# The Commercial Real Estate Ecosystem

Ralph S.J. Koijen* Neel Shah† Stijn Van Nieuwerburgh‡

August 15, 2025

**Abstract**

We develop a new approach to understand the joint dynamics of transaction prices and trading volume in the market for commercial real estate. We start from a micro-founded model in which buyers and sellers differ in their private valuation of building characteristics, such as size, location, and quality. Consistent with the decentralized nature of the commercial real estate market, we model the probability that a seller meets a particular buyer, where the meeting probability depends on the characteristics of the buyer, the seller, and the building. In equilibrium, the mapping from building characteristics to observed transaction prices depends on the identity of the buyer and the seller, an important property missed by traditional hedonic valuation models. We estimate the model using granular data on commercial real estate transactions, which contain detailed information on the identity of buyers and sellers. Our central finding is that the identity of buyers and sellers has a first-order effect on both property valuation and the likelihood of trade. The importance of investor characteristics for valuations remains true, in fact is amplified, in a rich machine learning model that allows for non-linearities and interactions. We show how the model can be used for out-of-sample predictability and for counterfactual analyses on investment flows and prices. Our methodology extends to other private markets, including private equity, private credit, and infrastructure.

**JEL:** G11, G12, R30, R33

**kywords:** private and real assets, demand-based pricing, commercial real estate

*Booth School of Business, University of Chicago, ralph.koijen@chicagobooth.edu.

†Columbia Business School, ns3481@columbia.edu.

‡Columbia Business School, svnieuwe@gsb.columbia.edu.

The authors would like to thank Dayou Xi, Luofeng Zhou, Zi Wang, Laurenz De Rosa, and especially Michael Leahy for excellent research assistance. We thank the Fama-Miller Center at Chicago Booth and the Paul Milstein Center (PMC) at Columbia Business School for financial support to acquire the MSCI Real Capital Analytics data. The PMC also provided the Compstak data. We thank Jeff Fisher for providing the NCREIF data. The authors thank the participants at the Inquire Europe conference in Valencia, the NBER Long Term Asset Management conference, the ABFER conference, and seminar participants at the University of Michigan, Columbia Business School, and Wharton for helpful comments. Koijen thanks the Center for Research in Security Prices at the University of Chicago and the Fama Research Fund at the University of Chicago Booth School of Business.

# 1 Introduction

Financial markets have shifted markedly from public to private over the past 25 years. The number of U.S. publicly traded stocks fell from about 8,000 in 1997 to roughly 4,000 in 2024. Over the same period, assets under management in private equity broadly defined grew from a small base to $7.1 trillion in the U.S. and $13 trillion globally.[1] Many large institutions have reduced public equity and fixed-income allocations, increasing exposure to alternatives.[2] The recent surge in private debt markets is part of this trend.[3] Using a broad definition including privately held businesses and collectibles, Goetzmann et al. (2021) estimate $84 trillion in private and real assets in 2020. This "great rotation" has major implications for finance research and practice. Private and real assets differ from public ones: they trade infrequently, often in bilateral search-and-match markets; prices are rarely observed; assets are lumpy and heterogeneous; location and other idiosyncrasies matter; and specialized investor ecosystems dominate. This paper develops a methodology to value such assets.

Under traditional rational asset pricing, buyer and seller identities are irrelevant: any absent participant is replaced by another with the same expected valuation.[4] Asset "hedonics" determine value, not investor composition. Real-world markets, especially private and real assets, often violate these assumptions. We focus on commercial real estate (CRE)—a large, well-documented private asset segment with rich data on buyer and seller characteristics, spanning a diverse and evolving participant mix.

Two CRE examples illustrate investor composition effects. First, a suburban mall loses anchor tenant Sears, hurting rents. Owned by a publicly-listed mall REIT with little capacity to repurpose it, the property may be worth more to an opportunistic REPE fund like Blackstone, with flexible capital and expertise. Second, large foreign institutional investors—sovereign wealth and pension funds—favor high-end "superstar city" CRE. Their 2015 and 2018 surges raise questions: How much did they push up prices? What assets did displaced domestic buyers acquire instead? How similar were those assets? Answering them sheds light on how foreign purchases spill over to other locations, CRE sectors, and asset features. Such questions are timely as cross-border investors became net sellers after 2018.

We systematically document CRE buyer–seller composition over 2001–2023 using MSCI Real Capital

[1] McKinsey Global Private Markets Review 2024. Private asset markets include buyout, venture capital, growth, and other private equity, private debt, real estate, infrastructure, and natural resources. AUM as of June 2023; expected to double by 2029.

[2] For example, public pension funds raised their private and real asset share from below 10% in the late 1990s to over 25% by 2020 (Mittal, 2024). Alternatives include private equity, hedge funds, real estate, infrastructure, commodities, and miscellaneous alternatives.

[3] Private credit refers to lending outside the traditional banking system, with strategies including direct lending, subordinated debt, distressed debt, and special situations. The market grew from $1 trillion to $1.5 trillion between 2020 and 2024; projected to reach $2.8 trillion by 2028 (Morgan Stanley, June 2024).

[4] Even in heterogeneous-agent models, absent frictions, investors' valuations shift sharply with price, producing large flows but little price movement (Gabaix et al., 2025).

Analytics (RCA) data, which covers 470,000 transactions worth $10 trillion across apartments, offices, retail, and industrial CRE. RCA's buyer–seller mapping reveals 325,000 unique investors holding $8.6 trillion in CRE by end-2023. We classify investors into eight types, noting segmentation: smaller investors focus on smaller, local assets; larger investors trade nationally in bigger deals.

We build a model reflecting decentralized private-asset trade. Each investor's valuation depends on asset characteristics and an idiosyncratic component (e.g., liquidity shocks). Valuations vary across investors for features like size, location, and cash flow. Allowing for investor-specific feature valuation greatly improves price fit: adding investor traits to a rich Light Gradient Boosted Machine (LGBM) hedonic model raises $R^2$ by 10–20 points across sectors, with final in-sample $R^2$ of 84–88% versus 48–59% for a linear hedonic model. Our main model reduces the unexplained variation of the best non-linear model without investor characteristics between 42-55%.[5] We follow best practices in choosing hyper-parameters, mindful of overfitting concerns. Out-of-sample $R^2$ reach 81–85%.

We then estimate probabilities of listing, meeting, and transacting, with meeting likelihood shaped by buyer size, similarity to seller, prior relationships, and asset-type fit. To make estimation of the meeting model feasible, we adapt the concept of negative sampling from the NLP literature (Mikolov et al., 2013), reducing computational burden while preserving the consistency of the estimator (Ma and Collins, 2018).

Having estimated the model parameters, we turn to the main counterfactuals. A central object is the *potential price distribution* of a listed property. The seller could have met many other buyers besides the one who completed the deal. By sampling (with replacement) potential buyers in proportion to their meeting probabilities, and discarding cases with no gains from trade, we obtain a distribution of possible transaction prices. This distribution is valuable in its own right—for assessing the seller's luck or broker's skill, setting future price targets, and managing risk—and serves as a key input in three applications.

The first application addresses building owners planning to sell in the next period, who seek to predict both transaction price and price risk. Our valuation model shows that building, market, and investor characteristics all influence observed prices, capturing nonlinear and interaction effects. Forecasting price and its uncertainty thus requires predicting these inputs, with investor characteristics being the hardest to anticipate. Here the meeting model is crucial, drawing potential buyers in proportion to their likelihood of engaging and completing a sale. Our benchmark model outperforms both the linear and hedonic LGBM models in out-of-sample tests, producing prices closer to actual transactions more often and with smaller percentage errors.

The second application examines heterogeneity in valuations through a 2007 episode, just before the

[5]These gains occur notwithstanding the possibility that a rich hedonic LGBM model without investor features may infer investor-related traits from property, market, and fixed-effect variables through *triangulation*, much like Fuster et al. (2022) show a random forest model can infer race from other borrower data in mortgage default prediction.

Global Financial Crisis, involving several large transactions between REIT sellers and REPE fund buyers. Operating in a strong fundraising environment, REPE funds may have been compelled to deploy capital quickly. We ask what prices would have been if these funds had valued the properties like other investors in 2007, or if they had been absent or less active. To measure the impact of a buyer group, we remove it from the potential matching pool and recompute the potential price distribution. The difference from the benchmark distribution quantifies its price effect, with key moments being the average price and inter-quartile range as a measure of price risk. We find a large impact of eager REPE buyers on both transaction volume and average prices conditional on a transaction.

The third application uses the model to analyze substitution patterns in deal flows across building size, sector, and location. This is relevant for both policymakers and CRE investors. Concretely, if many foreign investors purchase large high-end Manhattan office buildings, what happens to smaller low-end offices in Manhattan, local apartments, or offices in other cities such as Chicago, DC, Los Angeles, and San Francisco? The model traces transaction flows along these dimensions, with the counterfactual price distribution quantifying the impact of foreign investors on prices and price risk. We find that average transaction prices in the top-quartile of Manhattan office assets would have been 4.3% lower over the full sample and 5.5% lower over the post-2013 period in the absence of foreign buyers. These effects become larger for the top-10% of office properties, illustrating the large impact of foreign buyers on trophy office properties in gateway markets.

**Related Literature** Our paper contributes to the literature on valuing private assets. Practitioner approaches often ignore risk, simply comparing total cash outflows from operations and disposition to the initial investment (TVPI) or adjusting for timing and the time value of money (IRR). The finance literature emphasizes recognizing the systematic risk in cash flows, measured in increasingly sophisticated ways. The public market equivalent (PME) approach Kaplan and Schoar (2005) benchmarks against a stock index with identical cash-flow timing. The generalized PME of Korteweg and Nagel (2016) allows for equity market exposure different from one. The strip-by-strip approach of Gupta and Van Nieuwerburgh (2021) incorporates a broader set of public-market risk factors, including cross-sectional equity and interest rate risks, and allows exposures of operating and disposition cash flows to differ.[6] While Gupta et al. (2025) value CRE assets with richer cash-flow models, they, like earlier work, assume the Stochastic Discount Factor is inferred from public data. We depart from this by not assuming that public markets set the prices of risk for private market exposures; instead, our valuation model allows investor composition to influence prices.

We also add to the literature on asset prices using asset demand systems (Koijen and Yogo, 2019; Koijen

[6]Other contributions to measuring risk and return in private equity include (Korteweg and Sørensen, 2010; Driessen et al., 2012; Ang et al., 2018).

et al., 2024), sharing the goal of jointly explaining valuations, asset characteristics, holdings, and trades. This is the first demand system for private assets. Because such assets trade in bilateral markets, our approach differs sharply: instead of modeling portfolio shares of divisible public securities, we model the probability that a whole private asset changes hands in a buyer–seller meeting. We provide a microfoundation for the valuation equation suited to this market structure. Our framework extends naturally to other private or over-the-counter markets, including corporate and municipal bonds, private credit, private equity, infrastructure, natural resources, residential real estate, and collectibles.

The standard hedonic regression—valuing assets as a linear function of fixed-priced physical characteristics—has a long tradition (Lancaster, 1966; Griliches, 1971; Rosen, 1974; Witte et al., 1979; Wallace, 1996). Our model is richer: investor composition shapes how characteristics map to prices. Transaction prices reflect both buyer and seller valuations; changes in investor mix thus affect residual variance across time and markets. Time-varying coefficients in a standard hedonic may inadvertently capture investor composition effects. For example, "number of floors" may appear to gain value in Manhattan office data simply because foreign investors prefer tall buildings. Controlling for investor mix can "cleanse" hedonic price indexes of such effects.

We contribute to a large machine learning (ML) literature in finance (e.g., Gu et al., 2020; Francke and van de Minne, 2024, for stock and CRE markets, respectively). We model dynamic investor composition within a non-linear ML framework, capturing complex interactions that linear models may miss. We also enhance the hedonic model without investor features by adding variables—such as neighborhood quality—and moving to a rich non-linear form. Even relative to this richer hedonic LGBM model, investor characteristics substantially improve explanatory power.

The REIT literature uses standard empirical asset pricing to explain returns of publicly listed CRE firms (Van Nieuwerburgh, 2019, for a recent example), but REITs account for only 12% of transaction value in our data. As CRE trading shifts from public to private markets, understanding risk and return in private CRE becomes both harder and more important.[7] Peng (2016), Sagi (2021), and Giacoletti (2021) analyze risk and return at the property level, emphasizing liquidity. Our potential price distribution complements these by providing a new liquidity risk measure. Plazzi et al. (2008, 2010) examine cross-sectional variation in expected returns and cash flow growth across U.S. cities.

Recent studies highlight the role of investor heterogeneity in CRE valuation. Ghent (2021) finds that institutional investors like REPE funds cluster in markets with many similar institutions, raising valuations through added liquidity. Similarly, Cvijanović et al. (2022) show institutional CRE investors are highly

[7] Van Nieuwerburgh et al. (2015) discuss challenges in measuring returns for infrequently traded CRE and approaches to address them.

liquidity-sensitive. Badrinza et al. (2022) stress same-nationality buyer–seller matches as a mechanism for resolving the gravity puzzle. Badarinza and Ramadorai (2018) use differences in foreign concentration across London neighborhoods to identify cross-border purchase effects on home prices. Damen et al. (2025) shows landlord types segment rental markets by quality tier.

Finally, we contribute a new estimation technique to the matching literature. We adapt negative sampling from AI and large language model training (Mikolov et al., 2013; Ma and Collins, 2018) to handle the computational burden of many assets potentially trading each period to many buyers. Fox (2018) proposes a maximum score estimator to address dimensionality in many-to-many matching. Gabaix et al. (2024) applies LLM-inspired techniques to define stock similarity using portfolio data. Badarinza et al. (2024) estimates a housing-market matching function using unique search and broker request data. Unlike them, we do not observe meetings and instead estimate the meeting function.

The rest of the paper is organized as follows. Section 2 develops the theoretical framework, which contains a valuation and a matching model, bridging key methodological gaps in the current literature. Section 3 details the estimation procedure for the models. Section 4 describes the data and provides summary statistics. Section 5 shows results for the parameter estimates of the valuation model and Section 6 for the transaction model. Section 7 explores the three application outlined above. Section 8 concludes. The appendix contains additional details on the data and empirical methods.

# 2 Model

## 2.1 Notation

There are four building types, which we call sectors: office, retail, industrial, and apartments. We estimate separate models for each sector. Time is indexed by $t, t = 1, \ldots, T$. Buildings are indexed by $n$, $n = 1, \ldots, N$, and investors by $i, i = 1, \ldots, I$. If a building is sold, then we refer to the seller as investor $s$ and the buyer as investor $b$. The transaction price is $P_t(n)$.

Buildings are differentiated by a vector of asset characteristics, $x_{nt}$. The first eight are measured at the property level: log asset size (measured in square feet for office, retail, and industrial, and in number of units for apartments), an indicator for whether the building is located in a superstar city, an indicator for whether the property is located in the central business district (CBD), the number of floors, log building age, log renovation-adjusted age, building-level occupancy rate, and a property sub-type indicator.

The next five characteristics in $x_{nt}$ are local macro-economic variables, the first four of which vary at

the market-level:[8] log population (for apartments, from the BEA) or log employment (for the other sectors, BEA), log real personal income per capita (BEA), the occupancy rate (NCREIF), and the net operating income (NOI) growth rate (NCREIF).[9] The fifth local macro variable is the net effective rent (NER) per square foot (for office, retail, and industrial buildings) or the NOI per unit (for apartments). For simplicity, we will refer to this variable as NER for all four sectors. The NER is measured at the property level, if possible, or at the block level (small geographic area) where the property is located other wise. Appendix A.3 describes the details. All five local market series vary over time. The first one is a measure of market size, the second one an indicator of the average income level in the market, the next two are key metrics of the overall health of the specific CRE sector in that market. These four market characteristics are all natural candidates to be included in a valuation model, even though they often are not because they are not strictly asset features. The fifth one, the NER, measures differences in average cash flows per unit across properties or fine geographies. We interpret it as a (time-varying) measure of asset or neighborhood quality. This NER variable, which is the most important of the five, is new to the hedonic literature.

The final two characteristics that enters into $x_{nt}$ are two transaction type categorical variables: whether the asset belongs to a distressed sale and whether the asset belongs to an entity sale–where an entire legal entity that potentially owns an entire portfolio of properties is transacted.

Investor heterogeneity is captured by a vector of six characteristics, $z_{it}$. This vector includes information on the investor's type (e.g., REIT, PE fund or foreign investor), the log dollar value of the investor portfolio across all sectors, the share of the overall portfolio in superstar markets, the share of the portfolio in the same market as the asset in question, the share of the portfolio in the same sector, and whether the investor transacted as a part of a Joint Venture (JV).

## 2.2 Investors' Private Valuations

We first provide the main empirical specification of an investor's private valuation in Section 2.2.1, for which we provide a micro foundation in Section 2.2.2.

[8] We define 60 markets. Appendix A.2 shows the mapping from RCA markets to our 60 markets.

[9] The National Council for Real Estate Investment Fiduciaries (NCREIF) is a consortium of large institutional real estate investors that share data on their institutional-quality portfolios allowing for the construction of market-level cash flow and occupancy statistics. We thank NCREIF for making these data available to us. We build a bridge between the NCREIF markets and our 60 markets. Appendix A.2 provides the details. Net operating income is a standard cash-flow metric in CRE markets that subtracts operating expenses from rent revenues.

### 2.2.1 A characteristics-based model of investors' private valuations

We assume that investor $i$'s valuation of building $n$ is given by

$$v_{it}(n) \equiv \ln V_{it}(n) = h\left(z_{it}, x_{nt}; \gamma_t\right) + \epsilon_{it}(n), \tag{1}$$

where $\gamma_t$ is a vector of parameters that may vary over time. The error term, $\epsilon_{it}(n)$, captures liquidity or funding constraints, unobserved quality characteristics or differences in investors' beliefs. We assume that the valuation shocks are uncorrelated across investors and buildings, conditional on all characteristics (including investor-type and geography fixed effects), and normally distributed, $\epsilon_{it}(n) \sim N\left(0, \sigma_t^2\right)$.

In our empirical specification, we use a non-linear ML model that flexibly captures interactions between investor and asset characteristics without imposing a specific functional form. Market fixed effects are included to capture local heterogeneity.

### 2.2.2 A micro foundation of the private valuation model

Before proceeding, we provide a micro foundation of the valuation equation in (1) and, in particular, how an investor's valuation depends on characteristics, $h_{it}(n) = \beta'_{x,i} x_{n,t}$, with a loading vector (price vector for the hedonics) that itself depends on investor characteristics $\beta_{x,i} = \beta_x z_{i,t}$. To explain the key insight, we focus on a special case where $h_{it}$ and $\beta_{x,i}$ are linear.

We consider a two-period model, $t = 0, 1$. In period $t = 0$, the investor considers buying a building. In period $t = 1$, the investor receives the net cash flow and the resale value of the building. In mapping the model to the data, we think of $t = 1$ as the payoff of the building over a longer holding period.[10] The building in question may be part of a broader property portfolio or, in case of a pension fund or insurer, part of a broader portfolio that may include equities and fixed income assets. Without the new building, the broader portfolio generates a payoff $D_{1i}$. If the investor has no other assets, $D_{1i} = 0$.

At $t = 0$, the investor has access to cash $C_{0i}$ that can be used to purchase the building. Any remaining cash earn a gross risk-free rate that is normalized to one. This implies that, without purchasing the new building, the investor's wealth in $t = 1$ is $A_{1i} = D_{1i} + C_{0i}$.

The new building generates a time $t = 1$ payoff of $N_{1i}$. We assume that $(D_{1i}, N_{1i}) \sim N(\mu_i, \Sigma_i)$. We denote elements of $\mu_i$ and $\Sigma_i$ by their logical counterparts, e.g., $\mu_{Ni}$ and $\Sigma_{NDi}$. We allow investors to have heterogeneous beliefs about future payoffs and index all moments by $i$. If the investor adds the building to her portfolio, period $t = 1$ wealth equals $A_{1i}^P = D_{1i} + C_{0i} - P_0 + N_{1i}$, where $P_0$ is the purchase price

[10] At the expense of some additional notation, this micro foundation can be extended to a multi-period model.

of the property.[11] To complete the model, we assume that the investor has mean-variance preferences over terminal wealth:

$$\mathbb{E}_i[A_{1i}] - \gamma_i \text{Var}_i(A_{1i}),$$

where $\gamma_i$ is risk aversion.

As the building is a discrete purchase (as opposed to buying shares in a company), we compute the investor's valuation as the $t = 0$ price $P_0 = V_{0i}$ that makes the investor indifferent between purchasing the building or not. This valuation then solves the following equation:[12]

$$\mathbb{E}_i[D_{1i} + C_{0i}] - \gamma_i \text{Var}_i(D_{1i}) = \mathbb{E}_i[D_{1i} + C_{0i} - V_{0i} + N_{1i}] - \gamma_i \text{Var}_i(D_{1i} + N_{1i}),$$

where the left-hand side is the investor's utility if she does not buy the building and the right-hand side is the utility when she does buy the building. We use this equation to solve for the investor's private valuation:

$$V_{0i} = \mathbb{E}_i[N_{1i}] - \gamma_i \text{Var}_i(N_{1i}) - 2\gamma_i \text{Cov}_i(D_{1i}, N_{1i}), \tag{2}$$

which depends on the expected payoff of the property, $\mathbb{E}_i[N_{1i}]$, a discount for its variance, $\text{Var}_i(N_{1i})$, and a further discount or premium that depends on the property's covariance with other assets in the investor's portfolio, $\text{Cov}_i(D_{1i}, N_{1i})$.

The main valuation equation is (2). To obtain a characteristics-based model of investors' private valuations, we follow Koijen and Yogo (2019) and model the moments as functions of characteristics with investor-specific coefficients:

$$\begin{aligned}
\mathbb{E}_i[N_{1i}] &= \beta_{i0}' x_n, \\
\gamma_i \text{Var}_i(N_{1i}) &= \beta_{i1}' x_n, \\
\gamma_i \text{Cov}_i(D_{1i}, N_{1i}) &= \beta_{2i}' x_n,
\end{aligned}$$

Beliefs about first and second moments, risk aversion $\gamma_i$, and time-1 portfolio payoffs $D_{1i}$ are all allowed to be heterogeneous across investors. We model this heterogeneity across investors as a function of the investor characteristics including investor type, size of the investor portfolio, share of assets in superstar cities, share

[11] We assume throughout that the investor has deep pockets and $C_{0i} > P_0$. The model can easily be extended so that the investor finances a fraction $\phi$ of the property with a mortgage at a rate $R > 1$, or the investor's valuation is bounded by $C_{0i}$, which caps the valuation at $C_{0i}$.

