz elasticity also incorporates local land use regulations from the Wharton Residential Land Use Regulatory index. As house price growth increases local regulation through political economy mechanisms (Davidoff, 2016), this portion of the Saiz elasticity proxy is likely not exogenous. Also, local land use regulations predict house price increases (Davidoff, 2016). Thus, the first stage partial R^2 measures are higher in columns 1 and 2, which use Saiz elasticity, versus columns 3 and 4, which use LU within a 50k km circle of each CBSA's principal city centroid. In column 3, we also retain the original Saiz dataset (just 270 CBSAs). The coefficient estimates increase in all estimation periods (panels A - C) compared to column 2, but the standard errors also increase for the 1990-2017 and the 2000-2017 time periods.

Column 4 uses all CBSAs in the GMNS dataset (except those in AK and HI), and the excluded instrument remains LU calculated within a 50k km circle around each principal city centroid. The 2SLS elasticity estimates fall compared to column 3 and, in the case of the 1990-2017 and 2000-2017 periods, they do so considerably. This result indicates that housing wealth elasticity estimates may differ in recent data across the large CBSAs available in the Saiz dataset versus smaller ones available in modern housing datasets. We leave the estimation of differences in housing wealth elasticities between large and small CBSAs as an avenue for future research.

Column 5 shows our preferred elasticity estimates. Here, our LU instrument uses machine learning (ML) algorithms and out-of-sample cross-validation (CV) techniques to combine multiple LU proxies at various levels of disaggregation, separately by city (see section 3 and appendix B). In doing so, our approach accounts for the geographic heterogeneity across cities, such as differences in CBSA size or the importance of surrounding cities, while also recognizing that any individual LU proxy is likely measured with error. Our LU-ML measure thus yields a highly predictive, plausibly exogenous panel of house price predictors that addresses the GMNS criticisms of Saiz elasticity surrounding its homogeneous approach and low predictive power.

When using the LU-ML IV in column 5, first stage predictive power jumps markedly. Indeed, the first stage partial R^2 (the R^2 change from just adding the IV) in panel A increases to 0.10, a fivefold increase relative to the Saiz elasticity estimates in column 2. With this considerable rise in first stage predictive power, the second stage standard error falls precipitously. In fact, the precision of the 2SLS elasticity coefficient nears the precision of GMNS's preferred estimates, even though GMNS use a more elaborate IV (e.g., the sensitivity IV estimates in column 2 of table 3). Finally, note that GMNS produce a theoretical estimate of the housing wealth elasticity of 0.09. Thus, the coefficient estimate in column 5, panel A for the 1978-2017 sample is near the GMNS's theoretical results.

In column 6 of table 4, we employ both the LU-ML and the GMNS sensitivity instruments. This allows us to construct a standard overidentification test and test whether the second stage housing wealth elasticity estimates using either the GMNS sensitivity instrument or the LU-ML instrument are statistically different. The overidentification p-value in column 6, panel A is 0.01, indicating that for the 1978-2017 sample that we reject the null hypothesis that the LU-ML IV (e.g., column 5) and the sensitivity IV (table 3, panel A, column 2) produce the same second stage coefficients. This is not surprising as the housing wealth elasticity estimate is 0.081 when using the LU-ML IV, close to GMNS's theoretical estimates of 0.09, versus just 0.058 for the sensitivity instrument. Yet the overidentification p-values in panels B and C of column 6 indicate that we cannot reject the null hypothesis that the

LU-ML and the GMNS sensitivity IVs produce the same second stage estimates for the 1990-2017 and the 2000-2017 sample periods.

Column 7 uses an alternative ML approach. Instead of using ML techniques to combine multiple LU proxies into a panel (column 5), in column 7 we use out-of-sample cross-validation (CV) to choose the optimal polygon shape for which to calculate land unavailability for each CBSA. The result is a cross-sectional, CBSA dataset consisting of a land unavailability estimate for each CBSA. A full description of the algorithm is in appendix B. To use this cross-sectional dataset within a panel framework, we follow GMNS and multiply LU for each CBSA by national level house prices in year-over-year log first difference form. Relative to Saiz's original methodology that uses just a 50k km circle around each CBSA's principal city centroid, column 7 shows the benefits of allowing the polygon shape for which we calculate LU to vary across cities: The first stage partial R^2 in panel A increases sixfold when using this LU-CV instrument, from 0.01 for the Saiz method (columns 3 and 4) to 0.06 (column 7).

In column 8, we use the components of LU (land unavailability due to steep sloped terrain, water, and wetlands) as the excluded instruments. The optimal polygon shape for which we calculate land unavailability for each CBSA is chosen using the algorithm described for column 7. Using the LU components slightly increases first stage predictive power compared to total LU (column 7) and thus leads to a reduction in the second stage standard errors. Furthermore, for all reported sample periods, overidentification tests indicate that we cannot reject the null hypothesis that the LU components produce the same second stage estimates, supporting the exogeneity of LU and its use as an IV for house prices.

Finally, column 9 uses total LU (from column 7) and GMNS sensitivity as instruments. Similar to column 6, the overidentification p-values indicate that we reject the null hypothesis that the total LU and sensitivity instruments produce the same second stage results for the full, 1978-2017 sample (panel A), but not for the 1990-2017 (panel B) and the 2000-2017 (panel C) samples.

6 Buildable Land and Supply-Side Speculation

Nathanson and Zwick (2018) develop a theoretical model that documents how disagreement and supply-side speculation in housing markets can produce house price booms in traditionally supply elastic areas. Specifically, the model posits that homebuilders may view housing markets with intermediate amounts of land available for development (buildable land) as supply elastic in the short run but inelastic in the long run. When these homebuilders are optimistic about future prices (e.g., during a national housing boom), they acquire and subsequently bid up the prices of available land. Since land is a key factor in housing production, this raises house prices in markets with intermediate amounts of buildable land even in the face of large-scale construction. As a result, house prices boom in traditionally supply elastic housing markets. The supply-side speculation theory thus aims to explain the large and previously puzzling 2000s house price growth in areas like Phoenix, Las Vegas, Florida, and inland California.

Nathanson and Zwick (2018) provide several pieces of evidence in support of their theory. For example, they cite a Polte homes investor presentation that stated that the traditionally elastic markets of West Palm Beach, Orlando, Tampa, Ft. Myers, Sarasota, Las Vegas, and Chicago were surprisingly constrained. A more formal test of the supply-side speculation theory would require precise data on the amount of buildable land within housing markets. To our knowledge, no such dataset previously existed.

Therefore, this section exploits detailed satellite land cover and slope image files to construct a new dataset that precisely measures the amount of buildable land across the contiguous United States.

The basis of our computation is the 2001 USGS LandSat Land Cover Dataset. The LandSat Land Cover data classify land use in the United States at a spatial resolution of 30 meters. Figure 5 plots the LandSat Land Cover data for Florida. In the satellite image, red pixels correspond to developed land, where darker red pixels represent more dense development. Similarly, blue areas represent water and wetlands. The most developed area is downtown Miami (dark red in southeast Florida), and the map clearly shows how

water and wetlands restrict housing expansion in that market. Oppositely, other coastal and central Florida areas are comparatively at the intermediate stages of development. They have lower density and surrounding areas that appear to be available for construction.

We compute the land area available for development within each housing market by first removing developed land (e.g., red pixels on the Florida map) as well as water and wetlands (blue pixels). We also remove steep sloped terrain measured using USGS 1 arc-second DEM slope files (using no buffer for polygons in the shapefiles) and exclude regions designated as parks using a shapefile from data.gov. We then calculate the land area of the remaining, buildable land. In a sense, buildable land is the complement to land unavailability but additionally classifies start of period developed land (2001) and parks as unavailable as well.

