The Econometric Implications of Residualizing House Prices and Housing Units with Respect to Regional Income and Amenity Demand

A Methodological Note on Estimating the Effects of Land Unavailability

1 Introduction

When estimating the causal impact of geographic supply constraints—specifically, land unavailability (LU)—on local housing markets, researchers face the challenge of disentangling pure supply-side frictions from demand-side amenity premiums. Households endogenously sort into desirable locations, and richer households further improve city amenities, thereby pushing prices higher (**?**). To isolate the supply constraint, our baseline approach residualizes local house price growth with respect to regional personal consumption income (PCI) growth interacted with housing market cycles.

While this technique successfully absorbs the endogenous amenity premiums associated with regional wealth, it acts as a methodological double-edged sword. Because PCI is an equilibrium outcome inextricably linked to housing supply constraints, controlling for it introduces an attenuation bias. However, due to the specific direction of this bias, it ensures that any remaining measured effect of land unavailability is strictly conservative in nature. To further fortify this approach against critiques regarding exogenous geography (**?**), we introduce a dual-residualization strategy incorporating a natural amenity demand index.

2 The Mechanism of Endogenous Income Sorting

The limitation of residualizing house prices by PCI stems from the fact that regional income is not strictly exogenous; rather, it is a downstream consequence of the supply constraints being measured. The causal chain operates through a mechanical compositional shift in the local demographic makeup.

When a Core Based Statistical Area (CBSA) exhibits high land unavailability, the housing supply is highly inelastic. Consequently, when the area experiences a general demand shock, the inability of developers to construct new housing means the shock is absorbed almost entirely through price appreciation rather than an expansion of the housing stock.

This severe price spike initiates a compositional demographic shift that acts as a financial sieve:

1. **Displacement:** Lower-income households are outbid and forced to out-migrate to more affordable regions.

2. **Barrier to Entry:** Lower-income households from outside the CBSA are entirely priced out, preventing them from moving in.

3. **Concentration of Wealth:** The only households capable of affording the restricted housing stock are high-income earners.

Because PCI is measured as an average at the CBSA level, the substitution of low-income residents with high-income residents mechanically drives the regional PCI upward, even if individual wages remain stagnant. This influx of high-income households then triggers the endogenous amenity feedback loop, attracting premium retail, better schools, and enhanced services, which drives demand and prices up even further.

3 Endogenous PCI Growth and the Exclusion Restriction

The core concern is this: areas with binding geographic supply constraints (high land unavailability) have inelastic housing supply, which pushes up prices. Higher prices then select for wealthier households who can afford to live there. As these higher-income households move in, PCI growth rises—not because of amenities or demand fundamentals, but as a consequence of the supply constraint itself. PCI growth is therefore partly endogenous to land unavailability.

When house price growth is residualized with respect to PCI growth, the procedure strips out everything correlated with consumption growth—including this endogenous component. It inherently removes a segment of house price variation that was genuinely caused by the supply constraint channel (land unavailability → inelastic supply → higher prices → richer households sort in → higher PCI). That variation rightfully belongs to the land unavailability mechanism, but this procedure assigns it to the demand side and discards it.

This means the LU-ML measure captures less of the true supply-constraint effect on prices than it otherwise would. The estimate is attenuated—biased toward zero, not away from it. This is precisely what makes the baseline strategy conservative: if anything, the role of land unavailability as a supply constraint is being understated, not overstated.

For an instrumental variable framework, this is the safest side to err on. The primary danger with instruments is violating the exclusion restriction by capturing demand-side channels that independently affect the outcome. By aggressively purging all demand variation—even demand that is itself downstream of supply constraints—it becomes mechanically harder for the LU-ML proxy to show explanatory power. Whatever predictive power survives this aggressive residualization is cleanly attributable to structural supply constraints.

4 Addressing Exogenous Geography: A Dual-Residualization Strategy

While PCI residualization effectively isolates the endogenous income-sorting channel, it does not fully address a separate but equally critical demand-side confounder: strictly exogenous natural amenities. As noted by **?**, households intentionally sort into areas with high land unavailability because the very physical constraints causing the unavailability (e.g., oceans, coastlines, steep mountainous slopes) are themselves highly desirable amenities.

To completely isolate the supply-side constraint, we generate a second, dual-residualized LU-ML index. In this specification, we residualize house price growth with respect to both regional PCI growth *and* a static natural amenity index derived from Google Maps, with both vectors interacted with housing market cycles:

$$\Delta \log(HP_{i,t}) = \beta_1(\Delta \log(PCI_{r,t}) \times Cycle_c) + \beta_2(AmenityIndex_i \times Cycle_c) + \delta_t + \varepsilon_{i,t} \quad (1)$$

By extracting $\hat{\varepsilon}_{i,t}$ as the target for the machine learning algorithm, we purge the house price data of both the endogenous wealth sorting and the exogenous geographic premium.