[12] To focus on the key economics, we abstract from transaction costs here.

of assets in Office, etc. :

$$\beta_{ki} = \beta_k' z_i,$$

for $k = 0, 1, 2$. This yields the valuation equation in (1).[13]

## 2.3 Transaction Prices

There is a proportional cost of $cP_t(n)$ if investor $i$ and $j$ transact. The bargaining power of the seller is measured by $\beta \in [0, 1]$, which is used to split the proportional transaction cost and in setting the price. Investor $b$ is willing to buy the building when

$$V_{bt}(n) > P_t(n)(1 + \beta c),$$

or, equivalently,

$$v_{bt}(n) - \chi_b > p_t(n),$$

where $\chi_b \equiv \ln(1 + \beta c) \simeq \beta c$. Analogously, investor $s$ is willing to sell the building when

$$V_{st}(n) < P_t(n) - (1 - \beta)cP_t(n),$$

or, equivalently,

$$v_{st}(n) - \chi_s < p_t(n),$$

where $\chi_s \equiv \ln(1 - (1 - \beta)c) \simeq -(1 - \beta)c$.

We assume that the price is set by Nash bargaining (in logs),[14] implying

$$p_t(n) = \beta v_{bt}(n) + (1 - \beta)v_{st}(n) - c.$$

This implies that trade takes place when

$$v_{bt}(n) - v_{st}(n) > c.$$

In our empirical analysis, we set $c = 0$ and $\beta = 0.5$.

[13] The difference between a valuation in levels versus logs arises because we assume mean-variance preferences. Alternatively, we can assume CRRA preferences and log-linearize the model.

[14] Formally, the price is the solution to $p^* = \arg\max \left((p - c) - v_s\right)^{\beta} \left(v_b - (p + c)\right)^{1-\beta}$.

## 2.4 Transaction Probabilities

We now discuss the listing model, the meeting probability conditional on a listing, and the probability of transacting conditional on a meeting. At this stage, we assume that we know the parameters of the valuation model. We therefore have the component of valuations explained by characteristics, $h_b$ and $h_s$, for the buyer and seller.

We model the probability of transaction as:

$$\pi(b,s) = \pi_{l,t}\pi_{m,t}(b,s)\pi_{\tau,t}(b,s), \tag{3}$$

where the first term, $\pi_{l,t}$, is the probability of listing, $\pi_{m,t}$ the probability of meeting the listed building, and $\pi_{\tau,t}$ the probability that the transaction goes through (conditional on a meeting).

We assume that the listing probability is a function of time and sector, $\pi_{l,t}$.[15]

The second term is the meeting probability, the key object in our transaction model. We assume that the likelihood of a buyer $b$ meeting a building listed for sale depends on buyer and seller characteristics, past interactions, and the degree to which the property characteristics align with the buyer's prior property portfolio:

$$\pi_m(b,s) = \begin{cases} 0 & \text{if } b = s \\ \frac{\exp\left(\lambda_1 S_b + \lambda_2 N_b + \lambda_3 \Delta S_{b,s}^{-1} + \lambda_4 T_{b,s} + \lambda_5' \delta_{b,s}\right)}{\sum_{c \neq s} \exp\left(\lambda_1 S_c + \lambda_2 N_c + \lambda_3 \Delta S_{c,s}^{-1} + \lambda_4 T_{c,s} + \lambda_5' \delta_{c,s}\right)} & \text{otherwise,} \end{cases} \tag{4}$$

When $\lambda_1 > 0$, larger buyers are more likely to meet the asset, consistent with their higher trading activity and greater market presence.[16] When $\lambda_2 > 0$, buyers that already own multiple assets (i.e., $N_b = 1$) have a higher likelihood of being matched. When $\lambda_3 > 0$, matching probabilities increase when buyers and sellers have similar size (high $\Delta S_{b,s}^{-1}$). Investors may prefer to deal with investors of similar size to reduce deal execution risk. When $\lambda_4 > 0$, buyers are more likely to match with sellers with whom they have a prior transactional relationship ($T_{b,s} = 1$), reflecting the role of established counterparties in reducing search and execution frictions. The vector $\delta_{b,s}$ captures the buyer's consideration set. It measures the similarity between the characteristics of seller $s$'s building and the portfolio characteristics of buyer $b$. Similarity has four dimensions. First, asset size alignment, the similarity between the building's log dollar value and the

[15] More advanced models could model the listing decision as a function of the price dynamics in a certain market, for example to capture the importance of reference points (Genesove and Mayer, 2001; Andersen et al., 2022). In this case, the listing probability would increase in last period's price in that market relative to a running maximum of the price index. In such a model, if prices declined sharply in the prior period, buildings owners would be less likely to list. Additionally, listing probabilities could reflect market flow dynamics, such as the entry or exit of large investors or shifts in the composition of investor types.

[16] In reality, a seller may have bilateral meetings with multiple buyers, and choose among those making a bid for the asset based on multiple criteria ranging from offered price to certainty of execution. The buyer size as well as some of the other variables in (4) capture in reduced-form that larger buyers are more likely to emerge as winners from this process.

average log dollar value of the assets in the buyer's portfolio. Second, market alignment, the similarity between the building's geography and the share of the buyer's portfolio in that geography. Third, sector alignment, the similarity between the building's sector and the share of the buyer's portfolio in that sector. Fourth, quality alignment, proxied by the log NER of the building and the average log NER in buyer b's portfolio. Hence, $\lambda_5$ is a $4 \times 1$ vector.[17]

When $\lambda_i = 0$, $i = 1, \ldots, 5$, the matching probability is the inverse of the number of buyers and the model specializes to have random matching. When $\lambda_i \geq 0$, $i = 1, \ldots, 5$, (4) tilts the model away from random matching, towards directed search in ways that will be estimated from the transaction data.

The third term in (3) measures whether a transaction takes place conditional on a meeting: $\pi_\tau(b, s) = P(V_s < V_b)$. We assume that this is the case when the valuation of the buyer exceeds that of the seller:

$$P(V_s < V_b) = P(\epsilon_s - \epsilon_b < h_b - h_s) = \Phi\left(\frac{h_b - h_s}{\sqrt{2}\sigma}\right).$$

under our assumption that $\epsilon_b, \epsilon_s \sim N(0, \sigma^2)$.

The probability that a building does not trade is given by:

$$\pi_n(s) = (1 - \pi_l) + \pi_l \sum_b \pi_m(b, s)(1 - \pi_\tau(b, s)). \tag{5}$$

The first term, $1 - \pi_l$, is the probability that the seller does not list the building. The second term is the probability that the building lists, and thus a match will be formed, but the valuation of the buyer does not exceed the valuation of the seller, as captured by $1 - \pi_\tau(b, s)$. As $\sum_b \pi_m(b, s) = 1$, $\sum_b \pi_m(b, s)(1 - \pi_\tau(b, s)) \leq 1$ and $\pi_n(s) \leq 1$. The probabilities that the building transacts with some buyer and that it does not transact sum to one:

$$\pi_{l,t} \sum_b \pi_m(b, s)\pi_\tau(b, s) + \pi_n(s) = 1.$$

# 3 Estimation Procedure

The model is estimated sequentially, in two steps. In Section 3.1, we discuss the estimation of the valuation equation. We discuss the estimation of the listing and matching model in Section 3.2.

[17] For a given characteristic $x_s$ of the building and $x_b$, the average characteristic of the buyer's portfolio, similarity is derived from the distance $|x_b - x_s|$ by converting smaller distances into higher similarities. Specifically, we use 1 - $|x_b - x_s|$ for geography and sector alignment and $\frac{1}{1+|x_b - x_s|}$ for size and NER alignment.

### 3.1 Valuation Model

We begin by estimating the valuation model, which is done separately for each sector (A, O, I, and R). We rely on observable transaction prices to estimate the other parameters. The log price is given by

$$p_t(n) \quad = \quad \frac{1}{2}(h_{bt}(n) + h_{st}(n)) + \frac{1}{2}(\epsilon_{bt}(n) + \epsilon_{st}(n)). \tag{6}$$

We note that prices are observed only when $v_b > v_s$, which is when $h_b - h_s > \epsilon_b - \epsilon_s$. The difference in shocks is uninformative about the sum of the shocks because

$$\mathbb{E}\left[\epsilon_b + \epsilon_s \mid \epsilon_b - \epsilon_s\right] = 0,$$

when $\epsilon_b, \epsilon_s \sim N(0, \sigma^2)$. This implies that estimating the valuation model based solely on the observed transactions does not create a bias.

The investor valuations $h_{bt}$ and $h_{st}$ contain $\dim(x) + \dim(z) + 2$ predictors, where the last two predictors refer to the RCA Market and year categorical variables. To flexibly capture the complex relationships between asset and investor characteristics, we use LightGBM (Guolin Ke, 2017), a gradient-boosted decision tree model. LightGBM is well-suited for this task as it efficiently handles high-dimensional data, categorical variables, captures non-linear interactions without requiring pre-specified terms, and incorporates built-in regularization to mitigate overfitting. We use a two-step "coordinate descent" procedure with two separate LightGBM regressors for the buyer and the seller. In each iteration, we first hold the seller-side model fixed and update the buyer-side function by setting the training target to $2p(n) - v_s(x_n, z_s; \gamma)$. Next, we hold the buyer-side fixed and update the seller-side function using $2p(n) - v_b(x_n, z_s; \gamma)$ as the target. Alternating these steps ensures that, at convergence, the final predictions satisfy $p(n) = 0.5 \cdot \big(v_b(x_n, z_s; \gamma) + v_s(x_n, z_s; \gamma)\big)$ in a flexible, non-linear manner. The algorithm converges quickly in a few iterations, due to the high correlation between $v_b$ and $v_s$. We discuss the empirical implementation of the valuation model in Section 5.

### 3.2 Listing and Meeting Probabilities

Having estimated the valuation equation and obtained estimates of $h_{bt}(n)$ and $h_{st}(n)$, we turn to the estimation of the listing and meeting probabilities. In theory, we can use maximum likelihood. The contribution of the building of seller $s$ to the likelihood is:

$$\ell_s = \sum_{b=1}^{B} y_{b,s} \ln \pi(b, s) + (1 - \sum_{b=1}^{B} y_{b,s}) \ln \pi_n(s),$$

where $y_{b,s}$ equals one when a transaction take place and zero otherwise.

In practice, computing this likelihood is computationally infeasible due to the large number of buildings (and thus sellers) and buyers. We instead estimate the model using negative sampling, inspired by applications in the natural language processing literature (Mikolov et al., 2013). Intuitively, instead of using the full likelihood, we consider a classification problem. For each transaction, we sample another $K-1$ potential buyers $c$ that did not purchase the building from seller $s$, $c \in \mathcal{N}_s$, $|\mathcal{N}_s| = K-1$. We then compute the probability that the actual buyer $b$ purchases the building, conditional on one of the $K$ potential buyers (including the actual buyer) purchasing the property:

$$\pi_r(b,s) = \frac{\xi_{b,s}}{\xi_{b,s} + \sum_{k \in \mathcal{N}_s} \xi_{k,s}}, \tag{7}$$

where $\xi_{b,s} = \exp\left(\lambda_1 S_b + \lambda_2 N_b + \lambda_3 \Delta S_{b,s}^{-1} + \lambda_4 T_{b,s} + \lambda_5' \delta_{b,s}\right) \pi_\tau(b,s)$.[18] We then minimize the loss function over the observed transactions, i.e., $(b,s)$ pairs: $-\sum_s \ln \pi_r(b,s)$. This estimator is the ranking estimator in Ma and Collins (2018), who also develop the asymptotic theory. They show that the estimator is consistent for $K \geq 1$, asymptotically normal, and converges to the maximum-likelihood estimator as $K \to \infty$.[19]

Lastly, we estimate the listing probability, $\pi_{l,t}$. The total number of buildings that transact in a given sector-year in the data is denoted by $T_t$. We estimate the listing probability such that $T_t$ equals the expected number of transactions in the model in that same sector-year:

$$\sum_s \pi_{l,t} \sum_b \pi_m(b,s)\pi_\tau(b,s) = T_t, \tag{8}$$

which implies the sector-year listing probability:

$$\pi_{l,t} = \frac{T_t}{\sum_s \sum_b \pi_m(b,s)\pi_\tau(b,s)}. \tag{9}$$

### 3.3 Potential Price Distribution and Counterfactuals

For several applications discussed below, we are interested in computing the distribution of prices $f(p)$ at which a trade could have occurred for an asset that is listed for sale. To do this systematically, we introduce a new tool – the potential price distribution, the distribution of prices that could have cleared if the asset had met different buyers and if both sides' idiosyncratic shocks had realized differently. It delivers

[18] The common denominator, $\sum_c \exp\left(\lambda_1 S_c + \lambda_2 N_c + \lambda_3 \Delta S_{c,s}^{-1} + \lambda_4 T_{c,s} + \lambda_5' \delta_{c,s}\right)$, which creates the computational bottleneck, drops out, radically simplifying the objective function.

[19] Similarly, section 7 of the working paper by McFadden (1977) discusses limiting the number of alternative choices in the context of a multinomial logit model, and shows that consistent estimates can be obtained from random samples of alternatives (Theorem 2). Fox (2007) considers a semiparametric random utility model that produces a similar result.

two objects: the trade probability, and the distribution of prices conditional on trade. We summarize the latter with its mean and interquartile range as level and price-risk (liquidity) measures. The realized price's percentile within its own potential price distribution is an informative benchmark for luck versus selection or for assessing broker skill.

Once the parameters of the valuation and matching models have been estimated, we implement the following algorithm to get the potential price distribution. For each asset $n$ in the market at time $t$, we know the asset features $x_{n,t}$ and the systematic component of the sellers' valuation $h_{s,t}$. For each asset, we then draw a seller shock $\epsilon_s \sim N(0, \sigma_t^2)$, resulting in a seller valuation $v_{s,t} = h_{s,t} + \widehat{\epsilon}_{s,t}$. Next, we compute the probability $\pi_m(c, s)$, given by (4), that this seller meets a potential buyer $c$ from the pool of all investors. we randomly draw one potential buyer form the distribution $\{\pi_m(c, s)\}_{c=1}^{B}$. For the potential buyer that is drawn, we draw a latent demand shock $\epsilon_{c,t} \sim N(0, \sigma_t^2)$ and construct that candidate buyer's valuation $v_{c,t} = h_{c,t} + \epsilon_{c,t}$. Using the pricing rule from Section 3.1, the property transacts if $v_b > v_s$, in which case, the price is $\hat{p}_t(n) = \frac{1}{2}(v_{c,t} + v_{s,t})$. If the valuation of the buyer is too low, $v_{c,t} < v_{s,t}$, we set the price to missing. We can now compute the fraction of assets for which a trade takes place, $\tau = \frac{1}{N}\sum_n 1_{\hat{p}(n)>0}$, and the average price conditional on a trade $Ep = \frac{1}{N}\sum_n \hat{p}(n)1_{\hat{p}(n)>0}$. We repeat this procedure $C = 1,000$ times, drawing different potential buyers for each property, with replacement from the model-implied distribution $\{\pi_m(c, s)\}_{c=1}^{B}$. The resulting empirical distribution of these statistics is the potential price distribution.

As the procedure draws both buyer and seller shocks and simulates matches without conditioning on which buildings traded or which buyers met which sellers, it is not subject to the selection bias inherent in observed transactions. The procedure replays the same core ingredients –valuations, meeting probabilities, and the pricing rule that generate actual trades, but over the full set of feasible counterfactual matches. The implementation details can be found in Appendix A.4.

The potential price distribution reflects the alternative buyers the seller could have met and the prices that would have resulted from those meetings, if indeed the deal was consummated. The potential price distribution is a useful tool in private markets with bilateral price formation, with immediate applications for price setting strategy and risk management.

This algorithm is also how we compute counterfactuals. To understand the importance of a subset $G$ of investors, we recompute the potential price distribution removing that group from the buyer pool. The sampling of $C$ potential buyers is now from the set $B \backslash G$. Comparing the baseline and counterfactual means of potential price distributions $Ep$, help us see the effect of each group of buyers on average prices, and comparing IQRs provides a metric of how the buyer pool affects the price risk of an asset. Comparing trading probabilities $\tau$ allows us to assess the importance of group $G$ for the liquidity of an asset or a market.

Beyond examining the price impact of excluding certain buyer types from specific markets, we are also interested in studying spillover effects. When a particular investor type purchases a property, other potential buyers who might have acquired that asset are displaced, purchasing other properties instead. By identifying these displaced buyers, we can track which properties they actually ended up buying, providing insight into the reallocation of demand. This allows us to analyze the characteristics of the properties they purchased—such as sector, location, and asset quality—and assess how their alternative transactions shaped market outcomes. Comparing these patterns across different investor types and time periods helps us understand how competition for assets in one segment influences liquidity, pricing, and allocation dynamics across the broader market.

# 4 Data

Our data consists of the universe of commercial real estate transactions over $0.5 Million in the four main property sectors (apartments, office, retail, industrial) between 2001 and 2023, about 370,000 transactions representing $10 trillion in transaction value. The data are from MSCI Real Capital Analytics (RCA).

RCA has spent enormous effort unraveling the true buyer(s) and seller(s) involved in each transaction, as well as an investor type classification.[20] Appendix A.1 shows our data cleaning algorithm that leads to the final data sample used in the estimation.

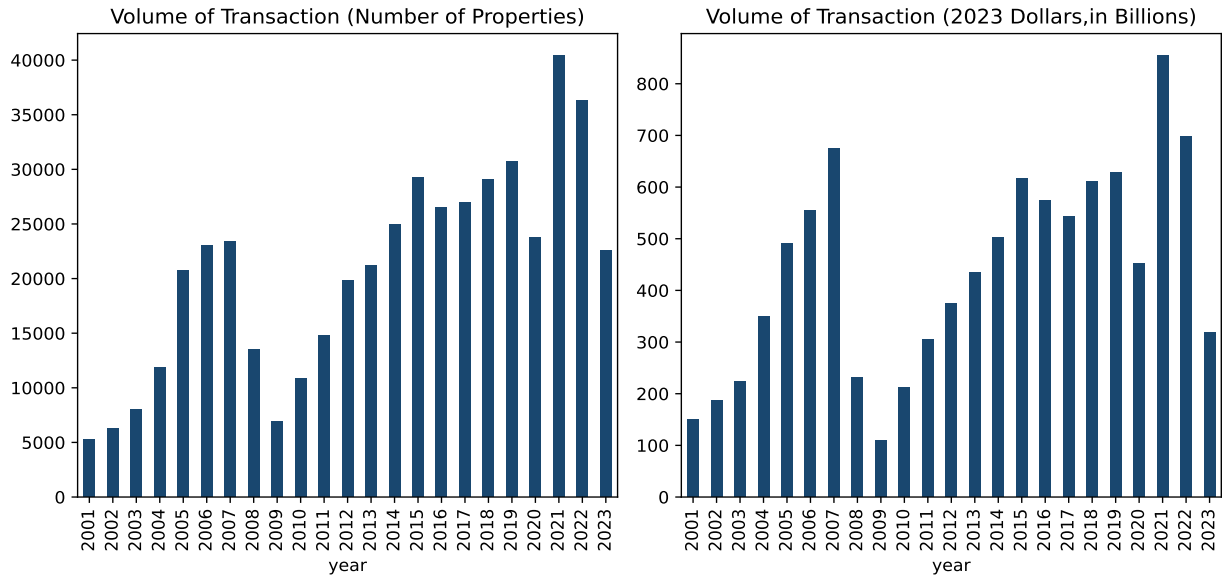
## 4.1 Summary Statistics of Transactions

Figure 1 shows the number of transactions (left) and the dollar volume of transactions (right) in our final sample. All dollar values in the paper are expressed in 2023 real dollars. Henceforth, we use the word volume to describe the dollar value of transactions. The graph shows three boom periods, from 2001 until 2007, from 2009 until 2019, and in 2021-22. It also shows three busts, the Great Financial Crisis in 2008-09, the Covid-19 recession in 2020, and the interest rate hiking cycle in 2022-23. Appendix Table A.3 shows transaction volume by year in table format. There are additional transaction dynamics that are sector specific, such as the decline in retail due to the rise in e-commerce and the decline in office due to the rise in working-from-home in the post-2020 period (Gupta et al., 2025). Figures A.1 and A.2 show transaction volume by sector and property subtype, respectively.

Property transactions are about evenly split between six size categories: >$250M, $100-250M, $50-100M, $25-50M, $10-25M, < $10M. Table A.4 shows the breakdown. Most investors typically stay within one of

[20] This is a difficult task since CRE properties are usually bought using a separate LLC for each building, typically with a non-descriptive name. The RCA team has spent years unraveling who is behind each LLC, often cross-checking each transaction using multiple sources.

Figure 1: Commercial Real Estate Transactions




*Notes:* The figure shows the aggregate number of transactions across all property sectors, i.e. Apartments, Office, Industrial and Retail, by year between 2001 and 2023. The left panel shows the transaction volume in terms of the number of properties traded while the right panel shows the total dollar value of properties traded, expressed in billions of 2023 U.S. dollars.

these size buckets.

Table A.5 reports transaction volume for the 16 largest markets, as well as all other 44 markets combined. Not surprisingly, Manhattan is the largest property market in the sample, followed by Los Angeles. There are large differences in the importance of each sector across markets. Office accounts for over 60% of volume in Manhattan, Washington DC, and San Francisco, but for only 34% of volume in Los Angeles and 24% in Dallas. About half of total transaction volume (and 40% of the number of transactions) is concentrated in the largest 16 markets.

About 75% of transactions accounting for 70% of volume are conventional sales, where one property changes hands. About 19% of transactions and 21% of volume are portfolio sales, where a group of properties is sold together. The remainder of the transactions are entity transactions, where an entire corporate entity changes hand, or distressed sales (foreclosures or debtor/trustee sales). In the case of portfolio and entity sales, RCA imputes the price of each property that is part of the transaction. We further improve on RCA's imputation as explained in Appendix A.1. Table A.6 shows the transaction breakdown by transaction type.