We compute buildable land within three-digit U.S. zip codes. U.S. zip codes were developed in the 1960s to have similar populations across geographical units. Hence, they better reflect pre-2000s housing boom U.S. populations and geographies, especially in the Western U.S. In contrast, counties or MSAs vary drastically in size and stem from geographic definitions dating back to the 1800s.[8]

To test the relationship between buildable land and 2002 - 2006 house price growth, we group three-digit U.S. zip codes into 2001 buildable land deciles. Summary statistics are in table 5. Column 1 shows the buildable land decile, and column 2 displays the average amount of buildable land for three-digit zip codes in that buildable land decile (thousands of square kilometers). As expected in column 2, buildable land is monotonically increasing over buildable land deciles. Column 3 shows the mean percentage of land that is buildable (relative to all available land) within each buildable land decile. Notice that there is minimal available buildable land in deciles 1 and 2. Three-digit zip codes in these deciles are the "inelastic" housing markets characterized by Nathanson and Zwick that likely have both high land unavailability and regulatory supply restrictions.[9] For other buildable land deciles, the percentage of buildable land is monotonically increasing. A potential concern when using

[8]For example, the land area of the Riverside-San Bernardino MSA is 260 percent larger than the land area of the entire state of Massachusetts.

[9]See also the references in Nathanson and Zwick (2018).

buildable land defined within three-digit zip codes, which can vary in size, is that buildable land may simply be a function of available land. We partially address this concern in column 4 and show the correlation between available and buildable land by buildable land decile. The correlations are wide ranging and only in buildable land decile 10 is the correlation with available land over 0.5. We return to this issue below.

Figure 6 maps three-digit U.S. zip codes by buildable land decile. Red areas signify buildable land decile 1 (least amount of buildable land), blue areas represent buildable land decile 5 (intermediate amount of buildable land), and yellow areas map buildable land decile 10 (largest amount of buildable land). Buildable land decile 1 indeed corresponds to housing markets that would traditionally be considered inelastic due to density, land unavailability, and regulatory constraints. These housing markets include New York City, Boston, Miami, downtown Tampa, New Orleans, downtown Chicago, downtown Milwaukee, coastal Los Angeles, and areas adjacent the San Francisco Bay. Three-digit zip codes in buildable land decile 5 (intermediate amounts of buildable land, blue) consist of suburban areas in inland southern California, central California, and northern California. Buildable land decile 5 also includes Las Vegas, Phoenix, Colorado Springs, suburban regions in central and coastal Florida, suburban Chicago, and several suburban housing markets in the northeast. Finally, yellow areas showing buildable land decile 10 are mainly rural areas in the Midwest and Texas.

Note that Nathanson and Zwick's supply-side speculation theory aims to explain housing markets with *intermediate* land supply. Thus, they concede that supply inelastic markets should also experience sizable house price growth during a boom (e.g., Saiz (2010)) and that the house price growth in inelastic markets is not the focus of their theory. Thus, the null hypothesis of interest is that house price growth in traditionally supply elastic areas with intermediate amounts of buildable land is equal to house price growth in areas with relatively smaller or relatively larger amounts of buildable land. A rejection of this null supports the supply-side speculation theory and would yield a hump-shaped, non-monotonic relationship between buildable land deciles and house price growth.

We evaluate the supply-side speculation theory in table 6 by regressing 2002 - 2006 three-digit zip code house price growth on 2001 buildable land decile indicator variables. Robust standard errors are clustered at the state level. Column 1 shows the mean house price growth within each buildable land decile. Not surprisingly, house price growth is largest in areas with the least amount of buildable land (buildable land decile 1 corresponding to inelastic markets), at 58.6 percent. Yet the second highest mean house price growth is in buildable land decile 5 at 44.1 percent, followed closely by buildable land decile 4 at 42.9 percent. House price growth in buildable deciles 2 and 3 is substantially smaller at 35 and 27 percent, respectively.[10] Similarly, house price growth is markedly lower for buildable land deciles 6 through 10. Note also that the R^2 of the regression is 25 percent, and thus 2001 buildable land deciles explain a large portion of the cross-sectional variation in house price growth during the 2000s. Together, this evidence suggests that inelastic areas and housing markets with intermediate amounts of buildable land experienced the largest house price growth during the 2000s boom.

Columns 2 and 3 statistically test the supply-side speculation theory. Here we exclude the indicator for buildable land decile 5 but retain the intercept. Thus, the intercept is house price growth for buildable land decile 5, and the regression coefficients are the difference in mean house price growth relative to decile 5. The coefficients on the indicators for buildable deciles 2, 3, and 6 - 10 are all negative and statistically significant in column 2. Hence, three-digit zip codes in buildable land deciles 2, 3, and 6 - 10 experienced noticeably lower house price growth than three-digit zip codes with intermediate amounts of buildable land. Similarly, column 3 shows that controlling for Bartik labor demand shocks does not affect our results.[11] Together, these regressions document that three-digit zip codes with intermediate amounts of buildable land experienced statistically larger house price growth from 2002 - 2006, congruent with the supply-side speculation theory.

As noted above, a potential concern with the construction of buildable land within three-

[10]Buildable land decile 2 also likely contains inelastic housing markets, perhaps accounting for its slightly higher house price growth relative to decile 3.

[11]The Bartik is demeaned relative to the entire sample so the intercept can be interpreted as the mean house price growth in decile 5 in a three-digit zip code with an average Bartik shock.

digit zip codes is that buildable land may be a function of available land. Hence, the amount of available land within a three-digit zip code may be driving our results. We address this concern with a falsification test. Specifically, we retain all three-digit zip codes outside of buildable land deciles 1 (inelastic areas) and 5 (intermediate buildable land areas). Of these remaining regions, we then collect the three-digit zip codes whose available land is within the range of available land for the original buildable land decile 5. This yields 294 (out of 607) three-digit zip code regions whose available land is within the range of buildable decile 5. The mean house price growth for these regions is 25.1 percent. All other three-digit zip codes outside of our original buildable land deciles 1 and 5 have a mean house price growth of 31.9 percent. The difference of -6.8 percentage points is statistically significant at the 1 percent level (robust t-stat $= -2.6$). Therefore, other three-digit zip codes whose available land is within the range of the available land for regions in buildable land decile 5 actually have *lower* house price growth. The results from this falsification test thus suggest that buildable land, and not available land, drive the above relationship between buildable land and house price growth.

7 LU and Self-Employment During the COVID-19 Pandemic

In a final application, we examine the relationship between house price growth and self-employment during COVID-19. We build on Adelino et al. (2015) who conduct a similar analysis during the 2000s housing boom. Our aim is to provide preliminary evidence as to the external validity of Adelino et al. (2015) during the pandemic and demonstrate the utility of our LU data for different geographic aggregations. The COVID-19 crisis was unique as substantial economic activity shut down, but house prices increased markedly in many markets. Thus, house price growth may have affected small firm counts via mortgage refinance or equity extraction, as lower debt-service payments or available cash may allow entrepreneurs to start new ventures or support existing businesses. Yet naive correlations between house price changes and self-employment or entrepreneurship are likely obscured by reverse causality. Thus, we employ LU as an instrument for house prices.

More specifically, we estimate the following regression separately by firm size bins (num-

ber of employees) using both counties and CBSAs as the geographic unit of aggregation:

$$\Delta^{2019\text{–}21} \ln NumFirms_i = \alpha + \beta \Delta^{2019\text{–}21} \ln HP_i + \varepsilon_i \tag{3}$$

$\Delta^{2019\text{–}21} \ln NumFirms_i$ is the log difference in the number of firms in county or CBSA i between 2019Q2 and 2021Q2. Firm counts are from the BLS QCEW.[12] Similarly, $\Delta^{2019\text{–}21} \ln HP_i$ is the log difference in the Zillow house price for county or CBSA i between 2019 and 2021. ε_i is an error term. As noted above, we estimate equation 3 separately for counties and CBSAs by firm size bins, measured by the number of employees per firm.

