# Identification Strategy: Isolating Physical Land Unavailability via Conditional Income Residualization

Methodology Report

## 1 Core Identification Strategy

To effectively train the Land Unavailability Machine Learning (LU-ML) algorithm, we must first isolate the portion of house price growth attributable to physical supply frictions (e.g., steep slopes, wetlands, fragmented parcels). We achieve this by residualizing house prices with respect to **Realized Per Capita Income (PCI)**, interacting the income elasticity with distinct housing market cycles.

The residualization equation is defined as:

$$\Delta \ln(P_{it}) = \sum_{k \in \text{Cycles}} \beta_k \left(\Delta \ln(I_{it}) \times \Bbbk_{t \in k}\right) + \delta_t + \epsilon_{it} \tag{1}$$

Where:

- $\Delta \ln(P_{it})$ is the log change in house prices for metro $i$.
- $\Delta \ln(I_{it})$ is the log change in realized Per Capita Income (PCI).
- $\delta_t$ represents Time Fixed Effects (absorbing national macro shocks).
- $\epsilon_{it}$ is the target residual representing the raw scarcity signal.

## 2 Justification via Diamond (2016)

Our decision to control for *Realized* PCI (rather than a Bartik instrument) is deliberate. It is designed to filter out price appreciation driven by "Amenity Sorting" and "Regulatory Scarcity," isolating the pure "Production Cost" signal of physical geography. Diamond (2016) provides the structural foundation for this approach.

### 2.1 Removing the Amenity Premium

Diamond (2016) demonstrates that high-skill, high-income workers sort into specific cities to consume desirable amenities, and they are willing to pay a premium for this access.

- Diamond finds that college graduates increasingly concentrate in high-rent cities because "the additional utility college workers gained from being able to consume more desirable amenities made them better off... despite the high local housing prices."
- In equilibrium, realized income is collinear with the value of local amenities. By residualizing house prices with respect to realized income, we mathematically remove the "amenity premium" that high-income workers pay. This filters out the signal from natural amenities (e.g., nice weather) which we do not wish to classify as "land unavailability" in the physical sense.

### 2.2 Breaking the Endogenous Amenity Loop

A key finding of Diamond (2016) is that amenities are not static; they are endogenously produced by the concentration of high-income residents.

- An increase in a city's college share (and thus income) "endogenously became more desirable places to live," fueling further in-migration and price increases.
- If we used an exogenous instrument (like Bartik), we would fail to capture this feedback loop. Using Realized PCI allows us to control for the *total* demand shock—both the initial labor demand and the resulting endogenous amenity improvement. This ensures the residual $\epsilon_{it}$ is free from the demand-side "price spiral" described by Diamond.

### 2.3 Separating Scarcity Rents from Production Costs

Diamond (2016) identifies that high-skill workers effectively accept lower *local real wages* (nominal wages minus housing costs) to access desirable cities.

- This "negative real wage differential" represents the **Scarcity Rent** paid for amenities and zoning-restricted access.
- Because this rent is a function of willingness-to-pay (Income), the relationship between Prices and Realized Income is tight ($R^2 \approx 0.49$ for rents).
- By regressing Price on Realized Income, we absorb these Scarcity Rents. The remaining residual $\epsilon_{it}$ captures price movements that *cannot* be explained by willingness-to-pay—specifically, the **Production Costs** associated with physical barriers (steep slopes, water) that make development expensive regardless of the resident's income.

## 3 The Case for Per Capita vs. Total Income

Based on the theoretical framework of Diamond (2016), we strictly residualize with respect to Per Capita Income (PCI) rather than Total Income ($PCI \times Population$). Using Total Income would introduce an endogeneity problem that contradicts the goal of isolating supply constraints.

### 3.1 The Amenity Mechanism is "Compositional," Not "Aggregate"

Diamond's central thesis is that local amenities are endogenous to the *skill mix* of the city, not the total size of the city.

- The "endogenous amenity index" is explicitly defined as a function of the city's college employment ratio.
- PCI is the direct financial proxy for this skill mix. As PCI rises, it signals that the composition of the city is shifting toward high-skill workers who fund and demand better amenities.
- Total Income can rise simply because a city adds low-skill workers (increasing $N$). This does not trigger the "endogenous amenity" feedback loop that drives up high-end housing prices. Controlling for Total Income would conflate "More Residents" with "Richer Residents," misidentifying the sorting mechanism Diamond describes.

### 3.2 Controlling for Population ($N$) "Controls Away" the Constraint

The goal of this study is to measure Land Unavailability, a supply constraint that functions by restricting the *quantity* of housing ($N$).

- Total Income equals $PCI \times N$. Including Total Income in the regression effectively controls for Population Growth ($N$).
- A highly constrained city (e.g., San Francisco) has high prices specifically because it cannot expand $N$. If we control for $N$ (via Total Income), we remove the most direct symptom of the supply constraint: the inability to add new residents.
- Diamond's equilibrium shows that while housing demand depends on total population, the *sorting* that drives price divergence is driven by the *willingness to pay* of the individual (PCI), which forces prices up when supply is inelastic.