## 4.2 Summary Statistics By Investor Characteristics

Next, we turn to investor characteristics. Table A.7 shows how we map RCA's investor types to a smaller group of eight investor types. Our breakdown creates meaningful heterogeneity in terms of constraints, investment objectives, geographic reach, et cetera. Table 1 shows the transaction breakdown by investor type. About 600 Real Estate Private Equity funds (REPE) account for 10-12% of volume. About 3,400 Institutional investors, made up of investment managers, pension funds, insurance, banks and finance companies, and open-ended funds, make up about 14% of volume. The two largest investment categories are local and national Owner-Operator-Developers ($OOD_L$ and $OOD_N$, respectively), accounting for about 15% and 30% of volume, respectively. These are typically smaller, more geographically-concentrated, private firms that are fully dedicated to commercial real estate. They tend to have substantial discretion in terms of investments, but are subject to tighter financial constraints. We define local OODs as those we see transact in only one market, and national OODs as those we see transact in more than one market. Individuals are made up of high-net worth buyers as well as non-traded REITs who serve high-net worth households, and account for about 3-4% of volume. REITS are publicly-listed commercial real estate vehicles and account for 12% of volume. We include real estate operating companies and listed funds in this category as well. As listed entities, REITs are subject to much stricter oversight from the regulatory and financial analyst communities, and face difficulties raising additional equity when their share price trades below net asset value. Foreign investors include sovereign wealth funds and foreign OODs, but also the foreign investors of all other investor types. They represent about 6-8% of volume.[21] Users consist of corporations, governments, non-profits, educational, and religious institutions and charities that own their own real estate. They account for about 5% of transactions. For about 1-2% of volume we do not know the investor type, and create an Unknown category.

Table A.8 shows that 87% of transactions by count and 73% by volume involve a single buyer and a single seller. A further 10% by transactions and 20% by volume involve one buyer and two sellers or two buyers and one seller. The remainder has multiple buyers and multiple sellers where we observe up to four buyers and up to four sellers. Table A.9 provides additional detail on these Joint Ventures, in terms of which investor types tend to collaborate with each other.

Investors are also heterogeneous in terms of their portfolio size, the size of the typical asset they trade, and their geographic, sector, and asset quality concentration. Figure A.3 shows the distribution of investor portfolio sizes at the end of 2023. The left panel of Figure 2 shows that $OOD_L$ predominantly hold smaller properties. As property size increases, the share of $OOD_L$ transactions diminishes, giving way to a more

[21] Table A.7 shows that there are some foreign entities among most investor types. For example, about 33% of the pension fund buys and 21% of pension fund sells are by foreign pension funds. Our foreign investor flag keeps track of this.

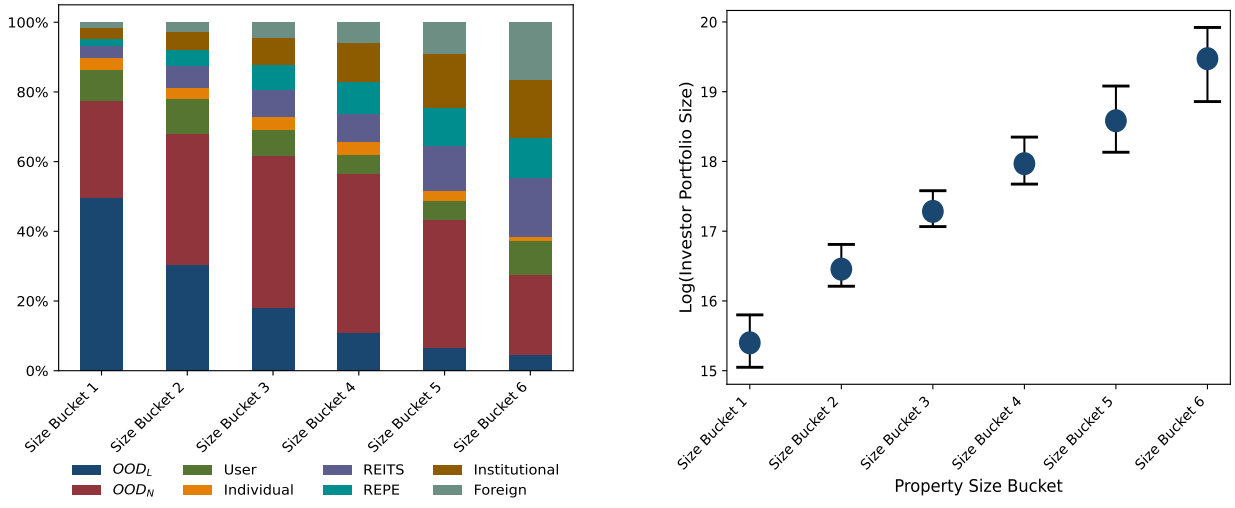
Table 1: Transactions By Investor Type

| | Buyer (#Trans) | Buyer ($ Vol) | Buyer (% Vol) | Seller (#Trans) | Seller ($ Vol) | Seller (% Vol) | Unique Investors |
|---|---|---|---|---|---|---|---|
| REPE | 28,853 | 1241 | 12.30 | 22,058 | 1031 | 10.22 | 596 |
| Institutional | 38,066 | 1479 | 14.66 | 38,148 | 1371 | 13.59 | 3,435 |
| $OOD_L$ | 147,030 | 1316 | 13.05 | 161,967 | 1656 | 16.41 | 238,140 |
| $OOD_N$ | 150,678 | 3083 | 30.56 | 131,966 | 3081 | 30.54 | 25,385 |
| Individual | 19,453 | 406 | 4.02 | 19,399 | 285 | 2.82 | 15,811 |
| REITS | 33,518 | 1182 | 11.72 | 35,135 | 1254 | 12.43 | 389 |
| Foreign | 17,606 | 844 | 8.37 | 13,055 | 616 | 6.11 | 2,782 |
| User | 28,771 | 418 | 4.14 | 33,663 | 554 | 5.49 | 29,845 |
| Unknown | 12,044 | 119 | 1.18 | 20,627 | 241 | 2.39 | 7,802 |
| Total | 476,018 | 10,088 | 100 | 476,018 | 10,088 | 100 | 324,185 |

*Notes:* The table shows the number and dollar volume of buy and sell transactions for various investor types, i.e REPE (Real Estate Private Equity), Institutional Investors, $OOD_L$ (Local Developers), $OOD_N$ (National Developers), Individual investors, REITs, Foreign Investors and Users from 2001 to 2023. Volume is expressed in billions of 2023 U.S. dollars. The table also reports the the share of buy and sell volume, which sums to 100% across investor types.

diverse mix of investor types in higher size buckets. In contrast, Foreign investors, Institutions, and REITs become increasingly prominent in the larger size buckets. This could indicate a preference or ability to invest in larger properties. The right panel of Figure 2 illustrates the average log of investor portfolio size by property size bucket, with error bars showing the inter-quartile range. A clear upward trend in log portfolio size is evident, suggesting that investors transacting larger buildings (by dollar value) tend to have larger portfolio sizes. This could reflect economies of scale or better access to capital for larger investors.

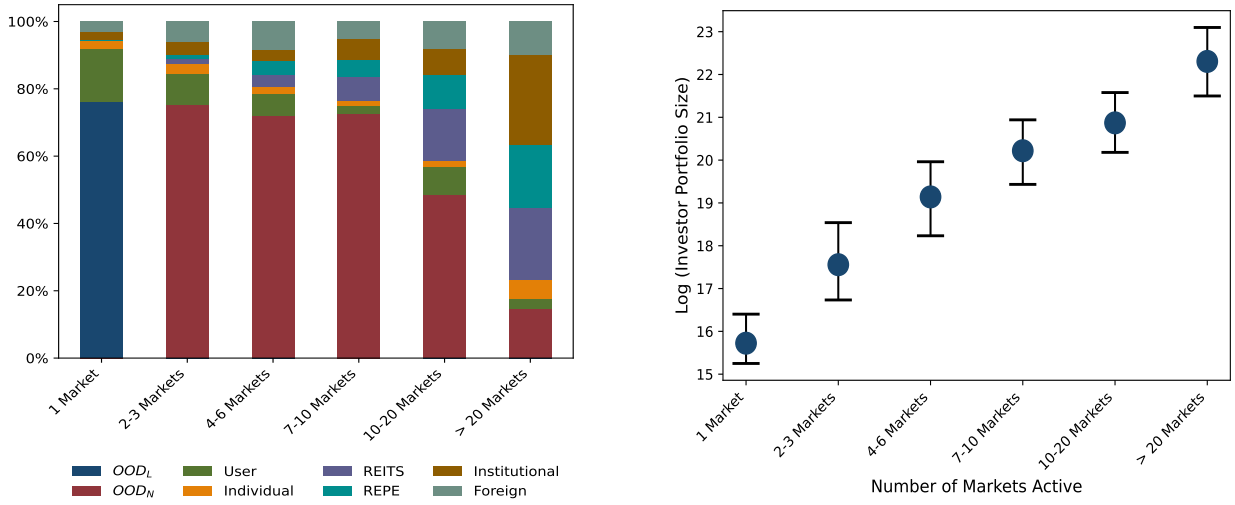
Figure 2: Size Concentration




*Notes:* This figure shows investor concentration by the aforementioned property size buckets at the end of 2023 for investors that own at least one property. The left figure shows the distribution of investor types by the size buckets while the right figure shows average log investor size (and inter-quartile range) of average log(investor portfolio size) for each property size bucket.

The left panel of Figure 3 summarizes the distribution of investor types by the number of markets the investor is active in. Investors of type $OOD_L$ are active in a single market by definition. Many Users also only show up in a single market. REPE, REITs, and institutional investors, on the other hand, are more likely to hold geographically-diversified portfolios. The right panel shows a strong positive correlation between investor size and geographic diversification. Overall, these facts suggest that there is a substantial amount of market segmentation among the investors in our sample.

Figure 3: Geographic Concentration




*Notes:* This figure shows investor concentration by geography at the end of 2023 for investors that own atleast one property[17]. The left figure shows the distribution of investor types by the number of aggregated RCA markets they are active in while the right figure shows average log investor size (and inter-quartile range) of average log(investor portfolio size) by the number of markets the investor is active in.

## 4.3 Flow Dynamics

Identifying buyers and sellers at the transaction level, and aggregating their buy and sell trades, allows us to analyze aggregate capital flows by investor type. Figure 4 shows net flows for select investor types by sector and year. It offers insights into how capital was allocated across CRE sectors over time.

One interesting trend is the large inflow of foreign investors from 2013 onwards. These inflows totaled $317 billion, with a significant concentration in Office ($123 billion, 39%) and Apartments ($106 billion, 33%). This surge included landmark deals such as Canada's Brookfield Property Partners' $18 billion acquisition of the REIT GGP in 2018 and its $11.4 billion acquisition of the REIT Forest City that same year. Singapore's sovereign wealth fund GIC made a major push into the US industrial market by acquiring IndCor Properties from the REPE fund Blackstone for $8.1 billion in 2015. Figure A.4 further reveals that most of this foreign investment in Offices and Apartments was concentrated in superstar cities, primarily

Figure 4: Net Flows by Investor Types




*Notes:* This figure shows the net acquisitions of Real Estate Private Equity (REPE), REITs, Institutional, and Foreign investors, calculated as the dollar value of purchases minus sales (in 2023 dollars) for each sector and year. The measure captures transaction activity while excluding changes in property values.

Manhattan, Los Angeles, San Francisco, and Boston. These patterns provide a compelling motivation to study the price and portfolio effects of such foreign capital inflows in Section 7.

Another notable pattern is the high volatility in REIT flows, particularly in the lead-up to the 2008 financial crisis. Several REITs aggressively offloaded properties in 2006 and 2007. Notable examples include Sam Zell's $39 billion entity sale of his Equity Office Properties REIT to Blackstone in 2007 and the $22 billion purchase of Archstone-Smith REIT by Tishman Speyer and Lehman Brothers. These major transactions, occurring on the eve of the Great Financial Crisis, provide a natural setting for counterfactual analyses, helping us examine how market outcomes might have differed had these REITs not liquidated their holdings.

Finally, we document trading patterns among types in Table A.10. We see significant trade within and across investor types. Notably, a fourth of all transaction volume involves buyers and sellers of the same type, reflecting segmentation and shared investment strategies among similar investor groups. There is substantial trading activity between local and national *OOD*s. Additionally, REITs and Users exhibit a higher propensity to trade within their own investor type, indicating a degree of segmentation in trading

preferences among these groups.

# 5 Estimation of the Valuation Model

We estimate the valuation model (6) from observed transaction prices. We impose that investors with the same features have the same valuation for an asset regardless of whether they are buyers or sellers.[22] The left-hand side variable is the log real price per square foot for sectors O, I, and R and the log price per unit for sector A.

## 5.1 The Linear Hedonic Valuation Model

To establish a reference point, we first consider the standard valuation model in the literature: the linear hedonic model (LHM) which assumes that log price per square foot or per apartment unit is linear in property characteristics. This model ignores the role of investor characteristics in valuation.

For each sector, the first column of Table 2 shows the results for this standard LHM including only asset characteristics: log asset size, occupancy rate, log age, log renovation-adjusted age, subtype indicator, number of floors, CBD indicator,[23] superstar city indicator,[24] entity sale indicator, and distressed sale indicator. All property characteristics enter with the expected sign and most "hedonic prices" are precisely estimated. The deal variables show that transactions that are part of an entity sale trade at a premium whereas properties that are part of a distressed deal trade at a large discount. We add year-fixed effects to capture dynamics in the investment opportunity set such as the interest rate environment or the overall liquidity of the market. This model explains 46% of the variation in apartment prices, 48% for industrial, 36% for offices, and 52% in retail prices.

In the next step, we enrich this model with the local macroeconomic variables: market occupancy rate, market NOI growth, log real per capita personal income, and log population for A and log employment for O, I, R. These macro variables vary by market and year, and the former two also by sector. We also add the NER variable, which is either available at the property level or at a fine geographic level, as explained above. The NER variable enters highly statistically significantly in all sectors, as does personal income. We add market fixed effects to control for any other omitted variables that differ by market.[25] The second column

[22] Relaxing this assumption is possible but leads to parameter proliferation without much gain in explanatory power.

[23] Since the office subtypes are CBD and suburban office, the CBD indicator is co-linear with the subtype category CBD, and we drop the CBD indicator from the model for Office. For Apartments, the subtype indicator is 1 for Mid/Highrise and 0 for Garden, for Industrial, it is 1 for Flex and 0 for warehouse and for Retail properties, it is 1 for shops and 0 for centers.

[24] Superstar cities are defined as those markets with office prices above $350 per square foot in 2023. The superstar dummy is one for the following markets (for all sectors): Washington DC, DC VA burbs, DC MD burbs, Chicago, Los Angeles, Orange Co, Manhattan, NYC Boroughs, San Jose, San Francisco, East Bay, Seattle, Boston, Austin, Miami/Dade Co, and San Diego.

[25] Market fixed effects render the superstar cities indicator redundant.

for each sector of Table 2 shows the results for this augmented linear hedonic model. The $R^2$ now range from 48% for office to 59.2% for Apartments.

This model is the hedonic model against which we wish to assess the role of investor covariates. We note that this is a much richer hedonic model in terms of covariates than the traditional hedonic model. In addition, market and time fixed effects could pick up liquidity fluctuations driven by investor composition. By including them in the hedonic model, we make it harder for investor characteristics to matter.

Table 2: Reference Point: Linear Hedonic Model

| | Apartment | | Industrial | | Office | | Retail | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| CBD Indicator | 0.132 | 0.165** | 0.461*** | 0.288*** | 0.279*** | 0.105** | 0.413*** | 0.269*** |
| | (0.133) | (0.060) | (0.071) | (0.060) | (0.096) | (0.049) | (0.094) | (0.052) |
| Age | -0.072*** | -0.078*** | 0.012 | 0.001 | -0.017* | -0.030*** | 0.011 | -0.004 |
| | (0.011) | (0.009) | (0.009) | (0.007) | (0.009) | (0.007) | (0.007) | (0.006) |
| Renovation Adj. Age | -0.039*** | -0.033*** | -0.101*** | -0.092*** | -0.094*** | -0.078*** | -0.105*** | -0.104*** |
| | (0.013) | (0.007) | (0.011) | (0.010) | (0.012) | (0.011) | (0.009) | (0.009) |
| Property Size | -0.162*** | -0.091*** | -0.347*** | -0.266*** | -0.264*** | -0.221*** | -0.402*** | -0.369*** |
| | (0.027) | (0.021) | (0.018) | (0.015) | (0.022) | (0.021) | (0.016) | (0.014) |
| Property Subtype | 0.150*** | 0.145*** | 0.187*** | 0.129*** | | | -0.000 | 0.027 |
| | (0.046) | (0.027) | (0.028) | (0.020) | | | (0.024) | (0.025) |
| No. of Floors | 0.158*** | 0.120*** | 0.083** | 0.023 | 0.144*** | 0.097*** | 0.197*** | 0.059*** |
| | (0.023) | (0.017) | (0.030) | (0.021) | (0.017) | (0.010) | (0.024) | (0.020) |
| Occupancy Rate | 0.827*** | 0.709*** | 0.029 | 0.162*** | 0.916*** | 0.835*** | 0.572*** | 0.568*** |
| | (0.153) | (0.140) | (0.073) | (0.038) | (0.070) | (0.052) | (0.067) | (0.055) |
| Superstarcity | 0.626*** | | 0.434*** | | 0.530*** | | 0.409*** | |
| | (0.099) | | (0.119) | | (0.079) | | (0.075) | |
| Entity Sale | 0.277** | 0.220** | 0.127 | 0.118 | 0.152 | 0.140 | -0.001 | 0.044 |
| | (0.106) | (0.091) | (0.107) | (0.094) | (0.098) | (0.090) | (0.119) | (0.109) |
| Distressed | -0.282*** | -0.229*** | -0.288*** | -0.226*** | -0.349*** | -0.292*** | -0.309*** | -0.276*** |
| | (0.040) | (0.029) | (0.047) | (0.032) | (0.043) | (0.033) | (0.046) | (0.040) |
| Market Occupancy | | 0.094 | | -0.140 | | 0.217* | | 0.006 |
| | | (0.306) | | (0.091) | | (0.117) | | (0.091) |
| NOI Growth | | 0.073 | | 0.013 | | 0.027 | | -0.069 |
| | | (0.093) | | (0.046) | | (0.033) | | (0.043) |
| Personal Income | | 0.603*** | | 0.288*** | | 0.359*** | | 0.431*** |
| | | (0.071) | | (0.066) | | (0.052) | | (0.031) |
| Population/Employment | | 0.025* | | 0.019 | | -0.027** | | 0.038*** |
| | | (0.013) | | (0.013) | | (0.012) | | (0.006) |
| NER | | 0.043*** | | 0.224*** | | 0.446*** | | 0.173*** |
| | | (0.007) | | (0.033) | | (0.062) | | (0.033) |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Market FE | × | ✓ | × | ✓ | × | ✓ | × | ✓ |
| Observations | 141,135 | 141,135 | 116,737 | 116,737 | 96,139 | 96,139 | 114,223 | 114,223 |
| Adj. $R^2$ | 46.0 | 59.2 | 47.9 | 58.5 | 35.9 | 48.0 | 52.4 | 58.4 |

*Notes:* This table shows the full-sample coefficient estimates and goodness of fit statistics for a linear hedonic valuation model. Standard errors in bracket are clustered by RCA Market and Year.

## 5.2 LGBM model Without Investor Characteristics

Next, as an interim step, we allow for the augmented set of hedonics, i.e., the asset and deal characteristics, market characteristics, year-fixed effects, and market fixed effects to affect valuations in a non-linear way.

We also allow for all these features to interact with one another. Importantly, we still leave out investor characteristics.

We estimate a Light Gradient Boosted Machine (LGBM) model, which provides a flexible, data-driven and scalable approach compared to traditional parametric models. LGBM is well-suited for this estimation as it efficiently handles large datasets, captures complex non-linear relationships without requiring manual specification, and is robust to overfitting through built-in regularization.[26] We estimate the valuation model for each sector and model specification separately, calibrating hyper-parameters in each case. We standardize continuous variables to facilitate comparison across sectors.[27,28]

The first four rows of the top panel of Table 3 show the results. The first row shows the full-sample $R^2$ from the LGBM model estimated on just the 10 asset and deal characteristics. The second row adds the 5 macro variables including NER. The third row adds the year-fixed effects, and the fourth row the market-fixed effects. We refer to the model in row 4 as the hedonic LGBM model. The hedonic LGBM model produces a $R^2$ that ranges from 65% for Office to 75% for Apartments.

This $R^2$ in row 4 is directly comparable to the $R^2$ in the second column of Table 2 for each sector. The difference between the two sets of $R^2$ statistics tells us how important non-linearities and interaction effects are. The improvement from the added flexibility in the functional form between hedonics and valuations boosts the $R^2$ by 16% points for Apartments, 12% points for Industrial, 17% points for Office, and 14% points for Retail. The substantial improvement in fit from non-linearities and interaction effects is a first main empirical result, and is new in the context of CRE valuation. Below, we analyze which interaction effects and non-linearities matter most.

## 5.3 The Main Valuation Model

We are now ready to add investor characteristics. Again, we proceed in steps. In the fifth row of Table 3, we add investor type-fixed effects, leaving out the indicator for the Unknown category: REPE, Institutional, $OOD_L$, $OOD_N$, Individual, REIT, Foreign, User. In the sixth and final row, we add the log portfolio value of the investor, the share of the total portfolio value that the investor has in superstar markets, in the same market as the focal building, in the same sector as the focal building, as well as the joint venture dummy. These portfolio variables are computed just prior to the transaction for the sellers and just after the transaction for the buyers, so that the transacted asset is always included for both the seller and the buyer.

[26] The hyperparameters are optimized using 10-fold cross-validation to minimize out-of-sample mean squared error. The optimization procedure searches over key parameters, including tree depth, the number of leaves, feature and bagging fractions, and regularization terms to ensure that the model effectively balances flexibility and generalization. Details on the tuning process and parameter search space are provided in Appendix B.2.

[27] Standardizing variables is not a requirement for tree-based models such as LGBM which make splits based on the relative ordering of feature values rather than their absolute magnitudes.

[28] We make use of the flexibility of LightGBM to incorporate categorical variables as predictors directly.