Figure 7 reports estimates from equation 3 for data aggregated to either counties (panel A) or CBSAs (panel B). Within each panel, we report both OLS estimates (red) as well as those that use LU as an instrument (blue). The horizontal axis lists the firm size bins, while the vertical axis reports the coefficient of interest, $\hat{\beta}$, from equation 3. Error bands correspond to ± 2.5 robust standard errors clustered at the commuting zone level, in line with Adelino et al. (2015).

To create the LU instrument, we use out-of-sample CV techniques to determine the optimal polygon shape for which to compute land unavailability from several candidates at various levels of disaggregation. This approach recognizes that different polygon shapes may result in more accurate predictions across heterogeneous geographies. The result is a highly predictive instrument for house prices that accounts for idiosyncrasies across local housing markets. A full description of our algorithm, applied separately to both counties and CBSAs, is in appendix B.

Panel A in figure 7 employs data from 2,097 counties, a cross-sectional dataset nearly three times as large as that used by Adelino et al. (2015). This notable increase in the number of cross-sectional observations arises due to the comprehensive coverage of our LU dataset compared to the limited availability of the Saiz elasticity proxy used by Adelino et al. Panel B aggregates counties to 802 CBSAs; substantially more cities than are usually included in housing market studies.

[12] The BLS QCEW reports the number of firms by industry (two-digit NAICS) and size at the state level. To generate the number of firms at the county or CBSA level, we use apply state-level industry by size shares to the county and CBSA data, respectively.

The OLS estimates (red) in both panels show that higher house price growth is correlated with increased firm counts across all firm size bins. Yet, as noted above, these estimates are likely plagued by reverse causality: Brightening local economic growth elevates house prices, while rising house prices improve local economic conditions.

Thus, in figure 7 we also employ LU as an IV, with the corresponding 2SLS estimates in blue. For counties (panel A), the first stage F-statistic is 64.08, over 3 times as large as the first stage F-statistic in Adelino et al. (2015) that uses Saiz elasticity. For CBSAs (panel B), the first stage F-statistic increases further to 89.24. These considerable first stage F-statistics highlight the strong predictive power of our LU IV and the benefits of employing land unavailability estimates computed from multiple levels of disaggregation.

Overall, our 2SLS results are similar across both counties and CBSAs (panels A and B of figure 7). These findings are also congruent with Adelino et al. (2015), supporting the external validity of their results for the COVID-19 period. Indeed, like Adelino et al. (2015), we find that higher house price growth increases the number of small firms (those with less than 5 employees) but has little impact on larger firms. As previously indicated, the channel underpinning these results likely relates to households refinancing into lower debt-service payments or using loans backed by housing collateral to start new businesses or buttress existing firms. Since these results are a preliminary description of the impacts of pandemic era house price growth, we leave untangling of the relevant channels for future research.

8 Conclusion

This paper combines high-resolution satellite imagery with machine learning techniques to provide new estimates of the geographic determinants of housing supply. Our land unavailability (LU) measure is a markedly more accurate house price predictor than the popular elasticity proxy of Saiz (2010). We also show that land unavailability is not positively correlated with local amenities, the share of college or foreign-born, housing density, or Bartik labor demand shocks, supporting the use of LU as an instrument for house prices.

In several applications, we highlight the utility of our land unavailability data. First, in the canonical problem of estimating housing wealth elasticities with respect to consumption,

we show that using LU as an instrument increases first stage predictive power by a factor of 5 relative to Saiz elasticity. As such, the precision of the second stage estimates rises precipitously, reducing uncertainty in the causal relationship between housing wealth and consumption. We next build a new dataset that measures the available buildable land within a geographic polygon. Buildable land is the amount of available land less previous development, parks, and the components of land unavailability (steeped sloped terrain, water, and wetlands). Using buildable land, we perform new statistical tests that provide support for the supply-side speculation theory (Nathanson and Zwick, 2018). The supply-side speculation theory contends that 2000s land and house prices rose in traditionally elastic markets as homebuilders and other agents were concerned that these markets would become inelastic in the long run. Last, we provide new evidence on the causal impact of house price growth on self-employment during the COVID-19 pandemic. In a 2SLS research design, first stage results again show that LU is a substantially more accurate house price predictor than Saiz Elasticity. Second stage estimates provide external validity for the findings of Adelino et al. (2015), as rising house prices increase small firm counts but have little impact on the number of larger firms.

As real estate continues to play a leading role in household and broader economic activity, researchers will likely find our land unavailability and buildable land datasets useful as house price instruments, predictors, control variables, or geographic proxies for supply constraints across regions.

References

M. Adelino, A. Schoar, and F. Severino. House prices, collateral, and self-employment. *Journal of Financial Economics*, 117(2):288–306, 2015.

A. Aladangady. Housing wealth and consumption: Evidence from geographically-linked microdata. *American Economic Review*, 107(11):3415–46, November 2017.

T. J. Bartik. *Who Benefits from State and Local Economic Development Policies?* Books from Upjohn Press. W.E. Upjohn Institute for Employment Research, November 1991.

N. Baum-Snow and L. Han. The microgeography of housing supply. *Working Paper*, 2020.

S. Büchler, M. v. Ehrlich, and O. Schöni. The amplifying effect of capitalization rates on housing supply. *Journal of urban economics*, 126:103370, 2021.

T. Chaney, D. Sraer, and D. Thesmar. The collateral channel: How real estate shocks affect corporate investment. *The American Economic Review*, 102(6):2381–2409, 2012.

J. N. Conklin, W. S. Frame, K. Gerardi, and H. Liu. Villains or scapegoats? the role of subprime borrowers in driving the us housing boom. *Journal of Financial Intermediation*, page 100906, 2021.

T. Davidoff. Supply constraints are not valid instrumental variables for home prices because they are correlated with many demand factors. *Critical Finance Review*, 5(2):177–206, 2016.

S. Gabriel, M. Iacoviello, and C. Lutz. A crisis of missed opportunities? foreclosure costs and mortgage modification during the great recession. *The Review of Financial Studies*, 34(2):864–906, 2021.

W. Gamber, J. Graham, and A. Yadav. Stuck at home: Housing demand during the covid-19 pandemic. *Working Paper*, 2021.

A. Gupta, V. Mittal, J. Peeters, and S. Van Nieuwerburgh. Flattening the curve: Pandemic-induced revaluation of urban real estate. *Journal of Financial Economics*, 2021. ISSN 0304-405X.

A. M. Guren, A. McKay, E. Nakamura, and J. Steinsson. Housing wealth effects: The long view. *The Review of Economic Studies*, 88(2):669–707, 2021.

A. Mian and A. Sufi. The consequences of mortgage credit expansion: Evidence from the us mortgage default crisis. *The Quarterly Journal of Economics*, 124(4):1449–1496, 2009.

A. Mian and A. Sufi. House prices, home equity–based borrowing, and the US household leverage crisis. *The American Economic Review*, 101(5):2132–2156, 2011.

A. Mian and A. Sufi. What explains the 2007–2009 drop in employment? *Econometrica*, 82 (6):2197–2223, 2014.

A. Mian, K. Rao, and A. Sufi. Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*, 128(4):1687–1726, 2013.

C. G. Nathanson and E. Zwick. Arrested development: Theory and evidence of supply-side speculation in the housing market. *The Journal of Finance*, 73(6):2587–2633, 2018.