5 The Time-Varying Premium of Time-Invariant Amenities

A critical methodological question arises regarding this dual-residualization approach: *If the Google Maps natural amenity index is purely static, how does it provide explanatory power beyond standard area fixed effects?*

The answer lies strictly in the interaction term. This procedure is not purging the baseline *level* of amenity capitalization. Cross-sectional variation in baseline desirability is naturally absorbed by the LU-ML measure itself. Rather, the residualization targets the *differential response* of house prices in high-amenity areas across distinct housing market cycles.

While physical geography is time-invariant, the willingness to pay for that geography is highly dynamic. During macroeconomic housing booms, wealth effects and relaxed credit constraints empower households to bid disproportionately higher premiums for scenic, highly-amenitized locations. During busts, this geographic premium contracts severely.

By interacting the static amenity index with housing market cycles, we allow amenity-rich areas to exhibit fundamentally different house price dynamics across macroeconomic shifts. We are therefore cleaning out the time-varying component of exogenous amenity demand, conditional on PCI growth already being modeled.

6 Disentangling Exogenous Geography from Local Income Dynamics

While both PCI and amenity residualizations address the demand side, they capture fundamentally distinct economic mechanisms. Relying solely on PCI controls for the endogenous "neighborhood gentrification" effect (?), but leaves the instrument vulnerable to macro-level shifts in geographic preferences that circumvent local labor markets (?). The inclusion of the cycle-interacted amenity index isolates three specific channels of variance that local PCI inherently misses:

1. **National Wealth Shocks and External Capital:** Highly amenitized, constrained markets (e.g., coastal cities) are heavily influenced by external demand. During economic booms, external capital flows into these areas via second-home buyers. This drives up local house prices without mechanically raising the *local* CBSA PCI, as the buyers' primary income is generated elsewhere.

2. **Macroeconomic Preference Shocks:** The societal willingness to pay for physical geography can shift rapidly (e.g., pandemic-era "Zoom towns"). These preference shocks cause immediate price spikes that front-run any measurable, long-term compositional shifts in regional PCI data.

3. **Pure Geographic Capitalization:** The **?** critique notes that the geographic constraint *is* the amenity. An area with stagnant PCI growth may still experience rapid house price appreciation simply because capital flows disproportionately toward naturally scarce, scenic land.

7 Constructing the LU-ML Supply Index: Residualizing Quantity

To directly test the theoretical predictions regarding housing supply elasticity, we construct a complementary LU-ML Supply Index. This index utilizes the exact same dual-residualization methodology and ML pipeline, but substitutes the log change in housing units ($\Delta \log(\mathit{Units})$) as the target variable.

7.1 Validating the Elasticity Test (Louie et al.)

The LU-ML Supply Index isolates the component of housing unit growth driven strictly by supply-side geography. We use this index to address a fundamental housing question: Do looser supply constraints dampen house price growth? Examining existing supply constraint proxies, **?** conclude that the answer is "No," contrary to standard theory, using the following specification across U.S. CBSAs:

$$\begin{aligned} \Delta \log(HP_i) = \alpha &+ \beta_1 \Delta \log(Income_i) + \beta_2 LessConstrained_i \\ &+ \beta_3 [\Delta \log(Income_i) \times LessConstrained_i] + \epsilon_i \end{aligned} \tag{2}$$

For a valid supply constraint proxy, standard theory dictates that β_3 must be negative, meaning looser constraints mitigate the house price effects of income shocks. The failure of existing proxies to produce this result is likely due to demand contamination. By explicitly purging PCI growth and amenity demand from housing unit growth prior to training the ML model, we guarantee that our $LessConstrained_i$ proxy represents pure physical capacity to build, mathematically divorced from endogenous demand dynamics.

7.2 The Extensive Margin and Demand-Driven Construction

Purging PCI is strictly necessary when the target is housing quantity due to the divergence between per capita wealth and extensive margin growth. Because PCI is measured in per capita terms, it does not mechanically capture extensive margin population shifts—a city can experience rising per capita income with a flat or declining population, generating no aggregate construction pressure.

Consider the contrast between an oil boomtown and a highly constrained coastal city like San Francisco. In an oil boomtown, a localized economic shock causes PCI to spike, attracting workers. The population surges, and a massive construction boom follows. However, that construction is fundamentally demand-driven, not a reflection of intrinsically loose geographic constraints. Without residualization, the XGBoost algorithm would observe the boomtown's high housing unit growth, note the relatively low land unavailability, and incorrectly attribute the construction boom to permissive geography rather than the exogenous demand shock.