Table 3: LGBM Valuation Model

**In-Sample performance**

| Sector | Apartment | | Industrial | | Office | | Retail | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | R2C | $R^2$ | R2C | $R^2$ | R2C | $R^2$ | R2C |
| Hedonic Model | 54.14 | | 53.94 | | 50.07 | | 63.15 | |
| + Macro Vars | 71.26 | | 67.18 | | 62.29 | | 68.66 | |
| + Year Dummies | 72.34 | | 68.1 | | 61.92 | | 70.05 | |
| + Market Dummies | 75.3 | | 70.13 | | 64.93 | | 72.01 | |
| + Investor Types | 78.91 | 14.62*** | 74.67 | 15.2*** | 69.87 | 14.09*** | 76.56 | 16.26*** |
| + Portfolio Vars | 85.57 | 41.58*** | 85.95 | 52.96*** | 84.46 | 55.69*** | 87.52 | 55.41*** |

**Cross-Validated Out-of-Sample performance**

| Sector | Apartment | | Industrial | | Office | | Retail | |
|---|---|---|---|---|---|---|---|---|
| | $R^2$ | R2C | $R^2$ | R2C | $R^2$ | R2C | $R^2$ | R2C |
| Hedonic Model | 49.56 | | 50.39 | | 47.11 | | 60.0 | |
| + Macro Vars | 67.14 | | 62.93 | | 58.37 | | 65.09 | |
| + Year Dummies | 68.21 | | 63.62 | | 58.53 | | 66.13 | |
| + Market Dummies | 70.91 | | 66.09 | | 60.57 | | 67.65 | |
| + Investor Types | 76.34 | 18.67*** | 72.3 | 18.31*** | 67.08 | 16.51*** | 73.83 | 19.1*** |
| + Portfolio Vars | 83.35 | 42.76*** | 83.63 | 51.73*** | 80.94 | 51.66*** | 85.24 | 54.37*** |

*Notes:* This table summarizes the performance $R^2$ from valuation model estimations in equation (6) using LightGBM. The two panels summarize the performances of our non-linear valuation model by successively adding variables as described in 5. Each row chooses the optimal LightGBM hyperparameters using cross validation. The top panel shows in-sample $R^2$, and the bottom panel shows out-of-sample $R^2$.

We compute the comparative R-squared (R2C) for the models in rows 5 and 6, relative to the model without investor characteristics in row 4. The R2C is the ratio of the difference in R-squared values between the alternative and null models to the unexplained variation in the dependent variable by the null model:

$$R2C = \frac{R^2_{alt} - R^2_{null}}{1 - R^2_{null}}$$

Kim (2024) develops the asymptotic distribution of R2C and a test of the null hypothesis that R2C equals zero. The stars on the R2C statistic reflect the p-values of this test.

Row 5 of Table 3 shows that the investor categories provide useful additional information for valuations. Relative to row 4, the $R^2$ goes up between 3.6 and 4.9% points, reducing the unexplained variation in model 4 by between 14 and 16% (R2C).

The gains from adding the other investor characteristics in row 6 are substantially larger. This model in row 6 is our main valuation model. The $R^2$ reach values of 84.5% points for Office, 85.6% for Apartments, 86.0% for Industrial, and 87.5% points for Retail. Comparing row 6 to row 4, investor variables increase the $R^2$ by 10.3% points (A), 15.8% points (I), 19.5% points (O), and 15.5% points (R). They reduce the

unexplained variation in model 4 by 41.6% for A, 53.0% for I, 55.7% for O, and 55.4% for R. These R2C gains are highly statistically significant in all four sectors and economically large.

In sum, investor characteristics are a crucial "new hedonic" in commercial real estate valuation models that strongly add to the valuation model's explanatory power. When different investors value asset characteristics differently, it matters for the transaction price which investors are in the market at a given point in time.

The second panel of Table 3 shows the out-of-sample (OOS) performance of the LGBM model. For the latter, we use a 10-fold cross-validation procedure where the $R^2$ is calculated as the average of the hold-out samples across all folds. Naturally, the main model in row 6 has a lower OOS than in-sample $R^2$. However, both the absolute level of the $R^2$ and the gain in explanatory power relative to the hedonic model in row 4 remain large. Investor characteristics explain an additional 12.4% (A), 17.5% (I), 20.4% (O), and 17.6% (R) points of the variation in property transaction prices, reducing unexplained variation between 42.8% and 54.4%. This gain remains highly statistically significant in all four sectors.

The valuation model also allows us to compute the residual standard deviation for buyers and sellers in each year. We allow for time variation but impose: $\sigma^2_{b,t} = \sigma^2_{s,t}$. We estimate this using $\sigma^2_{b,t} = 2\sigma^2_{p,t}$.

Table 2

## 5.4 Variable Importance Analysis

The previous analysis showed that (i) investor characteristics matter and (ii) that allowing for non-linearities and interaction effects matters. Our main model allows for traditional hedonics to be valued differently by different types or sizes of investors, for example. We systematically investigate the importance of each feature using SHAP values (Lundberg and Lee, 2017).[29]

Figure 5 highlights the most important features selected by the model, organized by category. Investor portfolio characteristics are the most important feature category, ahead of property characteristics and macro variables. Appendix Figure B.2 breaks out the features further. Among all variables, investor size

[29]SHAP (Shapley Additive Explanations) values provide a game-theoretic approach to interpreting machine learning models by quantifying the marginal contribution of each predictor to the model's predictions. For a given observation, SHAP values measure how much each feature pushes the model's output away from the mean prediction. Mathematically, for a prediction function $f(x)$ and a set of features $N$, the SHAP value for feature $j$ is calculated as:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} (f(S \cup \{j\}) - f(S)),$$

where $M$ is the total number of features and $f(S)$ represents the model output when only the features in subset $S$ are known. Feature Importance ($I_j$) is obtained by averaging its absolute SHAP value over all n observations.

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right|$$

SHAP values provide a more comprehensive importance measure by accounting for both direct and interaction effects, unlike traditional importance measures such as split frequency.

Figure 5: Valuation Model Feature Importance by Category




*Notes:* Figure shows the importance of features selected by the LightGBM Model. Feature importance is defined as the absolute SHAP (Shapley Additive Explanations) values for each feature summed across all observations. We aggregate the importance by feature types and normalize them such that the importance for each feature group sums to 1

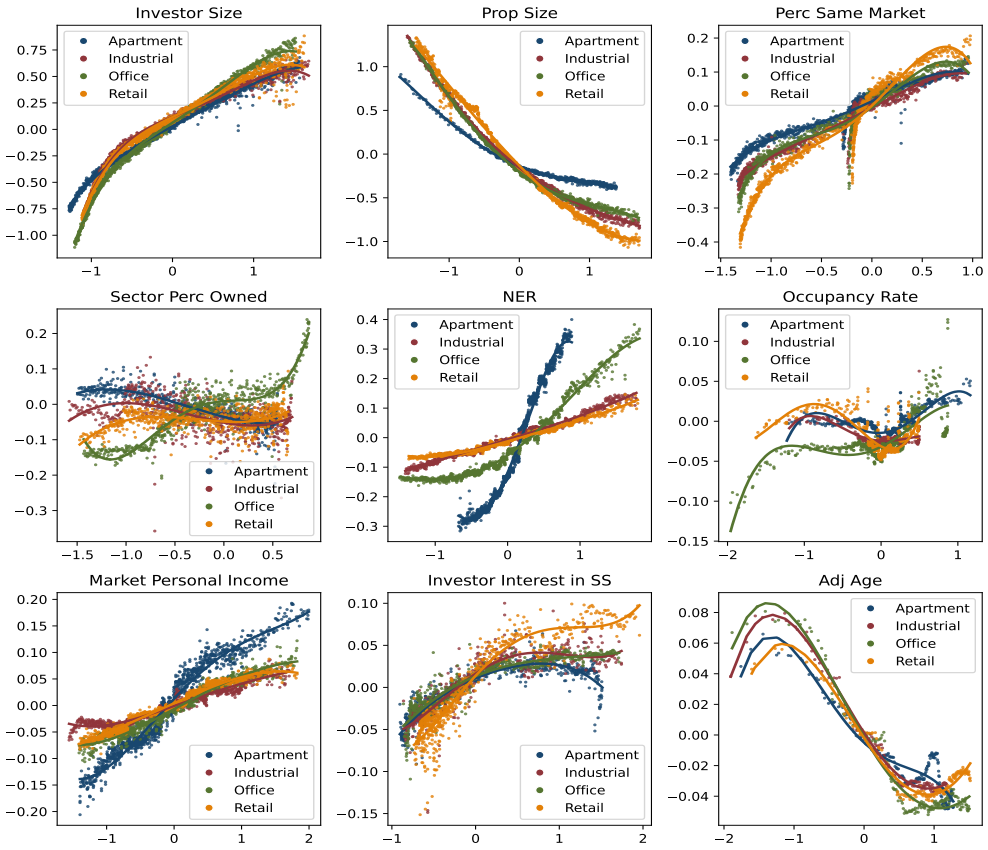
and property size are the two most important features in all sectors. The share of properties the investor has in the same market, the share of properties the investor has in the same sector, and the investor type are all among the ten most important features. NER is the most important macro variable and market fixed effects are important as well.

To analyze how transaction prices vary with predictors, we use SHAP dependence plots. For each feature, we compute the average SHAP value at different percentiles of the feature (from the 5th to the 95th percentile) and plot these against the corresponding feature values using a binned scatter plot. The X-axis represents the standardized feature value, while the Y-axis represents the standardized SHAP value averaged across the buyer and seller valuation models, ensuring comparability across sectors. Positive (negative) SHAP values indicate that a higher feature value increases (lowers) the price. Figure 6 shows the results for the nine most important continuous features, reported in the title of each panel.

Valuations are concave in investor size but convex in property size. Investors who specialize in the office sector have higher valuations for office, while sector focus matters much less for valuations in other sectors. A high NER (expensive neighborhood with high-quality assets) affects office and apartment valuations disproportionately. The graph illustrates that many of the effects are nonlinear in ways that would be difficult to uncover without the help of a ML model like LGBM.

To study time trends in valuations, we plot the SHAP values for the year fixed effects in Figure B.3. Valuations declined sharply in the GFC and recovered steadily after 2012 with industrial and apartments showing the strongest recovery. Retail valuations begin falling post 2016—coinciding with the retail apocalypse and rise of E-Commerce. All sectors saw their valuations fall in 2023 on the back of higher interest

Figure 6: SHAP Dependence: Non-linearities




*Notes:* This figure shows SHAP dependence plots for the nine most important continuous features in our valuation models. Each plot is a binned scatter plot with a best-fit polynomial trend line. The X-axis represents the standardized feature value, while the Y-axis represents the standardized SHAP value, averaged across the buyer and seller valuation models, to allow comparability across sectors.

rates.

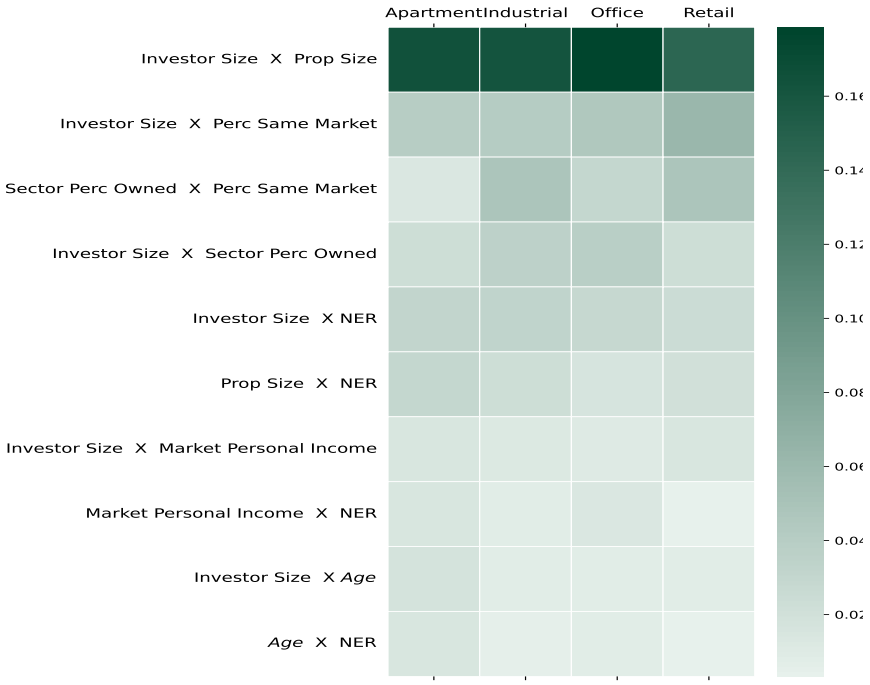
SHAP values capture both direct and interaction effects, making them ideal for analyzing feature interactions. Beyond attributing predictions to individual features, SHAP can be extended to decompose contributions into pairwise interactions. Following (Lundberg et al., 2020), we compute SHAP interaction values for all pairwise interactions.[30] Figure 7 shows the 10 most important interaction effects for each sector. Interactions between investor characteristics, notably investor portfolio size, and building features such as property size, NER, and building age are among the most influential. There are also several interactions of two investor characteristics in the top ten. This result reinforces the critical role of investor

[30]The SHAP interaction values quantify how a feature pair jointly influences predictions beyond their individual effects. For a pair of features $i$ and $j$ these values are defined as:

$$\phi_{ij} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M-1)!} \left[ f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \right]$$

where the term inside the square brackets captures the marginal effect of including both features $i$ and $j$ together compared to including them individually.

Figure 7: Top Ten Interaction effects




*Notes:* This figure presents the ten most important interaction effects, measured by averaging the absolute SHAP interaction values across all observations for each feature pair within a sector, averaged across buyer and seller valuation models. To facilitate comparability, values are normalized so that the sum across all feature pairs equals 1 for each sector.

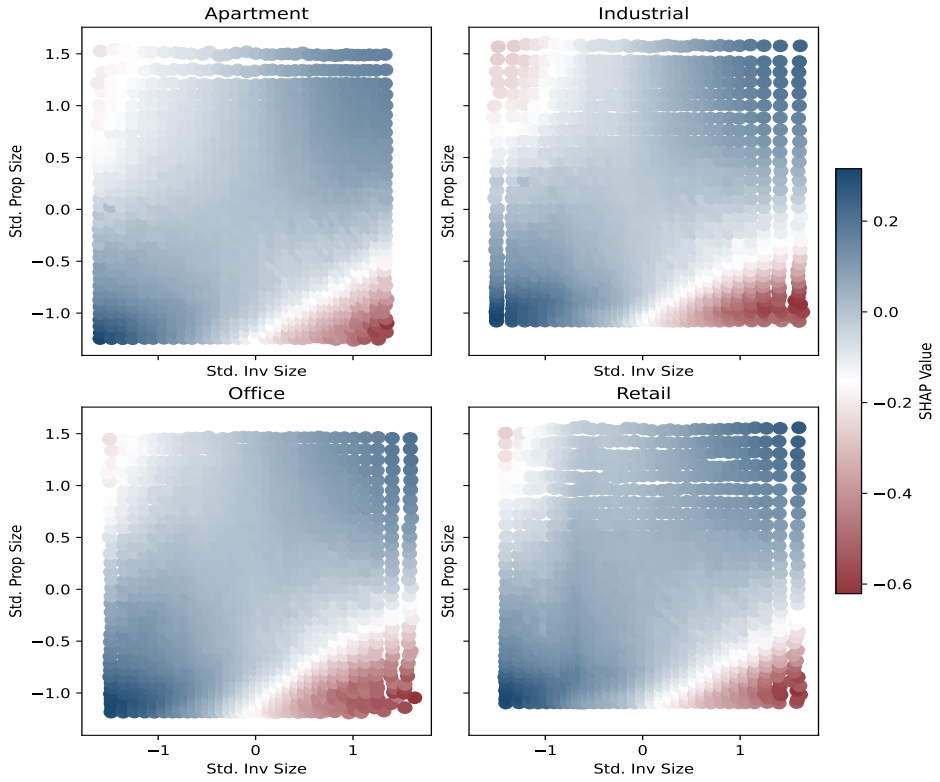
characteristics in the valuation model, as well as an important mechanism: different investors value different hedonics differentially so it matters which investors are in the market at a given point in time.

Figure 8 studies the interaction effects between investor size and property size, the most important interaction effect identified in Figure 7. For the pair, we compute the average interaction SHAP value at different percentiles of the feature (from the 5th to the 95th percentile) and plot these against the corresponding feature values using a binned scatter plot. The X-axis represents the standardized investor size, while the Y-axis represents the standardized property size. The color intensity reflects the magnitude and direction of the interaction SHAP values, where deep red indicates a negative interaction effect and blue represents a positive interaction effect.[31] A key observation is that large investors have lower valuations for small properties and small investors have dampened valuations for large properties in every sector.

To further examine interaction effects, we use SHAP values at the individual observation level to assess how the marginal effect of one feature varies across different levels of another. Since SHAP values decompose model predictions into contributions from each feature, we can analyze how the effect of a given feature depends on another feature. Appendix Figure B.4 analyzes how the impact of investor size on valuations

[31] Unlike the main SHAP values, which include both direct and interaction effects, this plot isolates only the interaction component, which measures how the effect of one variable changes depending on the level of the other variable.

Figure 8: SHAP Interaction Effects




*Notes:* This figure shows a binscatter plot to understand the interaction effects between investor size and property size, capturing how their combined influence deviates from the sum of their individual contributions. The color intensity reflects the magnitude and direction of the interaction SHAP values, where deep red indicates a negative interaction effect, while blue represents a positive interaction effect.

depends on other features such as asset size, investor type, or market. For example, the higher valuation that large investors have for office and retail properties is most pronounced for the largest assets, less pronounced for REITS, and is stronger in superstar cities than in smaller markets.

Figure B.5 explores interactions of NER with investor characteristics. Higher neighborhood quality (NER) is associated with higher valuations on average. The quality preference in office and retail markets is particularly strong for foreign, large, sector-focused, and geographically-focused investors.

## 5.5 Addressing overfitting concerns

To address concerns that the gains in explanatory power from the LGBM model reflect overfitting rather than genuine signal, we build in conservatism in our hyperparameter choices, including maximum tree depth of 5, as discussed in Appendix B. Table B.1 shows the output of a linear valuation model that adds five key investor characteristics and five key interaction effects, identified from the LGBM model. This table confirms the importance of investor characteristics and non-linearities. These variables boost the explanatory power

of the model and are estimated precisely. Figure B.1 shows successive model fits, starting from the simplest linear hedonic specification and adding asset features, and then investor type and portfolio features. We then augment the linear model with key nonlinearities and interaction terms identified from the SHAP analysis. Across these exercises, the performance patterns are consistent: investor characteristics, their interactions with property features, and selected nonlinearities materially improve predictive accuracy. The good fit of the simpler linear models suggest that the improvements stem from meaningful economic structure captured by investor characteristics and their interactions, rather than spurious fit.

# 6 Estimation of the Listing and Meeting Model

## 6.1 Point Estimates in the Meeting Model

Next, we estimate the parameters of the meeting model in (4). The variables that determine whether a potential buyer $b$ meets a building that has been listed for sale by a seller $s$ are: the buyer's log portfolio size $S_b$, an indicator $N_b$ which is one if the buyer owns more than two buildings, the log difference between the buyer and the seller's portfolio sizes $\Delta S_{b,s}^{-1}$, an indicator $T_{b,s}$ which is one if the buyer and the seller have traded with each other before, and four consideration set variables in $\delta_{b,s}$ which measure how similar the building for sale is to the assets in the buyer's portfolio in terms of asset size, sector, location, and NER. We estimate one model for each sector-year using all transactions from that sector-year. Table 4 reports the coefficient point estimates $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ that minimize the loss function $-\sum_s \ln \pi_r(b,s)$, with $\pi_r$ as defined in (7), and their standard errors. The parameter estimates $\hat{\lambda}_t$ are averaged over time $\hat{\lambda} = (T)^{-1} \sum_{t=1}^{T} \hat{\lambda}_t$. The standard errors for $\hat{\lambda}_t$ are obtained using bootstrap and then also averaged over time.

With the coefficient estimates for the valuation model in hand, we can compute the private valuation gap between buyers and sellers $\Delta_{b,s}^h$ and obtain $\pi_\tau$, the probability whether a transaction takes place conditional on a meeting. We estimate the models separately for each sector and year from 2002 to 2023, using all transactions in that year.[32]

Since most investors in our sample transact infrequently, we restrict the negative sample to those investors who have transacted at least once in the preceding five years.[33] This still leaves open the possibility that some potential buyers arrive in the transaction year with zero assets. In those cases, we assume that they own the median number of properties (which is 1) with characteristics equal to the average portfolio in that year. Since the asset listed for sale is included in the portfolio of the actual buyer, we also add it to the

[32] While the valuation model is estimated over 2001-2023, we begin estimating the meeting model in 2002 to ensure that all covariates, including the prior buyer–seller transaction variable $T_{b,s}$, are available.

[33] We explore robustness to this assumption in Appendix C.2.

portfolio of all other potential buyers to avoid a bias. Not doing so would make it artificially easy for the algorithm to select the true buyer from the set of all potential buyers.

As explained above, rather than including all potential buyers in the likelihood estimation, we use an important insight from the NLP literature showing it is sufficient to consider a negative set with small cardinality $K$. We explore values of $K$ between 50 and 1,500 and choose $K = 1,000$ for our main results, since parameter estimates stabilize around that value. We discuss robustness of our results to alternative values for $K$ in Appendix C.1.

Table 4: Meeting Model Parameter Estimates

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ |
|---|---|---|---|---|---|---|---|---|
| Apartment | 1.54 | 2.32 | 2.45 | 2.95 | 8.06 | 5.47 | 4.31 | 7.25 |
| | (0.03) | (0.08) | (0.17) | (0.29) | (0.24) | (0.15) | (0.15) | (0.32) |
| Industrial | 1.66 | 2.20 | 2.82 | 2.48 | 8.33 | 5.77 | 4.12 | 9.90 |
| | (0.04) | (0.09) | (0.19) | (0.4) | (0.31) | (0.17) | (0.15) | (0.50) |
| Office | 1.58 | 2.38 | 2.51 | 2.97 | 8.08 | 5.63 | 3.4 | 10.34 |
| | (0.04) | (0.08) | (0.19) | (0.39) | (0.29) | (0.17) | (0.16) | (0.54) |
| Retail | 1.52 | 2.24 | 2.5 | 3.01 | 8.17 | 5.44 | 3.98 | 7.92 |
| | (0.04) | (0.09) | (0.18) | (0.38) | (0.30) | (0.17) | (0.14) | (0.39) |

*Notes:* This table reports the average estimated values of the meeting model parameters along with their standard errors in bracket. The point estimates $\hat{\lambda}_t$ are obtained by maximum likelihood for each year, and averaged across years. Standard errors are calculated using bootstrap (with 1,000 samples) for each year, and averaged across years. (See Appendix C.1 for parameter estimates by year) The maximum likelihood estimation minimizes the loss function in (7). $\lambda_1$ is the coefficient of investor size, $\lambda_2$ is the coefficient of the indicator that equals one when the buyer owns more than 2 buildings, $\lambda_3$ is the coefficient on buyer-seller size difference. $\lambda_4$ is the coefficient of the indicator that equals one when the buyer and seller have a prior transactional relationship. $\lambda_{5,1}$ through $\lambda_{5,4}$ are the coefficients of asset, market, sector and NER alignment.