A. Saiz. The geographic determinants of housing supply. *The Quarterly Journal of Economics*, 125(3):1253–1296, 2010.

Y. Tan, Z. Wang, and Q. Zhang. Land-use regulation and the intensive margin of housing supply. *Journal of Urban Economics*, 115:103199, 2020.

A Tables and Figures

Table 1: 2SLS Housing Wealth Elasticity Estimates – 1978 - 2017

	Dependent variable:				
	YoY Log Diff in Retail Emp Per Capita				
	(1)	(2)	(3)	(4)	(5)
YoY Log Diff in HP Growth	0.082	0.143***	0.136**	0.081***	0.099***
	(0.054)	(0.053)	(0.057)	(0.018)	(0.022)
First Stage F-Stat	14.37	20.52	19.64	168.62	44.62
First Stage Partial R^2	0.02	0.01	0.01	0.10	0.07
OverID p-value					0.17
Instrument(s)	Saiz Elasticity	LU 50km Circles	LU 50km Circles	LU-ML IV	Slope,Water Wetlands
Num. CBSAs	270	270	376	376	376

Notes: Column 1 replicates Guren et al. (2021), but uses time × census region fixed effects, instead of the time fixed effects used in Guren et al. Column 2 uses the unavailable land in a 50km circle around each CBSA's central city centroid, following the method of Saiz (2010), and the number of CBSAs used by Guren et al. that are available for the Saiz data. Column 3 expands the number of CBSAs to match the full Guren et al. dataset (less AK and HI). The instrument in column 4 uses machine learning and cross-validation techniques to combine Land Unavailability estimates measured at multiple levels of disaggregation to yield a panel prediction of house prices that are then used as an instrument. The instruments in Column 5 are the components of LU (land unavailability due to steep sloped terrain, water, and wetlands), where the polygon shape is chosen specifically for each CBSA by a cross-validation algorithm. Controls include CBSA fixed effects, census region × time fixed effects, industry shares × time, and the prediction controls in Guren et al. Robust standard errors are clustered by CBSA and time. One, two, or three asterisks represent statistical significance at the 10, 5, and 1 percent levels, respectively.

Table 2: Zip and Zip3 LU Correlations with Housing Demand Proxies

	Dependent variable:			
	Amenities Index	College Share in 2000	Foreign Share in 2000	Housing Density in 2000
	(1)	(2)	(3)	(4)
Panel A: Zip Code				
LU	-0.012^{***}	-0.002	-0.043^{***}	-0.102^{***}
	(0.002)	(0.021)	(0.011)	(0.027)
Constant	0.287^{***}	18.244^{***}	5.912^{***}	7.646^{***}
	(0.084)	(0.700)	(0.742)	(1.629)
Observations	13,044	30,606	30,614	30,645
R^2	0.061	0.00001	0.015	0.011
Panel B: Zip3				
LU	-0.005^{***}	-0.020	-0.002	-0.060^{*}
	(0.002)	(0.029)	(0.024)	(0.032)
Constant	0.184^{**}	22.446^{***}	6.678^{***}	11.413^{***}
	(0.087)	(0.906)	(0.911)	(2.894)
Observations	727	832	840	867
R^2	0.017	0.002	0.00004	0.001

Notes: Zip code and three-digit zip code (Zip3) regressions of housing demand proxies on Land Uanvailability. Robust standard errors are clustered at the two-digit zip code level. One, two, or three asterisks represent statistical significance at the 10, 5, and 1 percent levels, respectively.

Table 3: Replication of Guren et al. (2021), Table 1

	Dependent variable:		
	YoY Log Diff in Retail Emp Per Capita		
	(1)	(2)	(3)
Panel A: 1978 - 2017			
YoY Log Diff in HP Growth	0.083***	0.058***	0.084*
	(0.007)	(0.017)	(0.048)
First Stage F-Stat		249.08	19.67
First Stage Partial R^2		0.16	0.03
Observations	59,999	59,999	42,710
Panel B: 1990 - 2017			
YoY Log Diff in HP Growth	0.081***	0.072***	0.141***
	(0.008)	(0.015)	(0.037)
First Stage F-Stat		440.81	20.41
First Stage Partial R^2		0.27	0.04
Observations	41,985	41,985	29,867
Panel C: 2000 - 2017			
YoY Log Diff in HP Growth	0.068***	0.055***	0.134***
	(0.008)	(0.014)	(0.035)
First Stage F-Stat		351.11	21.12
First Stage Partial R^2		0.31	0.05
Observations	26,884	26,884	19,116
Specification	OLS	IV	IV
Instrument		Sensitivity	Saiz Elast
Num. CBSAs	380	380	270
Yr-Qtr FE			✓
Region, Yr-Qtr FE	✓	✓	
CBSA FE	✓	✓	✓

Notes: Replication of Guren et al. (2021), table 1. One, two, or three asterisks represent statistical significance at the 10, 5, and 1 percent levels, respectively.

Table 4: 2SLS Housing Wealth Elasticity Estimates Using Saiz Elasticity and Land Unavailability

	Dependent variable: YoY Log Diff in Retail Emp Per Capita								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: 1978 - 2017									
YoY Log Diff in HP Growth	0.084*	0.082	0.143***	0.136**	0.081***	0.063***	0.109***	0.099***	0.064***
	(0.048)	(0.054)	(0.053)	(0.057)	(0.018)	(0.016)	(0.024)	(0.022)	(0.016)
First Stage Partial R^2	0.03	0.02	0.01	0.01	0.10	0.18	0.06	0.07	0.17
OverID p-value						0.01		0.17	0.00
Panel B: 1990 - 2017									
YoY Log Diff in HP Growth	0.141***	0.136***	0.166***	0.084	0.065***	0.070***	0.083***	0.079***	0.073***
	(0.037)	(0.040)	(0.058)	(0.063)	(0.018)	(0.014)	(0.024)	(0.022)	(0.014)
First Stage Partial R^2	0.04	0.04	0.02	0.01	0.16	0.29	0.09	0.11	0.29
OverID p-value						0.39		0.23	0.32
Panel C: 2000 - 2017									
YoY Log Diff in HP Growth	0.134***	0.121***	0.145**	0.050	0.054***	0.054***	0.067***	0.065***	0.056***
	(0.035)	(0.037)	(0.058)	(0.066)	(0.018)	(0.014)	(0.025)	(0.022)	(0.013)
First Stage Partial R^2	0.05	0.05	0.02	0.02	0.20	0.33	0.12	0.14	0.33
OverID p-value						0.91		0.14	0.24
Instrument(s)	Saiz Elas-ticity	Saiz Elas-ticity	LU 50km Circles	LU 50km Circles	LU-ML	LU-ML, Sensitivity	LU CV: Total	LU CV: Slope,Water Wetlands	LU CV: Total, Sensitivity
Num. CBSAs	270	270	270	376	376	376	376	376	376
Date FE	✓								
Region × Date FE		✓	✓	✓	✓	✓	✓	✓	✓

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: Buildable Land (BL) Summary Statistics by Decile

BL Decile	BL Mean (km^2, 000s)	BL Percent	Corr with Available Land
(1)	(2)	(3)	(4)
1	12.18	0.07	0.46
2	113.06	0.20	0.34
3	355.81	0.33	0.34
4	1013.92	0.41	0.34
5	2058.12	0.48	0.27
6	3516.77	0.58	0.46
7	5084.49	0.61	0.15
8	6883.83	0.67	0.26
9	9535.30	0.71	0.48
10	20188.21	0.75	0.83

Notes: Summary Statistics for Buildable Land (BL) Deciles based on three-digit zip codes. The computation of Buildable Land (BL) for each three digit zip code is described in the text.