# 4 Scarcity Rents and the Residualization Logic

Scarcity Rent (or "economic rent") is the premium paid for a good solely because its supply is limited relative to demand. In the housing context, it is the portion of the house price that is decoupled from the cost of construction—essentially an "access fee" to a constrained location.

## 4.1 The Basic Economic Mechanism

In an unconstrained market, house prices equal the Marginal Cost of Production. In a constrained market, supply cannot expand, so when demand rises, the quantity remains fixed while prices spike.

- **Left (Elastic):** Demand shifts up → Quantity expands → Price stays flat. (No Scarcity Rent).
- **Right (Inelastic):** Demand shifts up → Quantity is fixed → Price spikes. The gap between Price and Cost is the Scarcity Rent.

## 4.2 Diamond's Implicit Definition

Diamond (2016) finds that college graduates accept a "pay cut" in real terms (lower Real Wage) to live in high-amenity cities.

- This negative real wage premium is the Scarcity Rent paid to access exclusive amenities.
- Crucially, this rent is funded by Income. One cannot pay a scarcity rent without the PCI to support it. Therefore, Scarcity Rent is highly collinear with Realized PCI.

## 4.3 Why Residualization Removes It

This logic underpins the identification strategy. We aim to isolate **Physical Land Unavailability** (supply slope) from **Amenity Value** (demand shift).

$$\begin{aligned}
\text{Price} &= \text{Construction Cost} + \text{Scarcity Rent} \\
\text{Price} &\approx \text{Cost} + (\beta \times \text{Income}) \quad \textit{(Since Rent is function of Willingness to Pay)} \\
\text{Residual} &= \text{Price} - (\beta \times \text{Income}) \\
\text{Residual} &\approx (\text{Cost} + \text{Rent}) - \text{Rent} \\
\text{Residual} &\approx \text{Physical Cost Friction}
\end{aligned}$$

By controlling for Realized PCI, we mathematically subtract the Scarcity Rent—the "premium" paid for exclusivity and amenities. The remaining residual represents the friction caused by the physical difficulty of building (e.g., slopes, water) that exists independent of resident wealth.

# 5 Economic Justification for "Housing Market Cycle" Interactions

We interact the income coefficient $\beta$ with specific time periods ($k$) because the transmission mechanism from income to house prices is state-dependent. A single pooled $\beta$ would conflate credit-driven expansions with fundamental-driven expansions, biasing the residual.

## 5.1 The Defined Housing Market Cycles

**1. 1970s (The Inflationary Cycle)**

- *Justification:* Characterized by high nominal interest rates and the "Great Inflation." Housing was primarily a hedge against inflation rather than a pure income-derivative asset. The income elasticity of demand likely differs structurally here compared to the low-inflation "Great Moderation" that followed.

**2. 1980s (The Divergence Cycle)**

- *Justification:* As noted by Diamond (2016), this cycle marked the beginning of the "Great Divergence," where rust-belt cities (e.g., Detroit) declined while innovation hubs (e.g., Boston) appreciated. The sorting mechanism described by Diamond began accelerating here, justifying a distinct slope parameter.

**3. 1990s (The Stable Expansion)**

- *Justification:* A period of stable inflation and the tech boom. This provides a "baseline" elasticity for the relationship between income growth and house prices in a relatively unconstrained credit environment before the subprime expansion.

**4. 2000–2007 (The Credit Boom Cycle)**

- *Justification:* A critical interaction term. During this cycle, house prices decoupled from income fundamentals due to the expansion of subprime credit. A pooled model would underestimate the "true" income elasticity because prices rose everywhere regardless of income. Isolating this cycle prevents the "bubble noise" from contaminating the physical unavailability signal in other periods.

**5. 2007–2012 (The Bust & Deleveraging Cycle)**

- *Justification:* Income acted as a "binding constraint" differently in the bust. Prices fell due to liquidity spirals and foreclosures, often overshooting the drop in personal income. This asymmetric elasticity requires a separate coefficient to ensure residuals in this period reflect true supply conditions rather than distressed selling.

**6. 2012–2019 (The Supply-Constrained Recovery)**

- *Justification:* This cycle most closely resembles the theoretical equilibrium described by Diamond (2016), where sorting and supply constraints dominate. With credit standards tightened, price growth was strictly tethered to income growth and inventory shortages. This cycle likely provides the "cleanest" estimate of $\beta$ for identifying physical constraints.

**7. 2020s (The Pandemic & Remote Work Shift)**

- *Justification:* A structural break in housing preference. The "Space Race" and remote work shifted demand curves independent of local labor market income (e.g., people bringing NYC incomes to Florida). Interacting this cycle is crucial to account for the decoupling of *local* production wages from *local* housing demand.

# 6 The Necessity of the LU-ML Filter: Removing Non-Income Shocks

While residualizing with respect to Realized PCI removes the signal from amenity-driven sorting, the raw residual $\epsilon_{it}$ cannot be used directly as a proxy for physical supply constraints. The residual is a "catch-all" term that contains all price variation orthogonal to income, including significant noise from non-income demand shocks and idiosyncratic supply shocks. The LU-ML algorithm functions as a necessary filter to isolate the structural geographic signal.