The first column of Table 4 shows that the size of the buyer has a large positive effect on the meeting probability. A potential buyer which is 1% larger than another has about a 1.5% higher likelihood of meeting the seller. This parameter is identified from the fact that large investors, for example large REITS or REPE funds, are engaged in many more transactions than small investors. The positive and significant $\lambda_2$ suggests that owning more than two assets makes a buyer about 2.3% points more likely to meet the property for sale.

The positive and significant $\lambda_3$ estimate in the third column shows that buyers are more likely to meet sellers with a similar portfolio size. This is consistent with the positive assortative matching between investor size and asset size shown in the right panel of Figure 2. A positive $\lambda_4$ shows that a buyer who has transacted with the focal seller before is about 3% more likely to meet that seller than a buyer who has not traded with the seller before.

The consideration set variables (the vector $\lambda_5$ in columns 5–8) all have the expected positive sign, are large in magnitude, and are estimated precisely. Buyers are much more likely to meet assets that have similar

size, location, sector, and quality (NER) to the assets they already own. Property size alignment ($\lambda_{5,1}$) and asset quality alignment ($\lambda_{5,4}$) are the most important similarity measures. To understand the coefficient magnitudes, a potential buyers of apartments that is 1% more invested in the same market is 5.47% more likely to meet the seller. A 1% increase in the alignment of NER increases the chance of a buyer-seller meeting by 10.34% for offices. The consideration set variables suggest that specialization is an important feature of CRE portfolios.

## 6.2 Variable Importance in the Meeting Model

To better understand the importance of each of the variables in the meeting model, Table 5 successively leaves out each of the variables, reporting the resulting coefficient estimates and log likelihood. The first row of each panel is the baseline model, repeated for convenience. The eighth column shows $1 - \pi_{m_0}\prime/\pi_{m_0}$, the percentage reduction in the probability that the true buyer meets the seller when we exclude a specific variable from the estimation.

We find that buyer size is the most important determinant of the meeting probability. Excluding it from the estimation results in a 56-64% drop in the probability that the algorithm identifies the correct buyer across sectors. The likelihood of the restricted model is about 100 log points lower than that of the baseline model across all four sectors. Excluding buyer size leads to large changes (biases) in the remaining coefficient estimates. The second size measure, owning at least two properties, increases the probability of identifying the right buyer by 10–19%.

Size similarity between buyers and sellers is relatively less important. Excluding it lowers the probability of identifying the correct buyer by about 14–24%. Prior transactional relationships between buyer and seller also matter: excluding this variable reduces the probability of identifying the true buyer by 5–10%. Its overall impact is muted by the fact that such relationships are relatively rare.[34]

The consideration set variables are all quantitatively important, with some variation across sectors. The most important one in every sector is asset location similarity ($\lambda_{5,2}$). Excluding it lowers the the probability of identifying the correct buyer by 40–52% and the likelihood by 60–75%. Asset size similarity is the second most important consideration set variable in every section, with a 34-42% reduction in likelihood from excluding it.

Figure 9 displays the likelihood that the model identifies the true buyer from among the set of 1,000 potential buyers and the true buyer. That likelihood is consistently high, averaging over 20%. For comparison, random search would result in a 0.1% probability of identifying the true buyer. Directed search therefore delivers a 200x gain over random search. This underscores the model's ability to capture the key

[34] Empirically, about 8% of actual buyer–seller pairs in our sample have a prior transaction.

Table 5: Meeting Model Analysis

Apartments

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ | 1 - $\pi_{m0'}$ /$\pi_{m0}$ | Avg LL |
|---|---|---|---|---|---|---|---|---|---|---|
| All $\lambda$s | 1.54 | 2.32 | 2.45 | 2.95 | 8.06 | 5.47 | 4.31 | 7.25 | | -3.22 |
| $\lambda$s Excluding $\lambda_1$ | | 3.63 | 1.98 | 4.90 | 5.04 | 2.89 | 4.18 | 6.79 | 56.0% | -4.20 |
| $\lambda$s Excluding $\lambda_2$ | 1.90 | | 2.31 | 3.10 | 8.57 | 5.14 | 4.64 | 6.58 | 14.3% | -3.47 |
| $\lambda$s Excluding $\lambda_3$ | 1.46 | 2.19 | | 3.41 | 8.55 | 5.35 | 4.23 | 7.36 | 14.1% | -3.34 |
| $\lambda$s Excluding $\lambda_4$ | 1.58 | 2.33 | 2.62 | | 8.15 | 5.49 | 4.35 | 7.17 | 9.6% | -3.29 |
| $\lambda$s Excluding $\lambda_{5,1}$ | 1.07 | 2.26 | 3.52 | 3.57 | | 4.90 | 4.39 | 7.78 | 39.6% | -3.81 |
| $\lambda$s Excluding $\lambda_{5,2}$ | 0.99 | 1.51 | 2.15 | 3.05 | 7.01 | | 4.91 | 8.51 | 53.7% | -3.92 |
| $\lambda$s Excluding $\lambda_{5,3}$ | 1.35 | 2.24 | 2.38 | 3.41 | 7.69 | 6.03 | | 7.75 | 32.2% | -3.74 |
| $\lambda$s Excluding $\lambda_{5,4}$ | 1.53 | 2.03 | 2.64 | 2.93 | 8.36 | 5.93 | 4.52 | | 18.4% | -3.49 |

Industrial

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ | 1 - $\pi_{m0'}$ /$\pi_{m0}$ | Avg LL |
|---|---|---|---|---|---|---|---|---|---|---|
| All $\lambda$s | 1.66 | 2.20 | 2.82 | 2.48 | 8.33 | 5.77 | 4.12 | 9.90 | | -2.97 |
| $\lambda$s Excluding $\lambda_1$ | | 3.42 | 2.65 | 5.17 | 2.80 | 2.93 | 3.75 | 9.53 | 64.3% | -4.28 |
| $\lambda$s Excluding $\lambda_2$ | 1.91 | | 2.58 | 2.52 | 8.67 | 5.28 | 3.81 | 9.64 | 11.7% | -3.15 |
| $\lambda$s Excluding $\lambda_3$ | 1.59 | 1.92 | | 3.15 | 8.88 | 5.66 | 4.05 | 10.15 | 23.6% | -3.13 |
| $\lambda$s Excluding $\lambda_4$ | 1.69 | 2.19 | 2.95 | | 8.39 | 5.79 | 4.14 | 9.72 | 5.1% | -3.01 |
| $\lambda$s Excluding $\lambda_{5,1}$ | 1.16 | 2.07 | 3.75 | 2.94 | | 5.44 | 4.53 | 10.30 | 33.5% | -3.47 |
| $\lambda$s Excluding $\lambda_{5,2}$ | 1.24 | 1.04 | 2.82 | 2.69 | 7.97 | | 4.81 | 12.41 | 45.5% | -3.61 |
| $\lambda$s Excluding $\lambda_{5,3}$ | 1.61 | 1.43 | 2.84 | 2.79 | 8.67 | 6.26 | | 10.65 | 30.6% | -3.43 |
| $\lambda$s Excluding $\lambda_{5,4}$ | 1.66 | 1.94 | 3.15 | 2.35 | 8.60 | 6.51 | 4.43 | | 24.6% | -3.32 |

Office

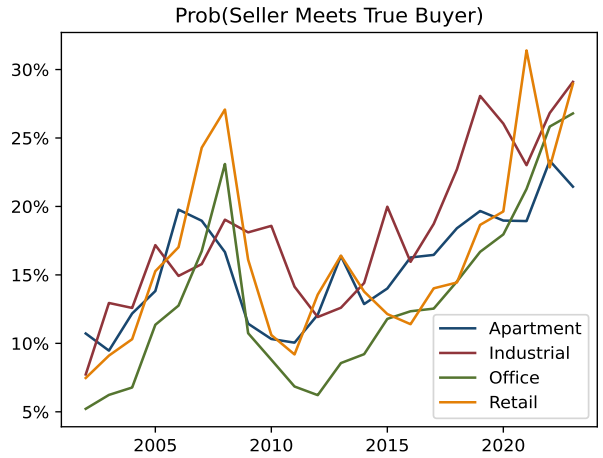
| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ | 1 - $\pi_{m0'}$ /$\pi_{m0}$ | Avg LL |
|---|---|---|---|---|---|---|---|---|---|---|
| All $\lambda$s | 1.58 | 2.38 | 2.51 | 2.97 | 8.08 | 5.63 | 3.40 | 10.34 | | -3.45 |
| $\lambda$s Excluding $\lambda_1$ | | 3.42 | 2.36 | 5.02 | 4.25 | 2.99 | 2.29 | 10.39 | 59.5% | -4.50 |
| $\lambda$s Excluding $\lambda_2$ | 1.85 | | 2.37 | 3.10 | 8.18 | 5.14 | 2.96 | 10.03 | 18.5% | -3.71 |
| $\lambda$s Excluding $\lambda_3$ | 1.52 | 2.22 | | 3.50 | 8.65 | 5.54 | 3.39 | 10.52 | 20.1% | -3.58 |
| $\lambda$s Excluding $\lambda_4$ | 1.61 | 2.39 | 2.65 | | 8.14 | 5.64 | 3.44 | 10.18 | 8.3% | -3.50 |
| $\lambda$s Excluding $\lambda_{5,1}$ | 1.06 | 2.09 | 3.53 | 3.55 | | 5.19 | 2.96 | 11.39 | 42.3% | -4.02 |
| $\lambda$s Excluding $\lambda_{5,2}$ | 1.08 | 1.35 | 2.34 | 2.99 | 7.28 | | 3.78 | 12.30 | 52.4% | -4.11 |
| $\lambda$s Excluding $\lambda_{5,3}$ | 1.39 | 1.87 | 2.52 | 3.21 | 7.58 | 5.95 | | 11.01 | 28.8% | -3.74 |
| $\lambda$s Excluding $\lambda_{5,4}$ | 1.57 | 2.17 | 2.74 | 2.88 | 8.58 | 6.18 | 3.70 | | 28.9% | -3.80 |

Retail

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ | 1 - $\pi_{m0'}$ /$\pi_{m0}$ | Avg LL |
|---|---|---|---|---|---|---|---|---|---|---|
| All $\lambda$s | 1.52 | 2.24 | 2.50 | 3.01 | 8.17 | 5.44 | 3.98 | 7.92 | | -3.12 |
| $\lambda$s Excluding $\lambda_1$ | | 3.42 | 2.43 | 5.11 | 3.67 | 2.73 | 3.86 | 7.37 | 60.8% | -4.18 |
| $\lambda$s Excluding $\lambda_2$ | 1.78 | | 2.31 | 3.12 | 8.46 | 4.80 | 3.74 | 7.59 | 10.3% | -3.32 |
| $\lambda$s Excluding $\lambda_3$ | 1.47 | 2.02 | | 3.48 | 8.59 | 5.34 | 3.90 | 8.04 | 20.0% | -3.25 |
| $\lambda$s Excluding $\lambda_4$ | 1.55 | 2.23 | 2.63 | | 8.24 | 5.44 | 4.03 | 7.79 | 7.0% | -3.17 |
| $\lambda$s Excluding $\lambda_{5,1}$ | 1.04 | 2.07 | 3.19 | 3.52 | | 5.08 | 4.37 | 8.27 | 37.6% | -3.63 |
| $\lambda$s Excluding $\lambda_{5,2}$ | 1.11 | 1.09 | 2.41 | 3.01 | 7.80 | | 4.44 | 9.82 | 40.5% | -3.68 |
| $\lambda$s Excluding $\lambda_{5,3}$ | 1.46 | 1.62 | 2.38 | 3.54 | 8.78 | 5.66 | | 8.27 | 34.6% | -3.61 |
| $\lambda$s Excluding $\lambda_{5,4}$ | 1.51 | 1.98 | 2.77 | 2.90 | 8.44 | 6.07 | 4.21 | | 22.8% | -3.44 |

*Notes:* Figure shows the point estimates of the meeting model for each of the four sectors. The last column reports the log-likelihood. The first row in each panel is the baseline model. The successive rows leave out one of the variables in the meeting model, each of which is a special case of the baseline model.

Figure 9: Meeting Probabilities




*Notes:* This figure shows the probability of meeting the actual buyer from among a set of $K = 1,000$ potential buyers (negative set) and the actual buyer, as implied by the benchmark model estimation.

determinants of buyer-seller matches in CRE markets. A drop in meeting probabilities during the financial crisis is evident across all sectors. This may reflect the heightened uncertainty and reduced liquidity in the market, disrupting the typical alignment between buyer preferences and assets available for sale. The graph also shows that the model's ability to identify the correct buyer rises from about 10% at the start of the sample to nearly 30% at the end of the sample. While the estimation is done year-by-year, the number of annual transactions in the data set is growing over time. Also, with every passing year of additional transaction data, we obtain a clearer picture of the investors' portfolios as holdings have to be built from past transactions.

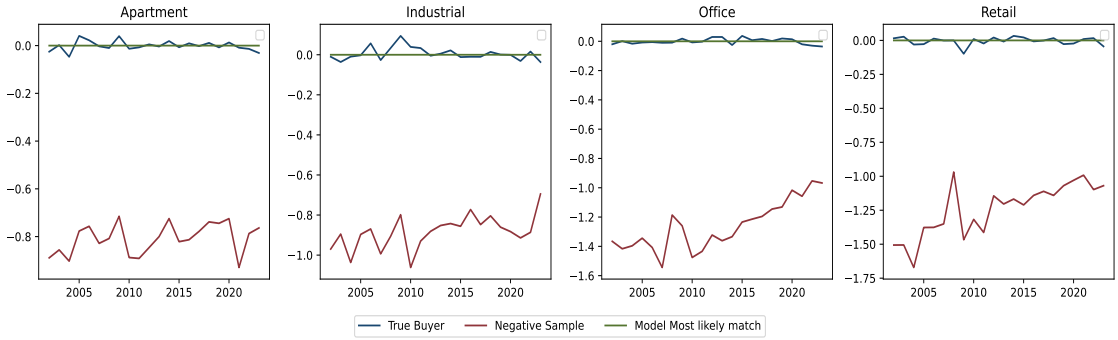
Identifying the correct buyer is a high hurdle. After all, in a pool of 1,000 potential buyers, there are usually several buyers with characteristics that are very similar to the true buyer, making exact identification difficult. What matters in practice is whether the model ranks the true buyer near the top. Figure C.2 shows that the true buyer appears among the top 1%, 2.5%, or 5% of predicted matches with high probability ($\sim$ 70% - 80%), highlighting the model's ability to zoom in on the relevant subset of potential buyers.

## 6.3 Transaction Probability

Having estimated the probability of a meeting, we next turn to the probability that a transaction takes place conditional on a meeting, $\pi_\tau$. The latter depends on the valuation gap between the buyer and the seller, and hence combine the estimates from the meeting and valuation models. Figure 10 plots the valuation gap $\Delta_{b,s}$ for the positive set (the actual buyer) and the negative set (the average among the $K$ potential buyers).

Across all sectors, the true buyers exhibit consistently higher valuation gaps than the other potential buyers. This reflects not only the model's ability to select the correct buyer, as noted above, but also reinforces the model's assumption that matches occur when buyers perceive gains from trade. True buyers have about a 3–8% higher valuation for the asset than the seller. For the negative sample (the non-buyers), the valuation gap is large and negative (between -80% and -150% on a log scale).

Figure 10: Valuation Gap



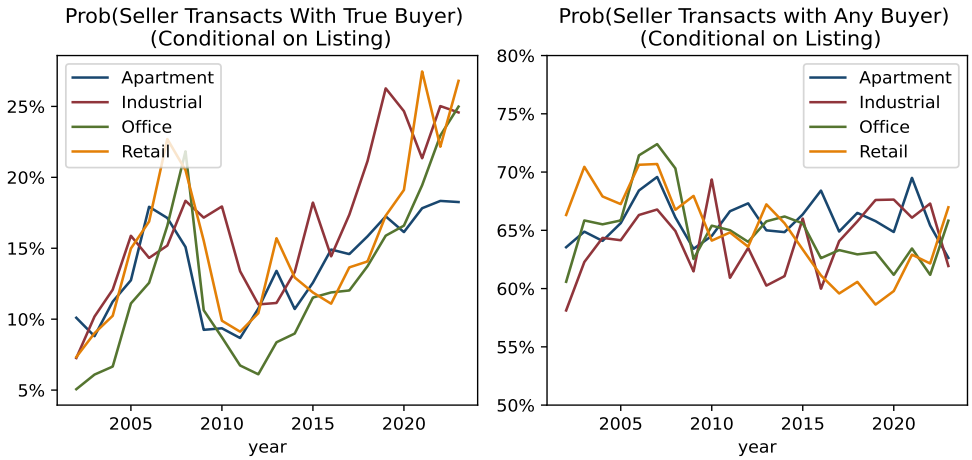

*Notes:* This figure shows the valuation gaps, i.e. difference between the log valuation of the buyer and the log valuation of the seller for the positive sample (the actual buyer) and the negative sample (the set of K - 1 = 999 potential buyers)

The left panel of Figure 11 shows the model-implied probability that the actual buyer transacts with the seller. This probability is consistently high, reflecting the model's ability to accurately capture directed search (identify the correct buyer) as well as the relative valuations of the buyer and the seller for the asset (the valuation model delivers a positive probability of sale conditional on a meeting). The right panel shows the model-implied probability of a transaction occurring with any buyer. These probabilities are 60-70%. They show that the model has a high likelihood of predicting a trade for assets that actually traded.

## 6.4 Listing Probability

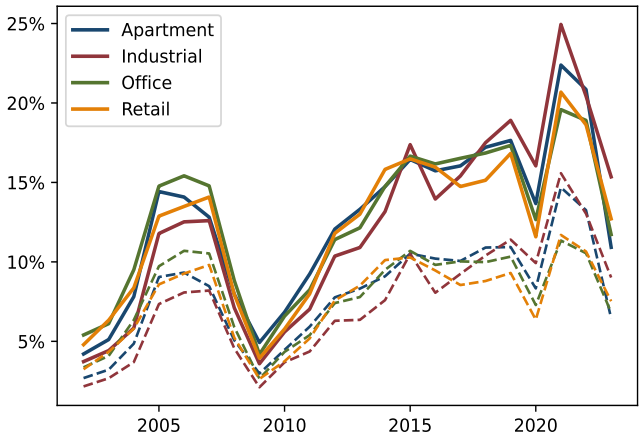
Finally, Figure 12 shows the listing probability, defined in equation (9). The listing probability reconciles the model-implied transaction probabilities with the observed transaction volumes in the data for each sector-year. For all sectors, the estimation results in a boom-bust pattern in listing probabilities in the 2001–2009 period and a rise in listing probabilities in the 2010–2019 period. Listing probabilities dip during the Covid year 2020 and again in 2022-23 when the Fed embarks on its interest rate hiking cycle. The model generates realistic average listing probabilities of around 15-20% per year, as inferred from actual transaction probabilities that around 7-10% per year (dotted lines).

Figure 11: Transaction Probabilities




*Notes:* This figure shows the model-implied probabilities of transactions. The left panel displays the probability that the seller transacts with the actual buyer, combining the meeting and transaction probabilities. The right panel shows the probability that the seller transacts with any buyer, reflecting aggregate transaction likelihoods across all potential buyers.

Figure 12: Listing Probabilities




*Notes:* This figure shows the listing probabilities $\pi_l$ for the benchmark model over time across all sectors (Apartment, Industrial, Office, and Retail), as defined in (9). Dotted lines, show for each sector the fraction of available properties that did transact in that year. The gap between the two lines being the likelihood that the seller listed the property, but failed to find a suitable buyer.

# 7 Applications: Prediction and Counterfactuals

Having estimated the valuation and matching models, we now turn to three applications of the model. In Section 7.1, we use the model can be used to predict the transaction price of a building, out-of-sample. In Section 7.2, we show how the model can be used to understand the dynamics of CRE prices, focusing

in particular on the period leading up to the 2008-09 financial crisis. We conclude in Section 7.3 with an application where we use the model to uncover investors' substitution patterns across building size, sector, and location. This analysis is an important input to understanding the impact on the CRE ecosystem of, for instance, an inflow of foreign buyers in the high-quality segment of the Manhattan office market. All three applications make use of the potential price distribution concept explained in Section 3.3, which in turn relies on both the valuation model and the transaction model.

## 7.1 Predicting transaction prices out-of-sample

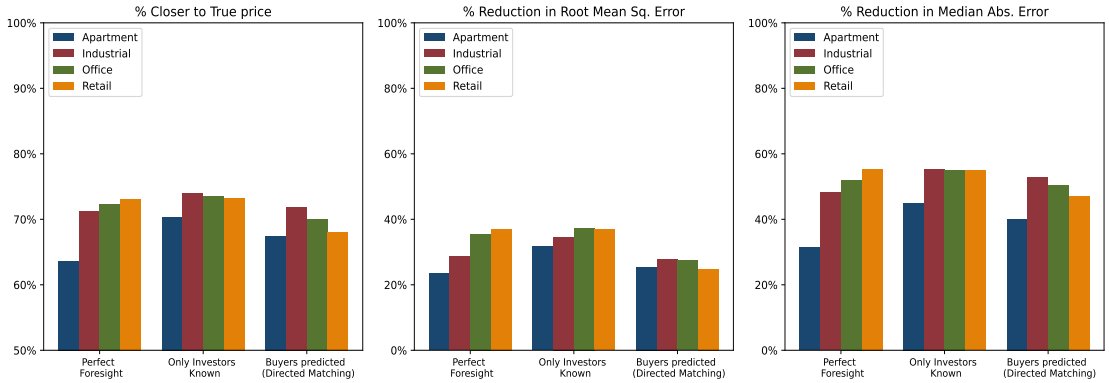
For an owner looking to sell their property in the next period, a key issue is to predict the transaction price and assess the price risk. Our CRE model is well positioned to answer both questions. The two main insights from the valuation model are that building, market, and investor characteristics all matter for observed transaction prices, and that it is important for accurate valuations to capture the nonlinear and interaction effects of those variables on prices. We now explore the model's ability to predict transaction prices out-of-sample for buildings that we know transacted.

Standing at time $t$, predicting prices $p_{t+1}$ for buildings that trade at time $t+1$ in the data amounts to (i) estimating the model based on data available until time $t$, and (ii) predicting the inputs in the valuation model $(x_{n,t+1}, z_{t+1}; \gamma_{t+1})$. To understand which inputs are most important, we proceed in three steps.