Table 6: 2002 - 2006 House Price Growth by Buildable Land Decile

	Dependent variable:		
	$\Delta(\ln \text{HP})_{\text{2002-06}}$		
	(1)	(2)	(3)
Buildable Land Decile 1	58.597***	14.518***	13.283**
	(4.887)	(4.611)	(5.322)
Buildable Land Decile 2	35.124***	−8.955***	−9.166***
	(4.717)	(3.276)	(3.484)
Buildable Land Decile 3	27.319***	−16.760***	−18.068***
	(3.164)	(4.628)	(4.419)
Buildable Land Decile 4	42.914***	−1.165	−1.819
	(3.664)	(2.918)	(3.166)
Buildable Land Decile 5	44.079***		
	(4.833)		
Buildable Land Decile 6	29.078***	−15.001***	−13.624***
	(3.060)	(3.781)	(3.755)
Buildable Land Decile 7	21.289***	−22.790***	−20.849***
	(2.356)	(3.971)	(3.906)
Buildable Land Decile 8	23.240***	−20.839***	−20.751***
	(3.066)	(3.798)	(3.742)
Buildable Land Decile 9	24.520***	−19.559***	−19.910***
	(3.819)	(4.662)	(4.667)
Buildable Land Decile 10	25.174***	−18.905***	−22.593***
	(4.171)	(5.802)	(5.862)
Bartik Labor Demand Shock$_{\textit{2002-06}}$			2.074***
			(0.637)
Constant		44.079***	44.481***
		(4.833)	(4.631)
Observations	757	757	757
R^2	0.250	0.250	0.280

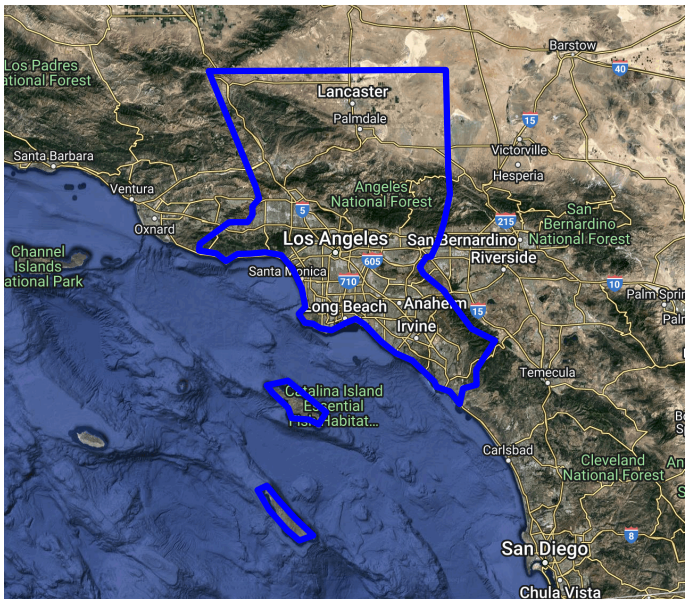
Notes: 2002 - 2006 house price growth means by Buildable Land Decile. In column (1), the intercept is excluded, and each coefficient represents the mean house price growth for the given Buildable Land decile. The excluded dummy in column (2) is Buildable Land decile 5, and thus coefficients represent the difference in means relative to decile 5. Robust standard errors are clustered at the state level.

Figure 1: Saiz Land Unavailabilty Coverage for Los Angeles

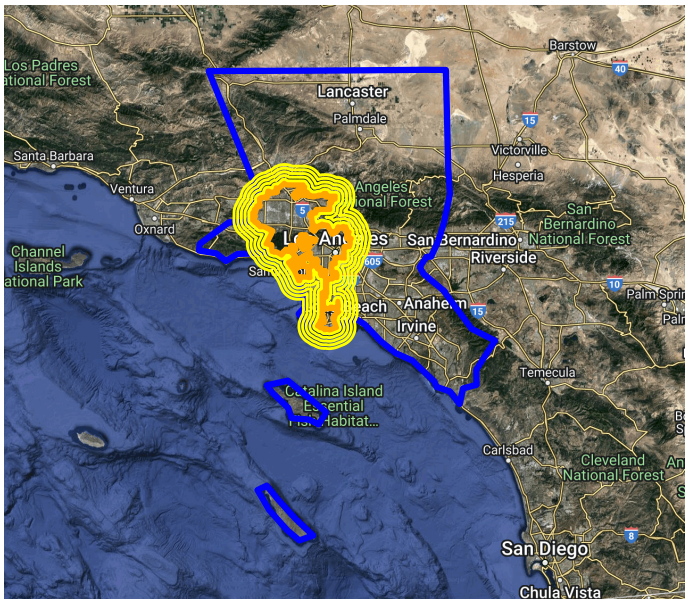
Notes: The blue line represents the polygon for the Los Angeles-Long Beach MSA. The orange lines signify the central cities within the Los Angeles MSA, and the red dots are the centroids for the central cities. The yellow circle has a radius of 50 kilometers and is centered around the polygon centroid for the first Los Angeles central city (Los Angeles).