## 6.1 Remaining Demand Shocks (The Leverage Wedge)

The raw residual $\epsilon_{it}$ contains demand shocks that shift purchasing power without changing income.

- **Credit Supply Shocks:** House prices are a function of *Purchasing Power* (Income × Leverage). A relaxation of lending standards (e.g., the 2004–2006 subprime expansion) allows buyers to bid up prices significantly without a corresponding increase in PCI. This leverage-induced appreciation appears in the residual as a "demand shock," which a raw residual approach would dangerously misclassify as a supply constraint.

- **Wealth Effects:** In high-equity markets, price appreciation is often driven by stock market gains or unrealized wealth. Since PCI measures realized income flows, these wealth-driven demand shocks remain in the residual.

- **Preference Shifts:** Structural shifts in preferences, such as the COVID-19 "Race for Space," increase the willingness to pay for housing independent of income changes. This creates a large positive residual driven by preference, not supply friction.

### 6.2 Remaining Supply Shocks (Non-Geographic)

The residual also contains supply-side noise unrelated to land unavailability.

- **Construction Cost Volatility:** Spikes in lumber or labor costs (e.g., 2021–2022) raise prices. These are temporal supply shocks, not structural land constraints.
- **Idiosyncratic Events:** Natural disasters or specific local policy interventions can cause price deviations that are uncorrelated with the underlying physical terrain.

### 6.3 The Machine Learning Filter

To address this, we do not use $\epsilon_{it}$ as the final measure. Instead, we train the LU-ML algorithm to project the residual onto a vector of strictly exogenous physical features ($X_{geo}$), such as slope, water bodies, and wetland proximity:

$$\hat{\epsilon}_{it} = f(X_{geo}) \tag{2}$$

This step acts as an instrumental variable approach. Since physical geography is orthogonal to credit cycles, wealth shocks, and construction cost volatility, the predicted value $\hat{\epsilon}_{it}$ retains *only* the portion of the residual explained by physical constraints. This effectively discards the economic noise (both supply and demand side) and isolates the structural friction of land unavailability.

## 7 Empirical Validation and Orthogonality

To verify that the LU-ML algorithm has successfully isolated the structural geographic signal and purged the macro-economic noise, we perform a rigorous post-estimation validation. This involves interpreting the projection magnitude and testing for orthogonality against the national business cycle.

### 7.1 Structural Significance of the ML Projection

The validity of the generated measure (lu_ml_xgboost) is confirmed by regressing the actual house price growth residuals on the ML-predicted values.

- **Coefficient Magnitude ($\approx 0.90$):** A coefficient near 1.0 indicates that the ML model has learned a physically meaningful relationship. It implies that for every 1% increase in "predicted pressure" from geographic constraints, actual house prices rise by approximately 0.9%. This near-unitary elasticity confirms that physical geography is a dominant driver of the residual variation, not a weak correlate.
- **Explanatory Power ($R^2 \approx 0.30$):** After controlling for all fixed effects and income sorting, the ML measure explains roughly 30% of the remaining idiosyncratic price variation. This is a substantial proportion for a cross-sectional variable, proving that physical land unavailability is a primary structural determinant of the "Great Divergence" in housing costs.

### 7.2 Dynamic Shadow Pricing (Scenario C)

We estimate the ML model separately for every time step $t$. This approach is economically critical because the "shadow price" of a physical constraint is not constant; it varies with aggregate demand intensity.

- **The Mechanism:** In a high-demand boom (e.g., 2005), the penalty for being land-constrained is high, leading to a wide dispersion in price growth between flat and rugged cities. In a bust (e.g., 2010), the penalty shrinks as demand slackens.
- **The Result:** By allowing the model to re-estimate the relationship at each time step, our measure captures the *intensity* of the constraint's bindingness. The measure $\hat{\epsilon}_{it}$ represents the "Wedge" caused by geography at that specific moment in time, rather than a static feature.

### 7.3 Orthogonality to the Macro Cycle

A potential concern with any residual-based measure is that it might accidentally capture the national business cycle (e.g., simply predicting higher values during booms). We test for this by calculating the correlation between the aggregate LU-ML measure and the National House Price Index.

- **Result:** The correlation is effectively zero ($\rho \approx -0.001$).
- **Implication:** This confirms that the LU-ML measure is orthogonal to the "tide" of the national economy. It effectively filters out the aggregate demand shocks (which lift all boats) and strictly measures the *cross-sectional dispersion* caused by supply frictions (the "rocks" the tide hits). This orthogonality allows the measure to be used as an exogenous structural variable in downstream analysis without fear of confounding from macro-credit cycles.

## 8 Conclusion

By combining the **Realized PCI control** (supported by Diamond, 2016) with **Housing Market Cycle Interactions** and the **LU-ML Filter**, we construct a measure that is orthonormal to amenity-driven sorting, macro-credit cycles, and idiosyncratic preference shocks. This leaves the final metric representing the pure, cost-push friction of physical land unavailability.