In the first step, we assume that the building, market, and investor characteristics can be predicted with perfect accuracy. We estimate the model using information up to time $t$ and predict the prices in $t+1$ using the observed inputs $(x_{t+1}, z_{t+1}; \gamma_{t+1})$. Figure 13 reports gains in OOS predictability of our baseline LGBM model relative to the linear hedonic model reported in Table 2, a natural benchmark in CRE valuation. The left panel shows the fraction of transactions for which the LGBM is closer to the true transaction price than the linear hedonic model. The middle panel shows the percent gains in terms of root mean squared error (RMSE), and the right panel reports the gains in median absolute pricing error (MAE). The first set of bars in each plot (labeled "perfect foresight") refer to the case where inputs are assumed to be known. The figure shows that the LGBM strongly outperforms the linear hedonic model out-of-sample. The LGBM model's forecast is closer to the true observed price than the linear hedonic model in about 60-75% of transactions across years and sectors. In terms of the magnitude of the pricing error, the LGBM model's predicted price is typically 20-40% closer to the true price than the linear hedonic model's prediction (RMSE), while MAE gains indicate a 30-55% error reduction rate across sectors.

In the second set of bars in Figure 13, we assume that the time-$(t+1)$ building and market characteristics and time fixed effects are unknown. We use their time-$t$ values as estimates; i.e., we assume that the macro

Figure 13: Comparing Out of Sample performance




*Notes:* This figure shows the relative out-of-sample performance of the LGBM model relative to the linear hedonic model (LHM). The left chart reports the fraction of transactions for which the LGBM prediction is closer to the true transaction price than the LHM prediction. The middle chart reports the percent reduction in root mean squared prediction error of the LGBM relative to the LHM model. The right chart reports the percent reduction in median absolute prediction error of the LGBM relative to the LHM model. While the prediction model is always estimated with information up until time $t$ in order to predict the price at $t+1$, the rows differ in how we determine the covariates $(x_{t+1}, z_{t+1})$ that enter in the prediction of the time-$t+1$ price. The first set of bars in each plot assumes that $(x_{t+1}, z_{t+1})$ are known. The middle group assumes that $z_{t+1}$ is known and that $x_{t+1}$ must be predicted. The right groups of bars assumes that $(x_{t+1}, z_{t+1})$ must both be predicted, and we use our directed matching model to identify the potential buyers.

variables and building occupancy rate follow a random walk: $E_t[x_{n,t+1}] = x_{n,t}$. We continue to assume that the buyer characteristics $z_{t+1}$ are known. The LGBM model continues to produce predictions that are closer to the true transaction price than the linear hedonic model's predictions in about 70% of transactions. The gains in RMSE and MAE are higher for Apartments and Industrial and comparable for Office and Retail when hedonics have to be predicted than when they are known.

In the third set of bars, we assume that all of the building, market, and investor characteristics need to be predicted. For the building and market characteristics and time fixed effects, we continue to use the time-$t$ values. For the buyer information, we use our directed matching model. Specifically, for each property that transacts, the forecaster who stands at time $t$ knows the identity of the seller, but not that of the buyer who materializes at time $t+1$, $z_{b,t+1}$. We draw $C = 1,000$ potential buyers in proportion with their likelihood according to our meeting model, estimated only on data available as of time $t$. We also do not know the ex-post characteristics of the potential buyers and seller and so we use their time-$t$ characteristics. For each potential buyer, we compute the price based on that buyer's fundamental valuation $h_b$ and a shock, and check whether the buyer's value exceeds that of the seller. This is similar to the process we use to obtain the potential price distribution (see Section 3.3), except here, we obtain potential prices for a given building $n$ and computing statistics across the $C$ potential buyer draws. The implementation details are in Appendix A.4. The third row of Figure 13 shows that the LGBM generates better out-of-sample predictions than the

linear hedonic model in about 65% of transactions. The average reduction in pricing errors relative to the linear hedonic model is about 25% in terms of RMSE and around 40% in terms of MAE. Figure D.1 shows that the gains in prediction power are consistent across years, while Figure D.2 decomposes the total out-of-sample performance gains into three components: the improvement from moving from the linear hedonic model (LHM) to the non-linear hedonic LGBM, the additional gain from adding investor characteristics to the LGBM, and the advantage of using our directed matching model for predicting buyers over a random matching benchmark. All three steps contribute meaningfully to the model's out-of-sample performance.

Even when investor identities and characteristics must be predicted, our joint valuation and transaction model delivers significant improvements in pricing accuracy fully out-of-sample. These results should lend credibility to the counterfactual exercises in the next two sections.

## 7.2 Understanding CRE price dynamics

A key insight of our analysis is that investors have heterogeneous valuations for properties. We now explore an application of the CRE model in which we quantify the role of heterogeneous valuations.

Figure 4 showed that a large number of offices and apartments changed hands from REITS to REPE funds and Institutions in the lead-up to the Great Financial Crisis. As our model indicates, some of these transactions may have reflected differences in beliefs between these types of investors about expected returns or differences in risk tolerance. But some may have arisen because REPE funds had strong fundraising years in 2005–2007 and may have been compelled to execute deals. To understand the impact of this heterogeneity across investors on Office and Apartment prices in 2006 and 2007, we compute the counterfactual experiment in the potential buyer pool for REIT holdings is different. Specifically, we explore what would have happened had REPE and Institutional investors not been around to buy these properties.

We follow the procedure outlined in Section 3.3 to compute potential price distributions. For every office and apartment building owned by a REIT in 2006 and 2007, we simulate $C = 1,000$ potential buyers from the model's meeting probability distribution.[35] If the buyer's valuation exceeds the seller's, the building transacts, and we record the price. We compute the fraction of assets that trade ($\tau$) and the size-weighted average price among the assets that transact.

Table 6 presents the results. In the first row of each panel, we show the fraction of REIT-owned properties that trade and the average price per unit (square feet for office, units for apartments) conditional on trade in the data, as well as the empirical composition of buyers for the properties that were sold by REITS. In the second row, we show the same statistics from a model simulation where all investors are present in the

[35]It is important to define the universe as the properties owned by REITs rather than those sold by REITS to avoid selection bias in which assets were transacted.

potential buyer pool. This baseline model predicts similar transaction probabilities than in the data for office and somewhat higher transaction probabilities for apartments than in the data. Prices conditional on sale are similar in model and data for 2006 offices and apartments, and lower in the model than in the data in 2007.[36] This benchmark simulation sets a baseline for studying the impact of removing specific groups of buyers.

Table 6: CRE Price Dynamics in the REIT sell-off of 2006-2007

2006 Offices (3,889 Buildings)

| Model | $\tau$ (%) | ppsf | Major Buyers (% Value of properties bought) |
|---|---|---|---|
| Truth | 22.0 | 261.7 | [REPE: 31.5, Institutional: 22.6, Foreign: 15.0, REITS: 14.1] |
| Benchmark (No Exclusions) | 25.2 | 249.1 | [Institutional: 27.7, REITS: 26.1, REPE: 17.4, $OOD_N$: 9.0] |
| Exclude REPE and Institutional | 20.1 | 232.4 | [REITS: 43.8, $OOD_N$: 22.8, Foreign: 12.9, $OOD_L$: 7.4] |
| Exclude REPE, Institutional and REITS | 14.3 | 222.9 | [$OOD_N$: 43.1, Foreign: 17.7, $OOD_L$: 14.0, Individual: 12.9] |

2007 Offices (3,997 Buildings)

| Model | $\tau$ (%) | ppsf | Major Buyers (% Value of properties bought) |
|---|---|---|---|
| Truth | 28.7 | 382.5 | [REPE: 60.9, Institutional: 14.9, $OOD_N$: 10.5, REITS: 10.3] |
| Benchmark (No Exclusions) | 26.7 | 306.5 | [Institutional: 27.6, REITS: 21.2, $OOD_N$: 17.9, REPE: 15.5] |
| Exclude REPE and Institutional | 20.3 | 236.8 | [REITS: 40.7, $OOD_N$: 27.2, Foreign: 11.3, Individual: 10.3] |
| Exclude REPE, Institutional and REITS | 15.3 | 228.5 | [$OOD_N$: 46.5, Individual: 18.8, Foreign: 16.1, User: 9.9] |

2006 Apartments (3,560 Buildings)

| Model | $\tau$ (%) | ppu | Major Buyers (% Value of properties bought) |
|---|---|---|---|
| Truth | 16.4 | 103,386 | [$OOD_N$: 39.2, Institutional: 32.6, REITS: 8.5, Foreign: 8.3] |
| Benchmark (No Exclusions) | 31.8 | 118,557 | [REITS: 38.7, $OOD_N$: 28.2, Institutional: 21.5, REPE: 5.5] |
| Exclude REPE and Institutional | 28.3 | 116,109 | [REITS: 54.1, $OOD_N$: 37.6] |
| Exclude REPE, Institutional and REITS | 20.4 | 101,150 | [$OOD_N$: 82.2, $OOD_L$: 9.2, Foreign: 5.1] |

2007 Apartments (3,375 Buildings)

| Model | $\tau$ (%) | ppu | Major Buyers (% Value of properties bought) |
|---|---|---|---|
| Truth | 18.8 | 200,547 | [REPE: 38.4, Institutional: 38.3, $OOD_N$: 15.0, Foreign: 5.1] |
| Benchmark (No Exclusions) | 29.3 | 155,308 | [REITS: 35.4, Institutional: 26.1, REPE: 16.4, $OOD_N$: 15.5] |
| Exclude REPE and Institutional | 24.7 | 128,177 | [REITS: 56.0, $OOD_N$: 34.6] |
| Exclude REPE, Institutional and REITS | 17.2 | 112,076 | [$OOD_N$: 80.2, $OOD_L$: 8.4, Foreign: 7.8] |

*Notes:* Figure shows the CRE Price dynamics for office and apartment buildings owned by REITs during 2006-2007. $\tau$, is average fraction of assets that trade in every draw as defined in 3.3. ppu/ppsf shows the price per unit for Apartments and price per square foot for Offices, conditional on a sale. The first row in every table refers to the actual data.

The third row shows transaction probabilities and average prices conditional on a transaction when we exclude REPE and Institutional investors from the pool of potential buyers. We find that there would have been 20.2% (24.0%) fewer office and 11.0% (15.7%) fewer apartment transaction in 2006 (2007) without REPE

[36]One possibility for the discrepancy between model and data is that the selection in the model, in terms of the average idiosyncratic shocks $\epsilon_b + \epsilon_s$ of the properties that actually trade, is smaller than in the data. Another possibility is that the model under-prices the specific set of office and apartment assets held by REITs in 2007.

and Institutional buyers. Average sales prices for offices sold by REITs in 2006 (2007) would have traded 6.7% (22.7%) lower, while apartment properties would have sold for 2.1% (17.5%) less in 2006 (2007). The reason for these price drops is that the model identifies REPE and Institutional investors as having particularly high valuations for these properties relative to other investors in those years (high $h_{b,t}$). Specifically, the last column shows that national owner-operator-developers ($OOD_N$) assume a much higher share of transactions in the counter-factual. These investors can be considered as the "value investors" in this market, looking for attractively priced assets. There are many of them; Table 1 showed that $OOD_N$ is the largest investor category by dollar volume, accounting for 30% of buys and sells in our data set.

We note that in this counterfactual, REITs also emerge as buyers of the assets sold by (other) REITs. As an additional counterfactual, we ask what would have happened if all REITs had stayed away from the office and apartment properties considered in this counterfactual. The fourth row of Table 6 shows that REITs would have sold a lot fewer assets than in the benchmark (between 36-43% drop in transactions across the four groups of assets) and that transaction prices would have fallen significantly more. The average price for the offices sold in 2006 (2007) would have been 10.5% (25.4%) lower, respectively than average observed prices in row 1. The average price for these offices would have been 10.5% (25.4%) lower in 2006 (2007), while the average price for apartments would have been 14.7% (27.8%) lower in 2006 (2007). In this counterfactual, local owner-operator-developers and end-users emerge as alternative buyers of offices and some foreigners step in to buy the apartments. Local OOD and Users have even lower valuations for these properties, possibly because of a size, location, or quality mismatch with their existing holdings or because of tighter financial constraints. This counterfactual illustrates that the presence of REITs as buyers lifted office values by 4% and apartment values by 13%. The model produces substantial heterogeneity in valuations, even within investor types such as REITs.

This exercise suggests that REPE funds and Institutions propped up prices just prior to the GFC. Their high sale prices, relative to the counterfactual valuation, may have blurred an early-warning signal of the impending financial crisis and led to larger losses for the buyers.

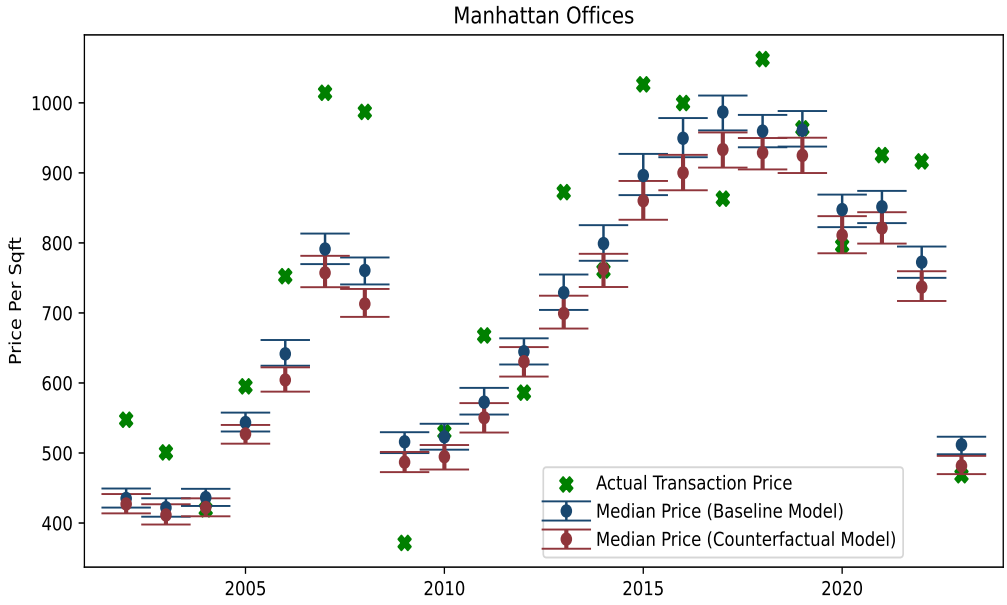
### 7.3 Substitution Patterns across Size, Sector, and Location

In our third application, we show how our model can be used to analyze substitution patterns in deal flows across building size, sector, and location. This is of interest to policymakers and CRE investors alike.

Concretely, we focus on the role of foreign buyers who became large net buyers of U.S CRE in 2015–2018 and again in 2021, as shown in Figures 4 and A.5. These foreign buyers, such as Middle Eastern sovereign wealth funds and large Canadian pension plans, had a strong preference for large, high-end properties in

superstar cities. In addition to understanding the price impact on those types of office assets, we also use the model to study substitution patterns. If a large number of foreign investors purchase large, high-end Manhattan office buildings, what is the impact on transaction volume and prices of smaller, lower-quality office buildings in Manhattan, on retail and apartment buildings in Manhattan, or on office buildings in other cities such as Chicago, Los Angeles, and San Francisco?

Figure 14: Manhattan Office Market: Counterfactuals




*Notes:* This figure shows the potential price distribution for the top-quartile of Manhattan office properties in the baseline model (in blue) and in a counterfactual model without foreign investors (in red) for each year between 2001 and 2023. The top (bottom) line segment indicates the 75th (25th) percentile of potential price distribution, and the dot, the median. The actual transaction price is given by the green cross. All prices are in 2023 dollars.

We start from the universe of high-end office buildings in Manhattan, defined as properties in the top quartile of the valuation distribution ($h_b + h_s | v_b > v_s$). Following the algorithm described in Section 3.3, we assess the impact of foreign buyers by comparing a benchmark model where all buyers are present to a counterfactual economy in which foreign buyers are excluded. We draw $C = 1,000$ potential buyers for each of the the high-end Manhattan office properties in proportion to their model-implied likelihood of buying the specific asset under consideration.

The blue line segments in Figure 14 show the potential price distribution for the Manhattan office properties bought by foreigners in each year between 2001 and 2023 according to the baseline model that includes Foreign buyers in the potential buyer pool. The top and bottom whiskers indicate the 75th and 25th per-

centiles and the blue dot the median of the potential price distribution. The green cross is the actual transaction price. The red whiskers show the potential price distribution for the same assets in a counterfactual world where we remove all Foreign buyers. The assets that in reality were bought by Foreigners are now bought by other types of investors, sampled in proportion to their model-implied transaction probabilities. We see that the counterfactual price distribution shifts down meaningfully from the benchmark one. Removing foreign buyers lowers high-end Manhattan office prices by 4.3% over the full sample period 2001–2023, and by 5.5% in the 2013-2023 subsample.

If we define the high-end of the Manhattan office market as the top decile instead of the top quartile of properties, the valuation impact of foreign buyers grows to 6.3% for the full sample period and to 7.5% post-2013.

Put differently, the foreign purchase activity meaningfully lifted the prices of Manhattan office assets, and especially at the top end of the market.

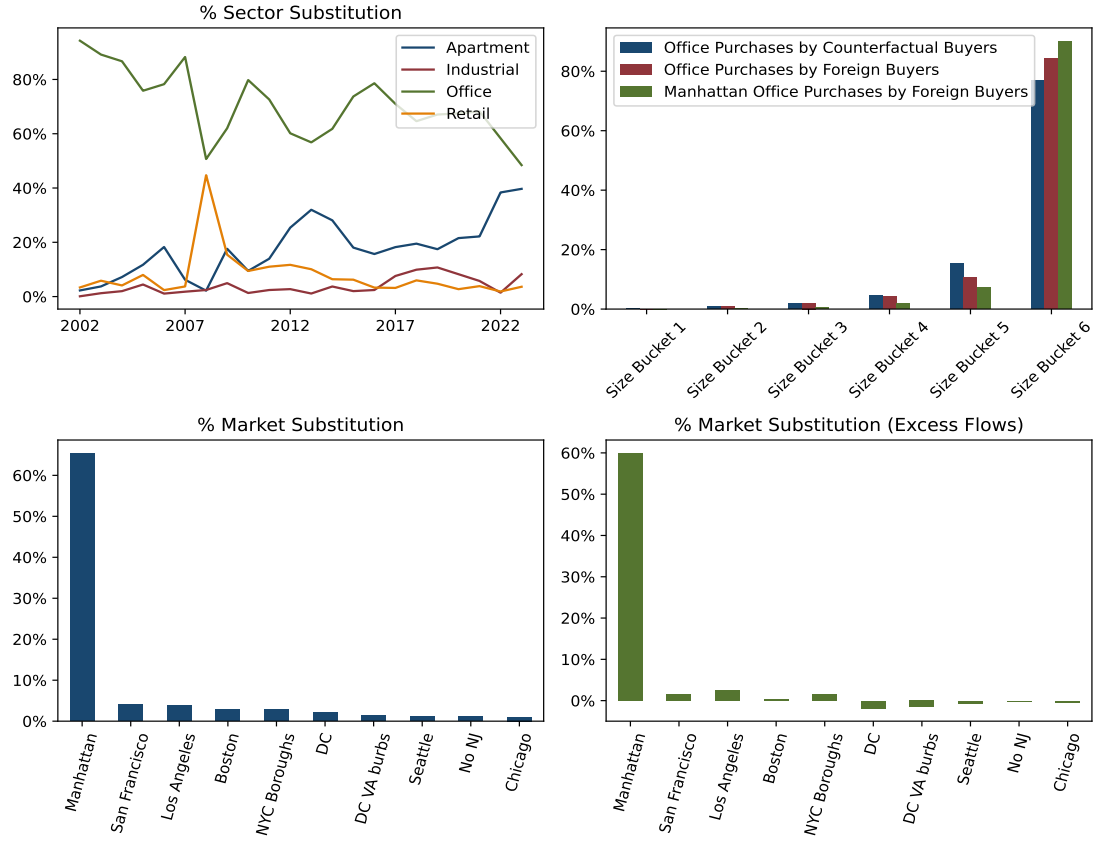
The counterfactual exercise gives us a list of potential buyers that would have bought the properties in question in the absence of foreign investors. With this list in hand, we explore whether those buyers bought another property in the same year or in year thereafter. By analyzing what kinds of properties these displaced investors bought instead, we can learn about the substitution patterns across property sizes, sectors, and locations.

The results are in Figure 15. The top left panel shows the sectoral composition of purchase volume by the counterfactual buyers. The majority of their purchases are in the office sector, about 60% on average. There is substitution to apartments, which represent about 25% of purchases by the counterfactual buyers post-2013.

The top right panel studies substitution across property sizes. For reference, the red (green) bars plot the size distribution of the actual foreign purchases of (Manhattan) office properties. The blue bars show the size distribution of actual office purchases made by the counterfactual buyers in those same years in any market. We find that foreigners predominantly purchased very large offices and pushed other office buyers towards smaller properties.

The bottom left panel shows the locations of the purchases made by the counterfactual buyers across all sectors. There is limited substitution across space. Over 60% of actual purchase activity by this group took place in Manhattan. Foreign buyers may have crowded them out of specific Manhattan office assets, but it did not fully crowd them out of Manhattan. The counterfactual buyers may themselves have strong preferences for (the types of assets available in) Manhattan, and have shifted to Manhattan apartments rather than to San Francisco offices. To the extent that there is geographic substitution, it is fairly evenly directed towards other superstar markets such as San Francisco, Boston, Los Angeles, Chicago and DC.

Figure 15: Manhattan Office Market: Substitution Patterns




*Notes:* This figure examines the purchasing behavior of the potential buyers who would have bought Manhattan office properties had it not been for the Foreign investors who actually bought them. The top left panel plots the share of transaction volume invested by these potential buyers across the four sectors. The top right panel shows the property size distribution of office purchases across all markets by the counterfactual buyers (blue bars) and compares them to the size distribution of Foreign purchases of Manhattan office properties (red bars). The bottom left panel plots the top ten markets by average percentage of purchase volume by the counterfactual buyers. The bottom right panel subtracts from the bottom left panel the average percentage of purchase volume by all manhattan office buyers.