Figure 2: LU Buffers and Circles for Los Angeles

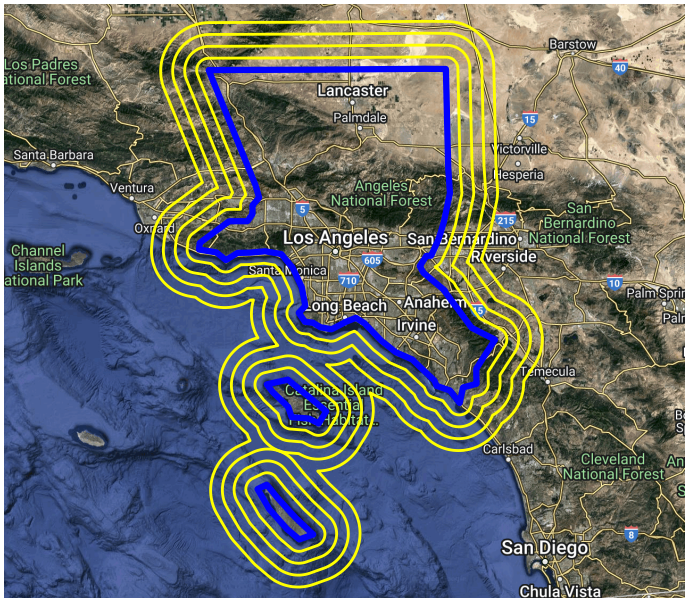
A: LA–Long Beach–Anaheim CBSA



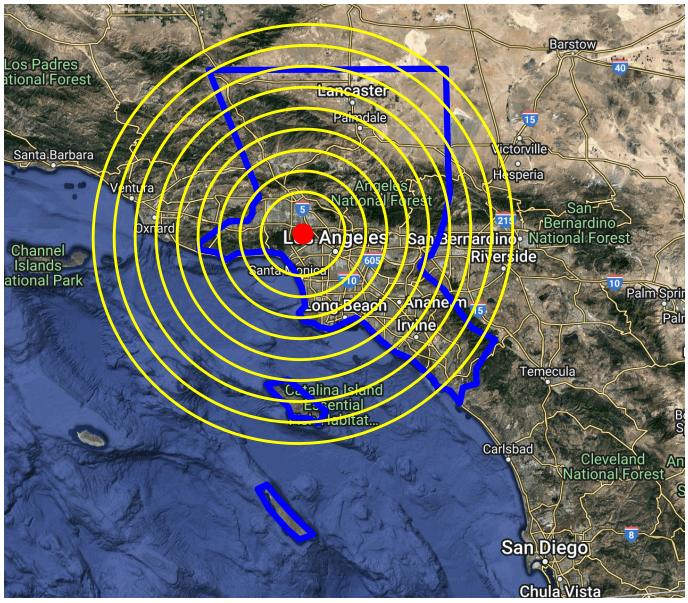
B: First Principal City Buffers



C: CBSA Polygon Buffers

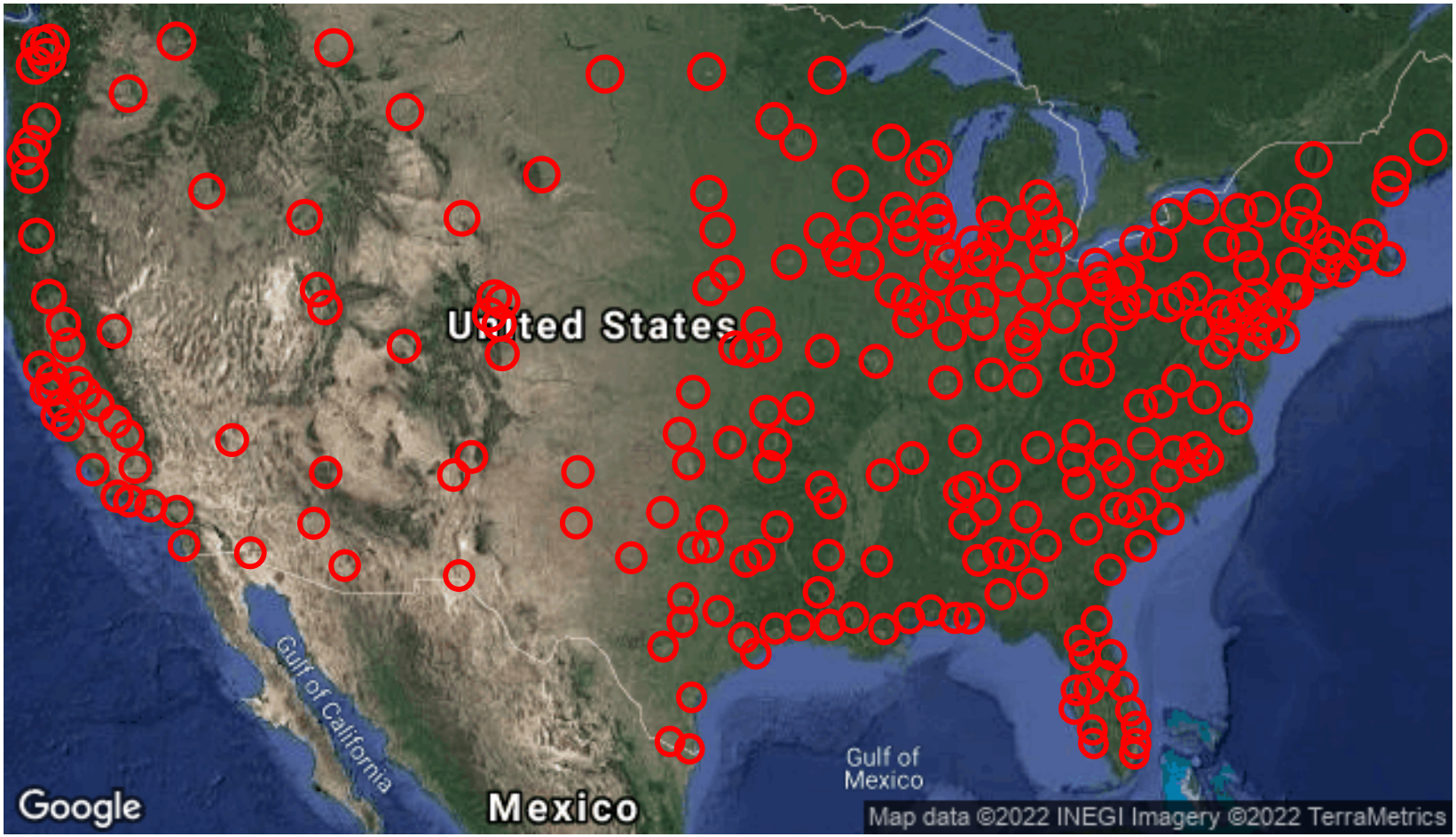


D: First Principal City Centroid Circles



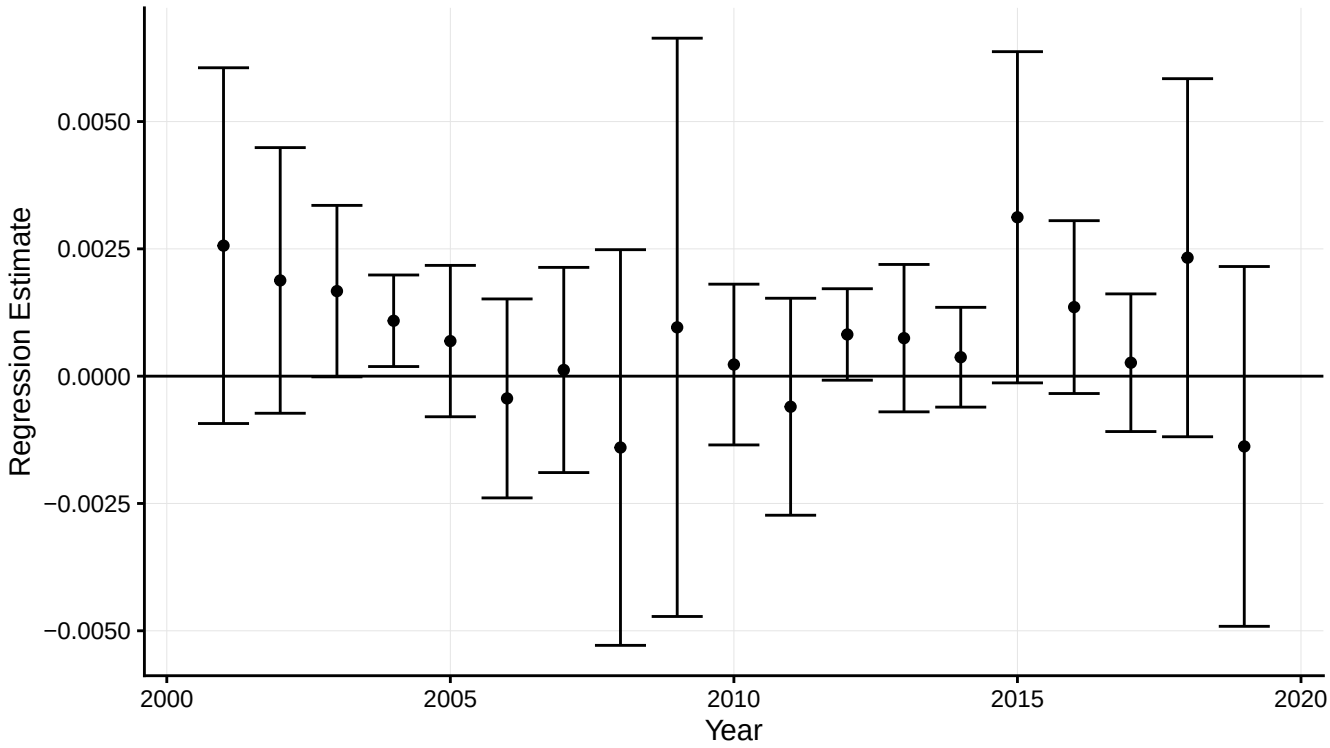
Notes: The blue line in all panels represents the polygon for the Los Angeles-Long Beach-Anaheim CBSA using 2015 delineations. Within each panel, the yellow lines represent the boundaries for which we calculate land unavailability. In panel B, the orange polygon signifies the first principal city (Los Angeles). The yellow lines correspond to 10 to 150 percent buffers (by 10 percentage point increments) of the first principal city polygon. Panel B shows 5 to 20 percent buffers (by 5 percentage point increments) of the CBSA polygon. The red dot in panel D is the principal city centroid. The corresponding yellow circles have a radius ranging from 20 to 100 kilometers (by 10 kilometer increments) centered at the first principal city centroid (red dot).