Conversely, San Francisco exhibits exceptionally high PCI growth alongside minimal new construction. Without residualization, the algorithm observes a massive income shock yielding no quantity response, making the supply constraints appear perfectly binding.

The residualization procedure correctly strips out the income-boom-driven construction component, ensuring the XGBoost algorithm isolates the construction variation strictly attributable to geographic supply conditions. Ultimately, this reframes the core econometric question underlying the LU-ML Supply Index: *Once the quantity response is made strictly orthogonal to local income shocks, how much of the remaining cross-sectional variation in housing construction is explained by physical geography?*

8 Orthogonalizing the Target Variable Prior to Machine Learning

In a standard Ordinary Least Squares (OLS) framework, researchers account for confounding variables by including them as covariates on the right-hand side of the equation. However, this "conditioning" approach is fundamentally incompatible with non-linear machine learning algorithms like XGBoost when the goal is to construct an exogenous instrumental proxy.

If regional PCI growth were fed directly into the XGBoost feature matrix alongside physical land features, the decision trees would construct complex, endogenous interactions (e.g., branching on high income growth conditionally interacted with steep slopes). This would explicitly conflate supply and demand within the proxy itself, violating the exclusion restriction.

To prevent this, our methodology relies on a two-stage pipeline analogous to the Frisch-Waugh-Lovell (FWL) theorem. By residualizing the target variable first, we mathematically force the target vector to be strictly orthogonal to the demand shock before the ML algorithm is initialized.

Consider two hypothetical CBSAs—one highly unconstrained (Flatland) and one highly constrained (Mountain-Coast). Both experience an identical demand shock, such as a 10% growth in PCI, which establishes an "expected" baseline of new housing construction. The unconstrained CBSA constructs the expected number of units, yielding a residual of roughly zero. The constrained CBSA, unable to build, yields a deeply negative residual (a construction deficit).

When these residuals are passed to the XGBoost algorithm, the model is tasked exclusively with mapping this orthogonalized variance to granular physical geography. The algorithm identifies that the negative residual correlates strongly with the steep slopes of the constrained CBSA, and the zero residual correlates with the developable land of the unconstrained CBSA. Because the target has already been purged of the demand shock, the resulting LU-ML index is an unequivocally supply-side measure, entirely devoid of demand-side influence.

9 Disentangling Frictional from Hedonic Geography

A critical conceptual distinction in our methodology is the separation of geography's two distinct economic roles: its frictional capacity as an impediment to construction, and its hedonic capacity as a driver of household demand. Standard urban economics literature frequently conflates the two, assuming that physically constrained areas inherently possess high natural amenity value (**?**).

However, land unavailability does not map perfectly to natural amenities. Geography often acts as a severe supply constraint without offering any corresponding amenity demand. For example, while the presence of water is a fundamental constraint on buildable land, a non-scenic industrial riverfront, a flood zone, or a protected inland wetland severely restricts housing supply but generates none of the amenity demand associated with a coastal beach like Santa Monica. Similarly, severe topological slopes may manifest as unusable, un-scenic ravines rather than highly desirable hiking destinations like Boulder, Colorado.

When constructing the LU-ML Supply Index, our target variable is housing unit growth ($\Delta \log(Units)$). If we were to train our machine learning model on raw housing unit growth, the algorithm would observe significant construction in highly constrained but highly amenitized areas (where developers are willing to incur exorbitant costs to overcome physical barriers and capture the massive amenity premium). Consequently, the algorithm might incorrectly infer that physical barriers such as water or steep slopes are not binding supply constraints.

To prevent this, we dual-residualize housing unit growth with respect to both regional PCI growth and the static Google Maps natural amenity index (interacted with decadal fixed effects). Residualizing by the amenity index acts as a precision filter. It explicitly partials out the "demand-pull" of hedonic geography—the aesthetic and recreational premiums that incentivize and finance costly construction—leaving behind the pure "supply-block" variance.

By mathematically stripping out the "Santa Monica premium" and the "Boulder premium," we force the XGBoost algorithm to evaluate the steep slopes and water boundaries of an amenity-rich market identically to the unusable ravines and swamps of a non-amenitized market—strictly as a physical and regulatory friction. Therefore, the variance that survives this dual-residualization process represents a rigorously purified measure of land unavailability. When deployed in our elasticity tests (Equation 2), it ensures that our proxy captures the true structural impediment to construction, completely orthogonal to both localized income shocks and natural geographic allure.