The bottom right panel of Figure 15 subtracts from the geographic portfolio shares of the counterfactual buyers, shown in the bottom left panel, the geographic portfolio shares of all investors who bought an office in Manhattan in the same period. It shows that the counterfactual buyer pool is much more heavily concentrated in Manhattan, relative to the average Manhattan office buyer. Put simply, foreigners crowded out Manhattan office specialists.

# 8 Conclusion

We provide a novel approach to determine the joint evolution of prices and volume in decentralized markets for private and real assets. We apply our approach to a large and granular data set of commercial real estate

transactions that contain detailed information not only on property characteristics but also on the identities of buyers and sellers. We find that using the investor composition in the commercial real estate ecosystem is critical to understanding asset prices. The meeting model, inspired by directed search and matching theory, demonstrates that buyer-seller interactions are shaped by factors such as investor size, portfolio similarity between buyers and seller, and asset feature alignment in terms of property size, location, sector, and quality. These insights underscore the role of matching frictions and heterogeneous preferences in shaping transaction dynamics. The counterfactual analyses highlight the pivotal role of investor composition in determining price outcomes. For example, foreign buyers increased Manhattan office prices by more than 5% in the post-2013 period. REPE fund purchases of large REIT sales in the office and apartment sectors in 2006 and 2007 increased property valuations by more than 10%. The commercial real estate price collapse in the Great Financial Crisis may have been less severe absent eager REPE buyers, emphasizing how specific investor groups influence market equilibria.

These insights extend beyond real estate. Our paper offers a powerful new methodology for understanding price formation and transaction dynamics in other private asset classes, such as private equity, private credit, infrastructure, and residential real estate, where buyer-seller interactions are similarly crucial.

Furthermore, our new estimation technique for matching functions, which leverages insights in the computer science literature, should have broad applicability to other search and matching settings such as labor and marriage markets.

# References


**Andersen, Steffen, Badarinza, Cristian, Liu, Lu, Marx, Julie and Ramadorai, Tarun**. (2022). 'Reference Dependence in the Housing Market', *American Economic Review* 112(10), 3398–3440.

**Ang, Andrew, Chen, Bingxu, Goetzmann, William N. and Phalippou, Ludovic**. (2018). 'Estimating Private Equity Returns from Limited Partner Cash Flows', *Journal of Finance* 73(4), 1751–1783.

**Arora, Abhishek and Dell, Melissa**. (2023). 'LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models'.

**Badarinza, Cristian, Balasubramaniam, Vimal and Ramadorai, Tarun**. (2024), In Search of the Matching Function in the Housing Market, Technical report, SSRN.

**Badarinza, Cristian and Ramadorai, Tarun**. (2018). 'Home Away from Home? Foreign Demand and London House Prices', *Journal of Financial Economics* 130(3), 532–555.

**Badrinza, Cristian, Ramadorai, Tarun and Shimizu, Chihiro**. (2022). 'Gravity, Counterparties and Foreign Investment', *Journal of Financial Economics* 145(2), 132–152.

**Cvijanović, Dragana, Milcheva, Stanimira and van de Minne, Alex**. (2022). 'Preferences of Institutional Investors in Commercial Real Estate', *The Journal of Real Estate Finance and Economics* 65, 321–359.

**Damen, Sven, Korevaar, Matthijs and Van Nieuwerburgh, Stijn**. (2025), An Alpha in Affordable Housing?, Working Paper 33470, National Bureau of Economic Research, Cambridge, MA.

**Driessen, Joost, Lin, Tse-Chun and Phalippou, Ludovic**. (2012). 'A new method to estimate risk and return of nontraded assets from cash flows: The case of private equity funds', *Journal of Financial and Quantitative Analysis* 47(3), 511–535.

**Fox, Jeremy T.** (2007). 'Semiparametric estimation of multinomial discrete-choice models using a subset of choices', *RAND Journal of Economics* 38(4), 1002–1019.

**Fox, Jeremy T.** (2018). 'Estimating Matching Games with Transfers', *Quantitative Economics* 9(1), 1–38.

**Francke, Marc and van de Minne, Alex**. (2024). 'Combining Machine Learning and Econometrics: Application to Commercial Real Estate Prices', *Real Estate Economics* 52(5), 1308–1339.

**Fuster, Andreas, Goldsmith-Pinkham, Paul, Ramadorai, Tarun and Walther, Ansgar**. (2022). 'Predictably Unequal? The Effects of Machine Learning on Credit Markets', *Journal of Finance* 77(1), 5–47.

**Gabaix, Xavier, Koijen, Ralph S.J., Mainardi, Federico, Oh, Sangmin Simon and Yogo, Motohiro**. (2025), Limited Risk Transfer between Investors: A New Benchmark for Macro-Finance Models, Technical report, NBER Working Paper 33336.

**Gabaix, Xavier, Koijen, Ralph S.J., Richmond, Robert and Yogo, Motohiro**. (2024), Asset Embeddings, Technical report, SSRN Working Paper Series.

**Genesove, David and Mayer, Christopher**. (2001). 'Loss Aversion and Seller Behavior: Evidence from the Housing Market', *Quarterly Journal of Economics* 116, 1233–1260.

**Ghent, Andra C.** (2021). 'What's wrong with Pittsburgh? Delegated investors and liquidity concentration', *Journal of Financial Economics* 139(2), 337–358.

**Giacoletti, Marco**. (2021). 'Idiosyncratic Risk in Housing Markets', *The Review of Financial Studies* 34(8), 3695–3741.

**Goetzmann, William N, Spaenjers, Christophe and Van Nieuwerburgh, Stijn**. (2021). 'Real and private-value assets', *The Review of Financial Studies* 34(8), 3497–3526.

**Griliches, Zvi**. (1971), Introduction: Hedonic Price Indexes Revisited, *in* **Zvi Griliches**., ed., 'Price Indexes and Quality Change: Studies in New Methods of Measurement', Harvard University Press, Cambridge, MA, pp. 3–15.

**Guolin Ke, Qi Meng, Thomas Finley Taifeng Wang Wei Chen Weidong Ma Qiwei Ye Tie-Yan Liu**. (2017). 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree', *Advances in Neural Information Processing Systems* 30, 3149–3157.

**Gupta, Arpit, Mittal, Vrinda and Van Nieuwerburgh, Stijn**. (2025). 'Work From Home and the Office Real Estate Apocalypse', *American economic Review* forthc.

**Gupta, Arpit and Van Nieuwerburgh, Stijn**. (2021). 'Valuing private equity investments strip by strip', *The Journal of Finance* 76(6), 3255–3307.

**Gu, Shihao, Kelly, Bryan and Xiu, Dacheng**. (2020). 'Empirical Asset Pricing via Machine Learning', *The Review of Financial Studies* 33(5), 2223–2273.

**Kaplan, Steven N. and Schoar, Antoinette**. (2005). 'Private equity performance: Returns, persistence, and capital flows', *Journal of Finance* 60(4), 1791–1823.

**Kim, Min**. (2024), Comparative R-Squared. SSRN Working Paper No. 3790200.

**Koijen, Ralph S.J., Richmond, Robert J. and Yogo, Motohiro**. (2024). 'Which Investors Matter for Equity Valuations and Expected Returns', *Review of Economic Studies* 91(4), 2387–2424.

**Koijen, Ralph S.J. and Yogo, Motohiro**. (2019). 'A Demand System Approach to Asset Pricing', *Journal of Political Economy* 127(4), 1475–1515.

**Korteweg, Arthur and Nagel, Stefan**. (2016). 'Risk-Adjusting the Returns to Venture Capital', *Journal of Finance* 71(3), 1437–1470.

**Korteweg, Arthur and Sørensen, Morten**. (2010). 'Risk and Return Characteristics of Venture Capital-Backed Entrepreneurial Companies', *Review of Financial Studies* 23, 3738–3772.

**Lancaster, Kevin**. (1966). 'A New Approach to Consumer Theory', *Journal of Political Economy* 74, 132–157.

**Lundberg, Scott M, Erion, Gabriel, Chen, Hugh, DeGrave, Avery, Prutkin, Jordan M, Nair, Bharat, Katz, Ronit, Himmelfarb, Jonathan, Bansal, Nisha and Lee, Su-In**. (2020). 'From local explanations to global understanding with explainable AI for trees', *Nature Machine Intelligence* 2(1), 252–259.

**Lundberg, Scott M and Lee, Su-In**. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30.

**Ma, Zhuang and Collins, Michael**. (2018). 'Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency', *CoRR* abs/1809.01812.

**McFadden, Daniel L.** (1977), Modelling the Choice of Residential Location, Discussion Paper 477, Cowles Foundation for Research in Economics, Yale University, New Haven, CT.

**Mikolov, Tomás, Sutskever, Ilya, Chen, Kai, Corrado, Greg and Dean, Jeffrey**. (2013). 'Distributed Representations of Words and Phrases and their Compositionality', *CoRR* abs/1310.4546.

**Mittal, Vrinda**. (2024), Desperate Capital Breeds Productivity Loss: Evidence from Public Pension Investments in Private Equity. SSRN Working Paper No. 4283853.

**Peng, Liang**. (2016). 'The risk and return of commercial real estate: A property level analysis', *Real Estate Economics* 44(3), 555–583.

**Plazzi, Alberto, Torous, Walter and Valkanov, Ross**. (2010). 'Expected Returns and the Expected Growth in Rents of Commercial Real Estate', *Review of Financial Studies* 23, 3469–3519.

**Plazzi, Alberto, Torous, Walter and Valkanov, Rossen**. (2008). 'The Cross-Sectional Dispersion of Commercial Real Estate Returns and Rent Growth: Time Variation and Economic Fluctuations', *Real Estate Economics* 36(3), 403–439.

**Rosen, Sherwin**. (1974). 'Hedonic Prices and Implicit Markets: Production Differentiation in Pure Competition', *Journal of Political Economy* 82, 34–55.

**Sagi, Jacob S.** (2021). 'Asset-Level Risk and Return in Real Estate Investments', *Review of Financial Studies* 34(8), 3647–3686.

**Takuya Akiba, Shotaro Sano, Toshihiko Yanase Takeru Ohta and Koyam, Masanori**. (2019). 'Optuna: A Next-generation Hyperparameter Optimization Framework', *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery  Data Mining* pp. 2623–2631.

**Van Nieuwerburgh, Stijn**. (2019). 'Why Are REITs Currently So Expensive?', *Real Estate Economics* 47(1), 18–65.

**Van Nieuwerburgh, Stijn, Stanton, Richard and de Bever, Leo**. (2015), 'A Review of Real Estate and Infrastructure Investments by the Norwegian Government Pension Fund Global (GPFG)'.

**Wallace, Nancy E.** (1996). 'Hedonic-based price indexes for housing: Theory, estimation, and index construction', *Economic Review* pp. 34–48.

**Witte, Ann D., Sumka, Howard J. and Erekson, Homer**. (1979). 'An Estimate of a Structural Hedonic Price Model of the Housing Market: An Application of Rosen's Theory of Implicit Markets', *Econometrica* 47(5), 1151–1173.

# A Data Appendix

## A.1 Data Cleaning Process

This appendix describes the process used to clean and filter the raw transaction data from RCA. The dataset includes transaction-level data for four core sectors—apartments, industrial, office, and retail—from 2001 to 2023.

We begin by removing transactions that are unlikely to be useful or may introduce noise. Using the transaction type variable, we retain only "Sale," "Construction," "Entity Level," and "Transfer" transactions, excluding land sales, air rights, and leases. Additionally, we drop incomplete transactions (e.g., contracts not completed) and those valued under $500,000. These filters eliminate approximately 5% of all transactions.

For each transaction, RCA provides up to four buyers and sellers, though less than 1% involve all four. We consolidate investor names with inconsistent spellings or formats using the method in Arora and Dell (2023). Specifically, we use two independent models to generate 20 closest matches per name, apply a 90% similarity threshold, manually discard false positives, and replace names with the most frequent version. This standardization allows us to track investors over time, aiding in two key tasks: (1) identifying property owners for each year (critical for the seller role in our meeting model), and (2) constructing investor-level portfolios to study asset evolution. The latter supports variables like investor transaction activity, size, and consideration set characteristics.

Transactions suspected of being non-arms-length (e.g., multiple deals involving the same property within 30 days) are consolidated. We merge such transactions into a single record, using the price and characteristics of the final transaction and assigning buyers and sellers from the first and last transactions, respectively. For properties that change sectors during the sample period due to conversions, we assign separate property IDs for periods before and after such conversions, to ensure the accuracy of analysis.

Transactions where the buyer and seller are the same entity are removed. For cases where one party transacts on behalf of another, we retain the primary investor and drop intermediaries (e.g., investment managers). RCA's investor names typically reflect the final entity; if data references a subsidiary, we replace it with the parent. If there are multiple buyers or sellers, we allocate equal ownership shares unless specific ownership structures are known (e.g., in Joint Ventures, where we adjust shares accordingly).

We consolidate RCA's raw investor types (around 20 groups) into eight core categories (Table A.7) to better capture differences in investment objectives, size, and geographic focus. We also manually correct certain investor attributes, such as incorrect country assignments or misclassified types. Next, we assign unique investor IDs based on the cleaned investor names. When an investor name is missing (approximately 5% of cases), we assign a distinct ID to that record. These IDs form the basis for tracking investor portfolios

Next, we fill missing property characteristics. Property age is calculated as the number of years between the transaction date and the construction year. Renovation-adjusted age uses the most recent renovation year; if unavailable, it is set equal to age. For occupancy rate, around 30% of properties have missing values. We first forward-fill for up to one year using the property's own most recent observation. For gaps longer than one year, we estimate a linear model using market-level occupancy, renovation-adjusted age, neighbourhood net effective rent, and the property's most recent observed occupancy (if available). We fill gaps with the model's fitted values, defaulting to the property's market-level occupancy rate when these cannot be produced.

We track property ownership over time to construct investor portfolios. Challenges such as missing or incorrect transactions are addressed by inferring missing links from prior and subsequent transactions. For

example, if investor A sells to B, and C later sells the same property to D, we estimate an intermediate transaction where B sells to C. We adjust buyer and seller shares to ensure total ownership sums to 100% without negative values. Approximately 20% of properties require such adjustments during the sample period. These corrections are used solely for portfolio construction and are excluded from valuation and meeting model calibrations, which rely strictly on actual transaction data.

Before computing portfolio variables, we first establish the current owner of each property. We assume that the most recent known owner held the property continuously until another transaction is observed. When a transaction occurs, ownership is updated accordingly. This step is critical for identifying the seller in each transaction, as the seller is simply the current owner immediately prior to the sale. Portfolio variables are then computed sequentially from the transaction history. For each investor, we calculate:

1. Portfolio size: Total market value of properties owned, adjusted for fractional ownership. The investor size is in 2023 dollars and is obtained by using the annual property price panels described below.

2. Number of properties: Count of unique holdings, adjusted for fractional stakes.

3. Sector and geographic composition: share of portfolio value in each RCA sector and aggregated market (see Section A.2).

4. Average property value: The average market value of properties in the investor's portfolio.

For entity sales, where RCA equally distributes transaction value across properties, we adjust property-level prices by assigning per-square-foot values proportional to RCA's sector-market price indices. This provides a more accurate estimate than assuming uniform prices across locations.

We construct property price panels for each year by combining RCA price indices with unadjusted CPI prices and adjusting them to 2023 dollars. CPI adjustments use the seasonally adjusted Consumer Price Index (CPIAUCSL) from FRED.

Market occupancy data is primarily sourced from NCREIF at the CBSA-sector level. For missing values, we calculate volume-weighted occupancy rates using RCA data, provided at least 15 properties in the sector-year contain valid occupancy information. Net Operating Income (NOI) data, also from NCREIF, is CPI-adjusted and annualized. NOI growth is calculated logarithmically if both the current and previous NOI values are positive; otherwise, we use a smoothed growth rate formula. Extreme outliers (top/bottom 1%) are replaced with median values.

Geographic variables (e.g., zip codes, FIPS codes) are filled where possible. Geographic variables (e.g., ZIP codes, FIPS codes) are filled wherever possible. For markets with intermittent coverage (at least 50%), we impute missing values using the average from available observations within the same market. Gaps are more common in remote areas and in retail or industrial properties, which tend to be farther from city centers. In the rare cases where no market-level information exists, we use sector-level national averages.

## A.2 Mapping Geographies

The RCA dataset provides geo-codes for every property, which we leverage to map properties to various geographic indicators. This is essential because the macroeconomic time series data we use comes from multiple sources, each often organized by different geographic units. Using the 3-digit rounded latitude and longitude data provided by RCA, we generate ZIP codes for each property. This is done using standard geolocation algorithms that match coordinates to corresponding ZIP codes.

We use the county-level Federal Information Processing Standards (FIPS) codes to map properties to Core-Based Statistical Areas (CBSAs). This mapping is based on a crosswalk file obtained from the National Bureau of Economic Research (NBER). RCA provides its own market classification system, which consists of 165 individual markets. To align with other datasets and enhance interpretability, we aggregate RCA's 165 markets into 61 broader geographic categories based on logical groupings and economic characteristics. The aggregated RCA markets enable direct mapping of properties to RCA price index files and cap rate data, both of which are reported at the level of these 61 aggregated markets. This standardization allows for consistent integration of property-level data with macroeconomic indicators. Table A.1 shows the mapping between RCA markets and our 60 geographic markets.

Table A.1: Mapping RCA Markets to Our Markets

| **Our Market** | **RCA Markets** | **Our Market** | **RCA Markets** |
|---|---|---|---|
| Atlanta | Atlanta | Manhattan | Manhattan |
| Austin | Austin | Memphis | Memphis |
| Baltimore | Baltimore | Miami/Dade Co | Miami/Dade Co |
| Birmingham | Birmingham (AL) | Minneapolis | Minneapolis |
| Boston | Boston | Nashville | Nashville |
| Broward | Broward | No NJ | No NJ |
| Charlotte | Charlotte | NYC Boroughs | NYC Boroughs |
| Chicago | Chicago | Orange Co | Orange Co |
| Cincinnati | Cincinnati | Orlando | Orlando |
| Cleveland | Cleveland | Palm Beach Co | Palm Beach Co |
| Columbus | Columbus | Philadelphia | Philadelphia |
| Dallas | Dallas | Phoenix | Phoenix |
| DC | DC | Pittsburgh | Pittsburgh |
| DC MD burbs | DC MD burbs | Portland | Portland |
| DC VA burbs | DC VA burbs | Raleigh/Durham | Raleigh/Durham |
| Denver | Denver | Richmond/Norfolk | Norfolk, Richmond |
| Detroit | Detroit | Sacramento | Sacramento |
| East Bay | East Bay | Salt Lake City | Salt Lake City |
| Hartford | Hartford | San Antonio | San Antonio |
| Houston | Houston | San Diego | San Diego |
| Indianapolis | Indianapolis | San Francisco | San Francisco |
| Inland Empire | Inland Empire | San Jose | San Jose |
| Jacksonville | Jacksonville | Seattle | Seattle |
| Kansas City | Kansas City | St Louis | St Louis |
| Las Vegas | Las Vegas | Stamford | Stamford |
| Long Island | Long Island | Tampa | Tampa |
| Los Angeles | Los Angeles | Westchester | Westchester |

| Our Market | RCA Markets |
|---|---|
| Tertiary Mid-Atlantic | All Others - DE,MD,VA,WV; Ann Arbor, Albany, All Others - NJ, All Others - NY, All Others - PA, Allentown, Buffalo, Camden, NJ, Eastern PA, Harrisburg, Rochester, Syracuse, Trenton, NJ |
| Tertiary Midwest | Lexington, Louisville, Akron, All Others - IL, IN, OH, Dayton, Kankakee, IL, Lansing Toledo, All Others - Plains States, Des Moines, Omaha, Wichita, All Others - MI,WI, Flint, MI, Grand Rapids, Madison, Milwaukee, Monroe, Racine, All Others - MN |
| Tertiary Northeast | All Others - CT, All Others - New England, New Haven, Portland, ME, Providence, Southern NH, Worcester |
| Tertiary Southeast | Polk Co, All Others - FL, Brevard Co, Daytona Beach, Florida Panhandle, Fort Myers, Gainesville, Martin/Saint Lucie, Sarasota, Tallahassee, All Others - GA,NC,SC,TN,KY, Charleston, Chattanooga, Columbia, Greensboro, Greenville, Knoxville, Myrtle Beach, Savannah, Winston-Salem, All Others - AR,LA, All Others - MS,AL, Baton Rouge, Huntsville, AL, Jackson, MS, Little Rock, New Orleans |
| Tertiary Southwest | Boulder, Greeley, Albuquerque, All Others - Southwest, Tucson, All Others - OK,TX, Corpus Christi, El Paso, McAllen, Oklahoma City, Tulsa, Co Springs, Fort Collins, Provo |
| Tertiary West | Santa Barbara, Ventura Co, All Others - Northwest, Spokane, Bakersfield All Others - Southern California, Reno, Fresno, Madera, CA, Modesto, Salinas, Vallejo-Fairfield, All Others - Northern California, Santa Rosa, Boise, Napa, All Others - ID,MT,WY; Hawaii, Honolulu |

*Notes:* This table provides the mapping we use from the raw RCA markets to the aggregated RCA markets we use.

## A.3 Net Effective Rent

This appendix describes how we construct net effective rent statistics of the property or the neighborhood of each property that we see transact in the RCA data. Net effective rents are observed for office, retail, and industrial properties at the lease level in the Compstak database. We first merge the RCA and the CompStak databases at the property level using the following procedure: for each RCA property, we consider CompStak buildings within a 1 km radius of its geo point. We assign each CompStak building a matching score based on the geographical distance, the difference between the years the properties were built, the difference in log building size, and the fuzzy matching score on property names. Specifically, the matching score equals the sum of: (1) $0.3 \times \max\{1 - \frac{\text{geo distance (km)}}{0.3}, 0\}$, (2) $0.3 \times \max\{1 - \frac{|\Delta\text{year built}|}{5}, 0\}$, (3) $0.3 \times \max\{1 - \frac{|\Delta\text{log building size}|}{5\%}, -1\}$, and (4) $0.4 \times \max\{\frac{\text{name similarity}-70}{30}, 0\}$, where name similarity is the highest partial Levenshtein similarity between elements in the RCA property name list and the CompStak name list (including all street addresses and building names of that property). Additionally, if the similarity score is exactly 100, we assign this part of the score as $0.4 \times 2$ instead of $0.4 \times 1$. We keep the CompStak building with the highest matching score. If the maximum score $\leq 0.3$, we deem that there is no match at the property level.