Figure 3: Saiz Land Unavailability Coverage for United States



Notes: Red circles have a radius of 50k km and are centered around the central city centroid of each MSA as in the Saiz (2010) dataset.

Figure 4: Annual Correlations between Bartik Shocks and LU



Notes: County-level correlations between LU and Bartik (1991) labor demand shocks from the following model estimated separately for each year: $Bartik_i = \alpha + \beta \cdot LU_i + \epsilon_i$. $Bartik_i$ represents the annual BLS QCEW Bartik shock for county i and LU is land unavailability for county i computed using a 5 percent buffer around each county polygon. Error bars correspond to ±2 robust standard errors clustered at the state level.

Figure 5: Florida 2001 LandCover Dataset

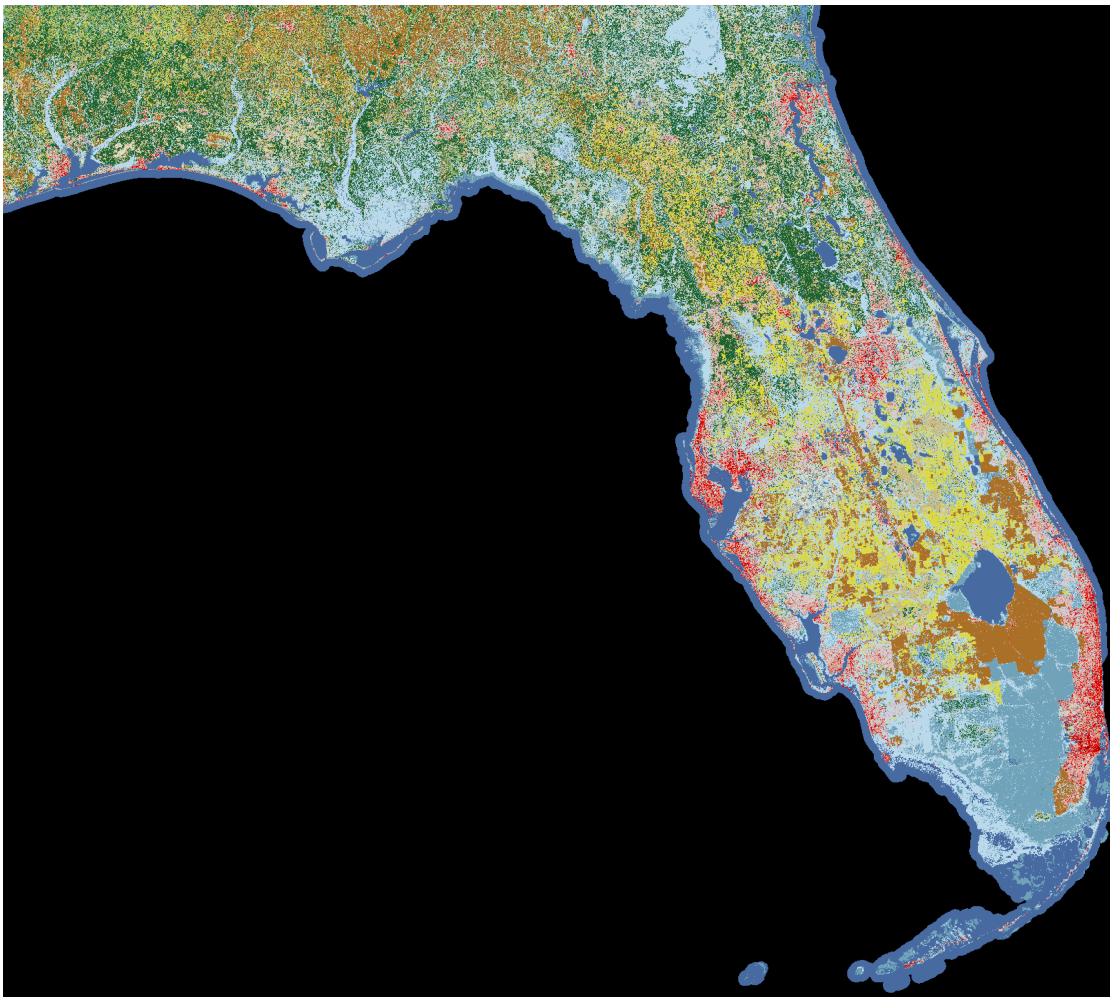
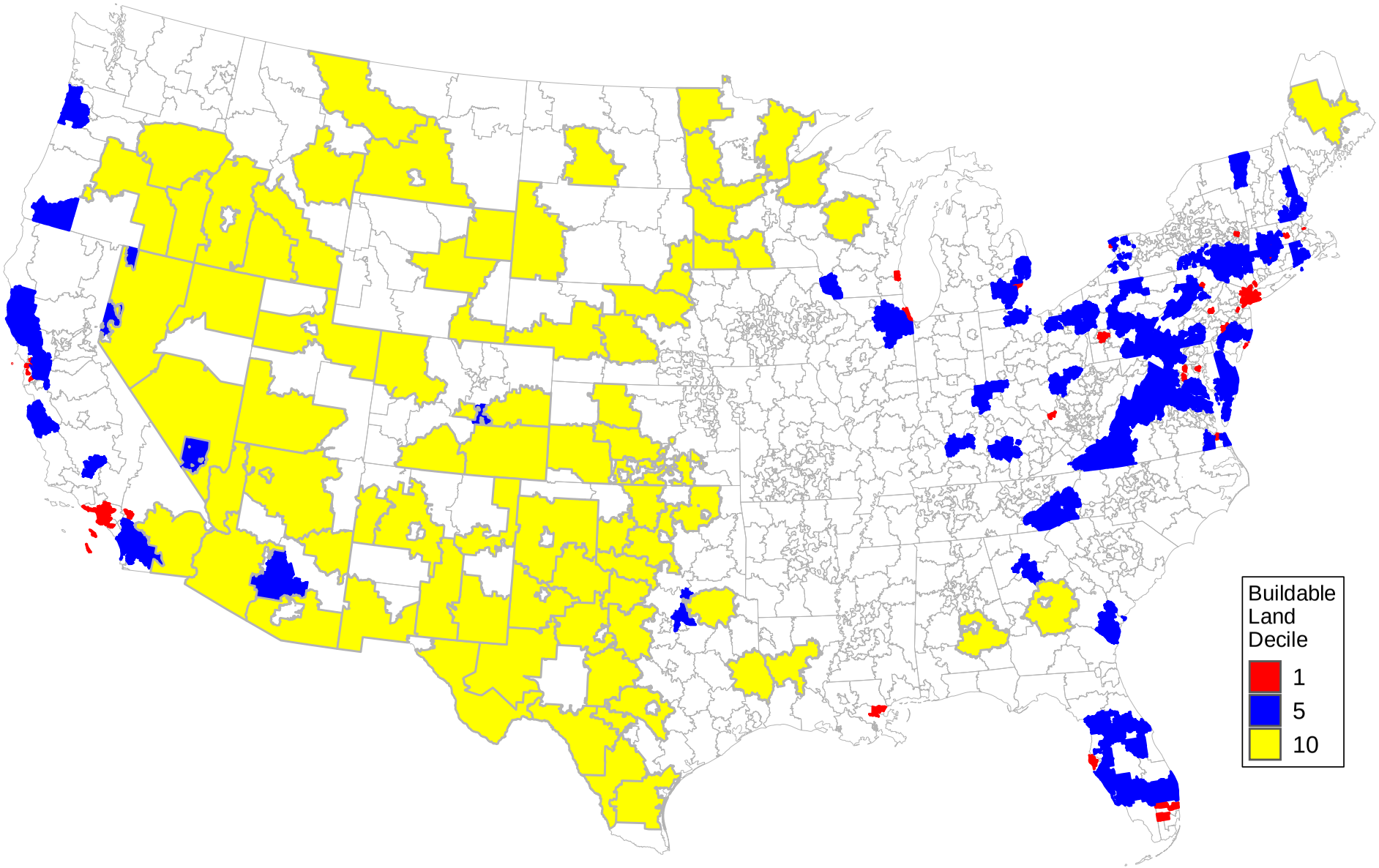
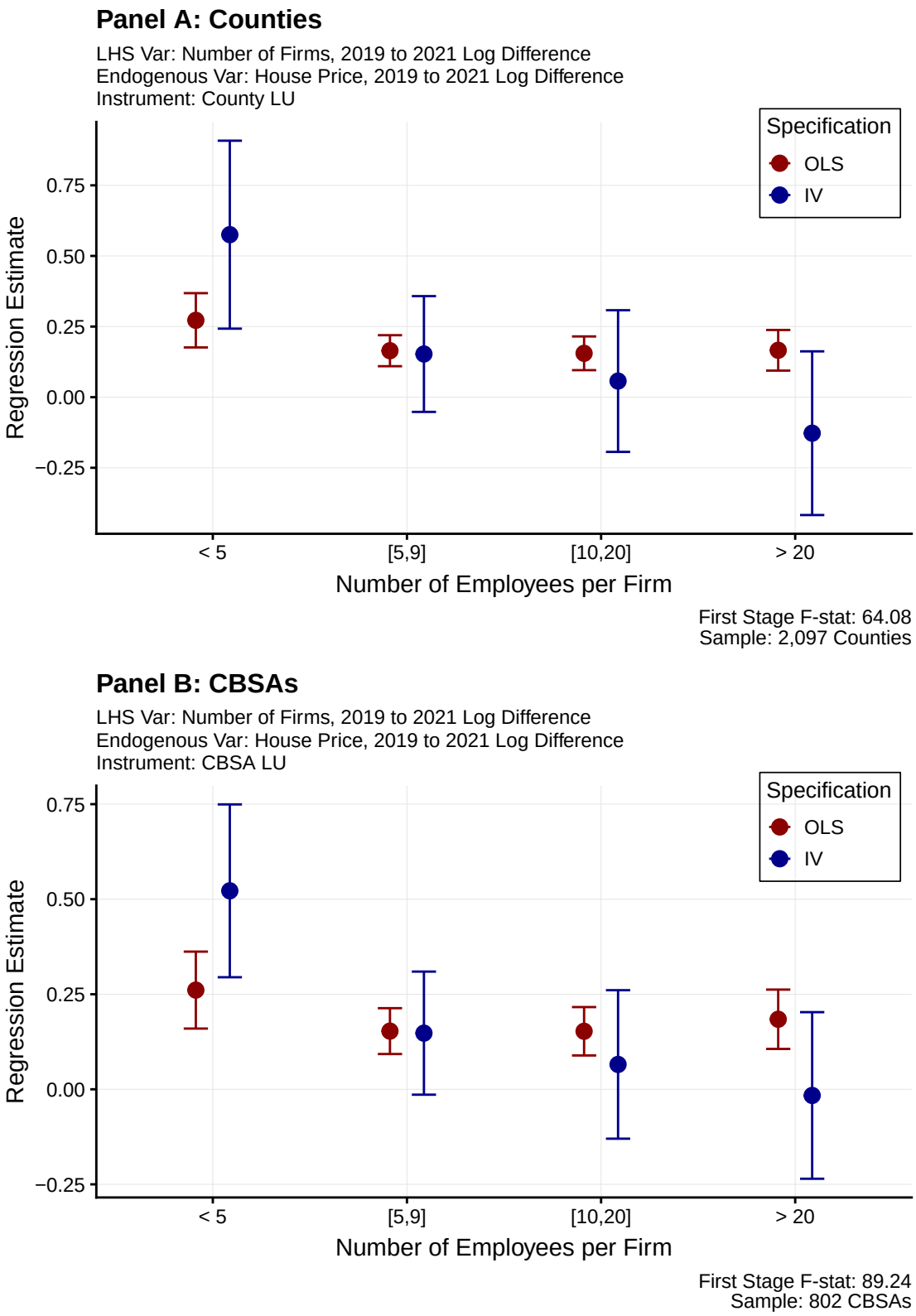


Figure 6: Three Digit Zip Codes and Buildable Land Deciles, 1, 5, and 10



Notes: Three digit zip codes. Red areas are Buildable Land Decile 1; Blue areas are Buildable Land Decile 5; and yellow areas are Buildable Land Decile 10.

Figure 7: COVID-19 Era HP Growth and Firm Counts by Firm Size



Notes: Error bands correspond to ±2.5 robust standard errors clustered at the commuting zone level.

B Appendix: LU ML Algorithms

Output is either (1) a panel dataset of house price predictions where we use machine learning algorithms to predict a panel dataset of house prices; or (2) a cross-sectional dataset with a single land unavailability estimate for each geographic unit (e.g., one LU observation for each CBSA), where the optimal geographic polygon for which to calculate land unavailability is chosen by cross-validation. (1) is useful when the econometric setup employs panel data (e.g., GMNS and equation 1), while (2) can be used when the econometric equation of interest is in long differenced form (e.g., Adelino et al. (2015) and equation 3). The below algorithm extends to LU calculated at any level of disaggregation.

The ML procedure is as follows:

- Multiply all LU proxies (e.g., those described for CBSAs in figure 2) by national house prices to create a panel as in GMNS.
- Residualize both the panel of house prices and the LU proxies with respect to fixed effects corresponding to the geographic level aggregation (e.g., for CBSAs, use CBSA fixed effects; for counties, use county fixed effects).
- Set up training and test sets. The training sets are rolling windows of time observations, inclusive of all geographic units. The test set for each training window is all time observations not used in that training window, inclusive of all geographic units.
  - In the GMNS application for the Freddie Mac house price dataset that begins in 1978, we use 10 year rolling windows. When the requested output is a panel dataset of LU-based house predictions, we also include separate training and test sets for each geographic unit's Census division.
  - In our Adelino et al. (2015) application where the Zillow data begin in 2011, we use 5 year rolling windows.
- For each training set, estimate an ML model. Obtain the fitted values for each geographic unit (e.g., for each CBSA) using the test set. Record the out-of-sample root mean-squared error for the corresponding test set predictions for each geographic unit. Repeat this step for all ML algorithms.
  - When the desired output is (1) a panel dataset of LU-based house predictions, algorithms include OLS, lasso, random forest, and XGBoost. For OLS, we build separate models for each LU proxy. For the other ML algorithms, we use all LU proxies and allow the ML algorithm to generate the prediction.
  - When the desired output is (2) a cross-sectional LU dataset where the optimal LU proxy is chosen separately for each geographic unit (e.g., each CBSA), we construct separate OLS models for each LU proxy.

- For each geographic unit (e.g., each CBSA), calculate the average out-of-sample RMSE across all test sets. For each geographic unit, choose the model with the lowest out-of-sample RMSE.
- Generate the final dataset
  - When the desired output is (1) a panel dataset of LU-based house predictions: For each geographic unit (e.g., each CBSA), run the LU proxies using all time periods through the chosen model to generate house price predictions.
  - When the desired output is (2) a cross-sectional LU dataset where the optimal LU proxy is chosen separately for each geographic unit (e.g., each CBSA): For each geographic unit, choose the LU proxy associated with the OLS model with the lowest average out-of-sample RMSE.