We manually check the matching result for the top 100 RCA office buildings transacted after 2015 in Manhattan: the correct matching rate is 92%, wrong matching rate is 3%, and the missing rate is 5%. Wrong matches occur when very similar and adjacent buildings have characteristics closer than the true match, while the missing matches occur because CompStak does not provide coverage for all RCA properties. The full-sample building-level matching rates are 55.1%, 39.3%, and 48.3% for office, retail, and industrial properties, respectively. Matching rates at the building-year level are naturally lower as we do not necessarily observe new leases in the matched buildings in every year.

For RCA properties that are not matched with a CompStak building, or matched with a CompStak building lacking sufficient recent lease data to compute measures, we create variables from the lease-level data at a fine level of geography (call it a "block"), and match each property we see transact in RCA to its block-level variables.

The finest level of geography we consider uses the first 2 digits of the latitude and longitude. For example, for a property with latitude and longitude (40.754932,-73.984016), we define a 2d area as [40.75,40.76;-73.99,-73.98]. This corresponds to an area of roughly 0.6 miles by 0.6 miles (1km by 1km). We collect all the leases in this block to calculate the average rent measures.

Given that leasing activity is sparse, we use all leases signed in the past three years to calculate the property-year or block-year average rent. We weigh leases by their transaction sqft, and assign a lower weight for leases signed in the more distant past. In the baseline, we assign weights of 1, 1/2, and 1/3 to leases signed one, two, and three years before, and then renormalize to make the weights sum to one. As a robustness check, we also compute a variable using the past 5 years of leasing data, with weights on past leases of 1, 1/2, 1/3, 1/4, and 1/5, renormalized to sum to one.

When there are fewer than 5 buildings in the block, or when there are no leases in the past three/five years, we use the average rent calculated for a larger block. The next-larger block size we consider is defined based on the first decimal place of the latitude and longitude. For example, a property at latitude and longitude (40.754932,-73.984016) belongs to block [40.7,40.8;-74.0,-73.9]. The typical size of this 1-digit block is 6 miles by 6 miles (10km by 10km).

When the average rent is still missing in this 1-digit block, we fill it with the value in the block defined based on the unit digit of the latitude and longitude (0-digit block). In our example, this would be the area

given by [40,41;-73,-74], an area of roughly 60 miles by 60 miles.

For the remaining missing values, we sequentially fill them with the average rent at the RCA market, average rent at our own market definition (which combines several RCA markets, see appendix A.2), and if necessary, average rent measured at the national level.

For apartment properties, we observe the net operating income per apartment unit at the loan level in the Fannie Mae's multifamily mortgage origination data. We follow a similar approach to calculate the NOI (referred to as NER below for simplicity). Given that loan origination activity is also sparse, we use all originations in the past three years to calculate the average NER. We first use the same matching algorithm to conduct a property-level matching, obtaining a matching rate of 45.8% at the building level, and then use the block-year-level NER to fill in missing values. We start with the 2-digit blocks, use the same weighting scheme explained above, and fill missing values sequentially with values computed on increasingly coarse geographic levels (1-digit, 0-digit, RCA market, our market, national).

Table A.2 reports the proportion of the final NER observations for each sector (columns) that come from each level of geography (rows). Most of the NER are measured at the 1-decimal level (6 miles by 6 miles). For office NER of the last three years, we have 24.9% of observations measured at the even finer 2-decimal level (0.6 miles by 0.6 miles) and 20.8% at the property level.

Table A.2: NER Information Source

**Panel A:** NER (Last 3 Years)

| **Level** | **Office Sector** | **Industrial Sector** | **Retail Sector** | **Apartment Sector** |
|---|---|---|---|---|
| Property level | 20.8% | 11.7% | 6.5% | 8.4% |
| 2-decimal lat/lon | 24.9% | 25.3% | 15.4% | 31.2% |
| 1-decimal lat/lon | 38.9% | 40.2% | 49.5% | 50.1% |
| unit-digit lat/lon | 10.8% | 15.2% | 21.3% | 9.5% |
| RCA Market | 2.5% | 3.3% | 4.4% | 0.8% |
| Market | 1.7% | 2.9% | 2.4% | 0.0% |
| National | 0.4% | 1.5% | 0.6% | 0.0% |
| Regression $R^2$ | 69.4% | 59.7% | 44.3% | 45.2% |

**Panel B:** NER (Last 5 Years)

| **Level** | **Office Sector** | **Industrial Sector** | **Retail Sector** | **Apartment Sector** |
|---|---|---|---|---|
| Property level | 25.9% | 15.9% | 9.3% | 12.5% |
| 2-decimal lat/lon | 23.2% | 24.7% | 15.9% | 35.7% |
| 1-decimal lat/lon | 37.4% | 38.8% | 49.6% | 44.3% |
| unit-digit lat/lon | 9.7% | 14.1% | 19.0% | 7.0% |
| RCA Market | 2.2% | 2.9% | 4.0% | 0.5% |
| Market | 1.4% | 2.5% | 1.9% | 0.0% |
| National | 0.2% | 1.0% | 0.4% | 0.0% |
| Regression $R^2$ | 66.9% | 54.1% | 43.2% | 43.2% |

To understand how well the block-level NER measures captures property-level variation in NER, we estimate a linear regression of property-level NER in the Compstak/Fannie Mae data on the block-level NER variable, where we use the most spatially granular variable available for that property. The $R^2$ of those property-level regressions is reported in the last row of Table A.2.

We have explored several more variables available in the Compstak data, such as occupancy, quality grade (class A, B, C), remaining lease duration, tenant size (number of employees of the tenant), tenant concentration (Herfindahl index computed from the leases in a building), but none of these variables add

explanatory power to the valuation model, once the NER variable is included.

## A.4 Potential Price Distribution: Process

This appendix describes the practical steps for constructing the potential price distribution (PPD) described in Section 3.3. For the counterfactuals in Sections 7.2 and 7.3, we study how removing specific investor groups affects (i) the probability that a asset trades and (ii) the expected price conditional on trade. For the out-of-sample prediction in Section 7.1, we use the same machinery to forecast next-period transaction prices, conditional on transacting.

We consider the market defined by a calendar year $t$ and a property sector. To reduce clutter, $t$ subscripts are omitted in what follows. For each asset $n$ with known seller $s$ (current owner), we observe asset characteristics $x_n$, seller characteristics $z_s$, and the set of active buyers $B$ (investors that transacted at least once in the preceding five years) with characteristics $z_b$. This allows us to estimate systematic valuations $h_s(n)$ and $h_b(n)$ from the valuation model, meeting probabilities for every potential buyer $b'$: $\{\pi_m(b' \mid s, n)\}_{b' \in B}$, and a common idiosyncratic shock scale $\sigma$ with $\epsilon_s, \epsilon_b \sim N(0, \sigma^2)$ (see Section 5.3). We then follow the steps below:

1. Draw a seller shock $\epsilon_s \sim N(0, \sigma^2)$. Then set $v_s = h_s(n) + \epsilon_s$.
2. For each asset $n$, draw a candidate buyer from $B$ with replacement using weights $\{\pi_m(b' \mid s, n)\}_{b' \in B}$.
3. For each potential buyer, draw a latent demand shock $\epsilon_{b'} \sim N(0, \sigma^2)$ and set $v_{b'} = h_{b'}(n) + \epsilon_{b'}$.
4. If $v_{b'} > v_s$, record the price $\widehat{p}(n) = \frac{1}{2}\big(v_{b'} + v_s\big)$; otherwise set $\widehat{p}(n)$ to missing (no trade).
5. Compute (i) the mean price conditional on trade, $\mathbb{E}[\widehat{p}]$ (average over non-missing $\widehat{p}(n)$), and (ii) the fraction of assets that trade, $\tau = \frac{1}{N}\sum_n \mathbb{1}\{\widehat{p}(n) \text{ observed}\}$.
6. Repeat steps 1–5 for $C = 1000$ replications, obtaining $\mathbb{E}[\widehat{p}]$ and $\tau$ for each run. The empirical distribution of these statistics is the market-level PPD. Asset-level PPDs are obtained by retaining outcomes for each $n$ across replications.
7. To understand the importance of a subset $G$ of investors (e.g., foreign buyers in Manhattan), compute the PPD after removing that group from the buyer pool: set $\pi_m(b' \mid s, n) = 0$ for $b' \in G$, renormalize over $B \setminus G$, and repeat steps 1–6. Compare the distributions of $\mathbb{E}[\widehat{p}]$ and $\tau$ to the baseline.

For market-level counterfactuals, we compare baseline and alternative buyer-composition outcomes by reporting, across replications, the distributions of the mean simulated price conditional on trade ($\mathbb{E}[\widehat{p}]$) and the fraction of assets that trade $\tau$. We interpret the differences as effects on prices and liquidity.

For out-of-sample prediction (predicting prices in $t{+}1$ from the vantage point of $t$), we do not observe next year's asset characteristics $x_{n,t+1}$, investor characteristics $z_{t+1}$, or macroeconomic conditions. We therefore use their time-$t$ values to construct a time-$t$ prediction of the seller's next-period deterministic valuation $h'_{s,t+1}(n)$. The price-shock variance for $t{+}1$ is also unknown; we set $\sigma^2_{p,t+1}$ to the historical estimate $\sigma^2_{p|t} = \mathrm{Var}_t(\epsilon_p)$ computed from $\epsilon_p$ over 2001–$t$. (Equivalently, in simulations with buyer/seller shocks, use $\sigma^2 = 2\,\sigma^2_{p|t}$ since $\mathrm{Var}(\epsilon_p) = \frac{1}{2}\sigma^2$.) We then get the asset-level PPD and set $\widehat{p}^{\,\mathrm{PPD}}_{t+1}(n)$ to the Monte Carlo mean of non-missing simulated prices for asset $n$ (conditional on trade). Forecast improvement

over the linear hedonic model (LHM) is measured on the set $\mathcal{T}_{t+1}$ of assets that actually trade in $t$+1. For instance, the reduction in mean-squared-error is:

$$1 - \frac{\sum_{n \in \mathcal{T}_{t+1}} \left(\widehat{p}_{t+1}^{\text{PPD}}(n) - p_{t+1}(n)\right)^2}{\sum_{n \in \mathcal{T}_{t+1}} \left(\widehat{p}_{t+1}^{\text{LHM}}(n) - p_{t+1}(n)\right)^2},$$

where $\widehat{p}_{t+1}^{\text{LHM}}(n)$ is the linear hedonic benchmark, and $p_{t+1}(n)$ is the realized price.

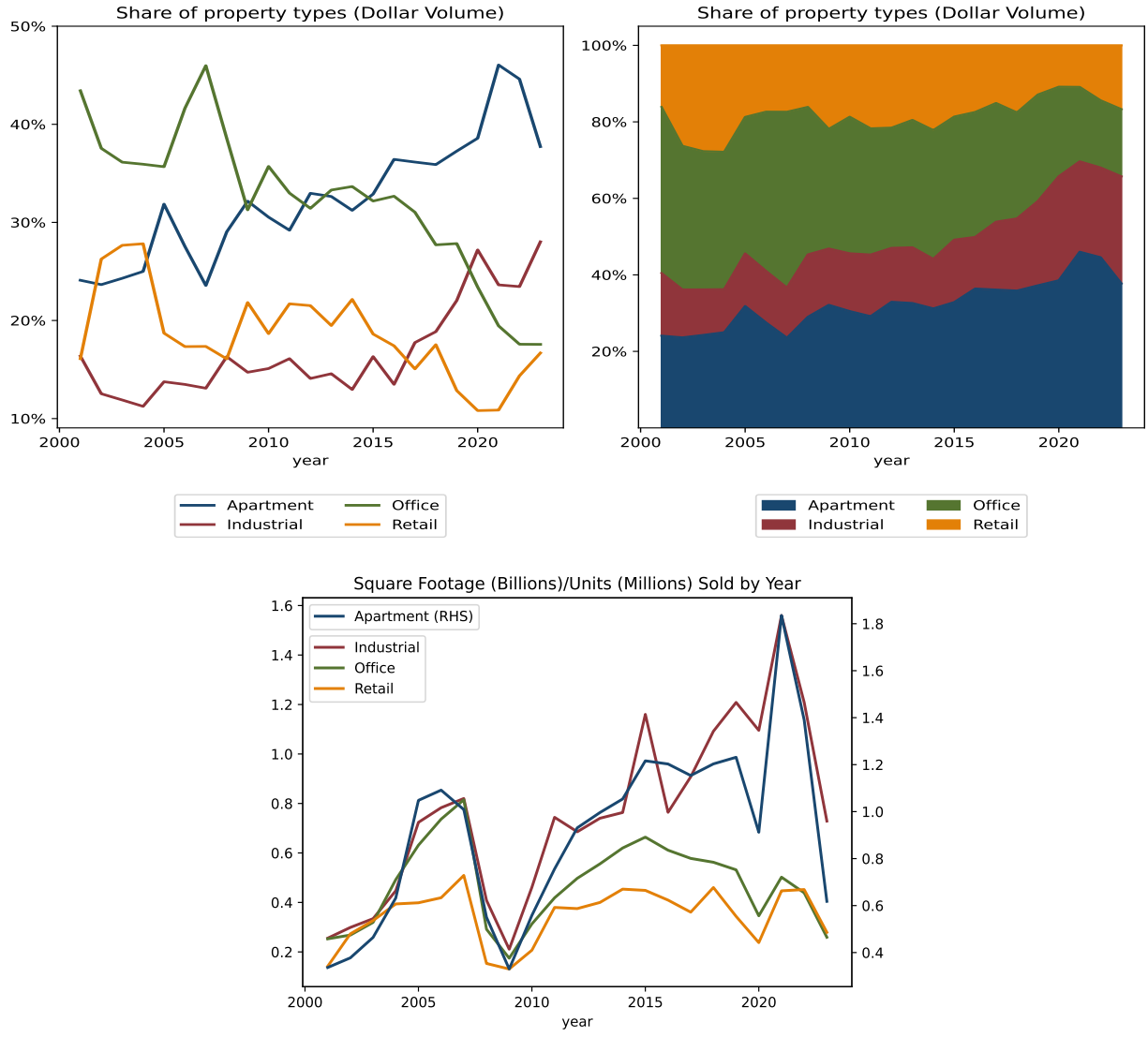
## A.5 Transaction Summary Statistics

Table A.3: Statistics by Transaction Year

| Year | # Trans | % Trans | Cum. % Trans | Volume | % Vol | Cum. % Vol | % Foreclosures | % EntitySales |
|---|---|---|---|---|---|---|---|---|
| 2001 | 5,272 | 1.11 | 1.11 | 150.71 | 1.49 | 1.49 | 0.14 | 6.82 |
| 2002 | 6,287 | 1.32 | 2.43 | 186.11 | 1.84 | 3.34 | 0.06 | 2.19 |
| 2003 | 7,977 | 1.68 | 4.10 | 223.00 | 2.21 | 5.55 | 0.21 | 2.96 |
| 2004 | 11,821 | 2.48 | 6.59 | 348.93 | 3.46 | 9.00 | 0.18 | 7.44 |
| 2005 | 20,725 | 4.35 | 10.94 | 490.05 | 4.86 | 13.86 | 0.15 | 3.14 |
| 2006 | 23,025 | 4.84 | 15.78 | 555.20 | 5.50 | 19.36 | 0.13 | 9.39 |
| 2007 | 23,385 | 4.91 | 20.69 | 675.31 | 6.69 | 26.05 | 0.34 | 21.01 |
| 2008 | 13,531 | 2.84 | 23.53 | 230.60 | 2.28 | 28.34 | 2.70 | 2.83 |
| 2009 | 6,884 | 1.45 | 24.98 | 109.09 | 1.08 | 29.42 | 17.25 | 0.38 |
| 2010 | 10,873 | 2.28 | 27.26 | 211.82 | 2.10 | 31.52 | 11.88 | 0.33 |
| 2011 | 14,749 | 3.10 | 30.36 | 304.41 | 3.02 | 34.53 | 6.64 | 8.58 |
| 2012 | 19,785 | 4.16 | 34.52 | 374.61 | 3.71 | 38.24 | 4.94 | 5.74 |
| 2013 | 21,200 | 4.45 | 38.97 | 434.17 | 4.30 | 42.54 | 2.30 | 5.71 |
| 2014 | 24,974 | 5.25 | 44.22 | 503.24 | 4.99 | 47.53 | 2.27 | 5.90 |
| 2015 | 29,252 | 6.15 | 50.36 | 616.83 | 6.11 | 53.64 | 0.48 | 10.14 |
| 2016 | 26,540 | 5.58 | 55.94 | 574.22 | 5.69 | 59.33 | 0.71 | 4.87 |
| 2017 | 26,996 | 5.67 | 61.61 | 542.27 | 5.37 | 64.70 | 0.73 | 4.33 |
| 2018 | 29,044 | 6.10 | 67.71 | 611.31 | 6.06 | 70.76 | 0.66 | 12.52 |
| 2019 | 30,679 | 6.44 | 74.16 | 627.41 | 6.22 | 76.98 | 0.36 | 1.26 |
| 2020 | 23,793 | 5.00 | 79.15 | 452.95 | 4.49 | 81.46 | 0.48 | 4.37 |
| 2021 | 40,410 | 8.49 | 87.64 | 854.42 | 8.47 | 89.93 | 0.55 | 4.27 |
| 2022 | 36,266 | 7.62 | 95.26 | 697.79 | 6.91 | 96.84 | 0.86 | 8.10 |
| 2023 | 22,554 | 4.74 | 100.00 | 318.54 | 3.16 | 100.00 | 1.68 | 2.95 |

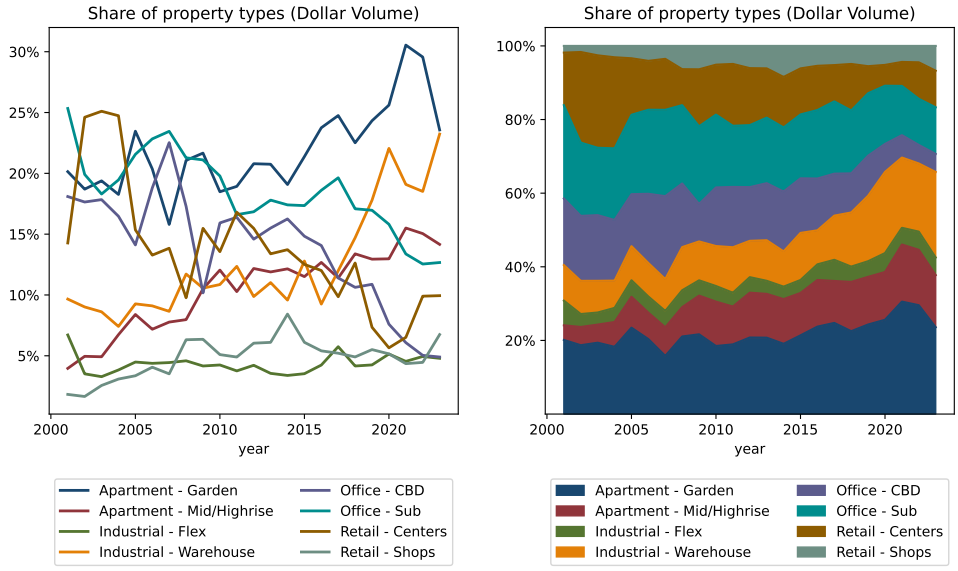
*Notes:* This table shows the number of transactions, the share and cumulative share of total transactions, the volume (dollar value), share and cumulative share of volume by transaction year. The last two columns displays the share of transactions that are foreclosures and the share of transactions that are entity sales. Volume is defined in billions of 2023 U.S. dollars.

Figure A.1: Transaction Volume by Property Type




*Notes:* The top panel shows share of properties transacted by property types, i.e. Apartments, Office, Industrial and Retail, from 2001 to 2023. Value is defined in terms of 2023 USD. The bottom panel shows the total square feet transacted in billions of Industrial, Office and Retail space, and the number of Apartment units bought in millions.

Figure A.2: Transaction Volume by Property Subtype




*Notes:* The top panel shows share of properties transacted by property subtypes., i.e. Apartments, Office, Industrial and Retail, from 2001 to 2023. Value is defined in terms of 2023 USD. The bottom Panel shows the total square footage bought (in billions) of Industrial, Office and Retail space, and the number of Apartment units bought (in millions) from 2001 to 2023.

Table A.4: Transactions By Transaction Size

| | # Trans | % Trans | Cum. % Trans | $ Vol | % Vol | Cum. % Vol |
|---|---|---|---|---|---|---|
| Above 1 Bil | 269 | 0.06 | 0.06 | 327 | 3.24 | 3.24 |
| 500 Mil - 1 Bil | 701 | 0.15 | 0.20 | 374 | 3.71 | 6.95 |
| 250-500 Mil | 2,368 | 0.50 | 0.70 | 704 | 6.97 | 13.92 |
| 100-250 Mil | 12,525 | 2.63 | 3.33 | 1,726 | 17.10 | 31.02 |
| 75-100 Mil | 9,301 | 1.95 | 5.29 | 772 | 7.65 | 38.68 |
| 50-75 Mil | 19,926 | 4.19 | 9.47 | 1,181 | 11.71 | 50.38 |
| 25-50 Mil | 52,693 | 11.07 | 20.54 | 1,814 | 17.97 | 68.35 |
| 20-25 Mil | 22,517 | 4.73 | 25.27 | 496 | 4.91 | 73.26 |
| 15-20 Mil | 33,779 | 7.10 | 32.37 | 578 | 5.72 | 78.99 |
| 10-15 Mil | 57,414 | 12.06 | 44.43 | 695 | 6.89 | 85.87 |
| 5-10 Mil | 135,100 | 28.38 | 72.81 | 951 | 9.42 | 95.30 |
| Below 5 Mil | 129,429 | 27.19 | 100.00 | 474 | 4.70 | 100.00 |

*Notes:* This table shows the number of transactions, the % of transactions, Value and % of Value for transactions by various transaction size buckets. Dollar volume is defined in terms of Billions of 2023 US Dollars.

Table A.5: Transactions By Major Market

| | # Trans | % Trans | $ Vol | % Vol | %A | %I | %O | %R |
|---|---|---|---|---|---|---|---|---|
| **Manhattan** | 12,617 | 2.65 | 733.41 | 7.27 | 26.15 | 0.85 | 63.82 | 9.18 |
| **Los Angeles** | 30,892 | 6.49 | 578.30 | 5.73 | 29.03 | 20.25 | 33.95 | 16.77 |
| Dallas | 18,720 | 3.93 | 448.09 | 4.44 | 44.79 | 18.30 | 24.20 | 12.72 |
| **Chicago** | 19,060 | 4.00 | 405.98 | 4.02 | 21.55 | 24.14 | 36.34 | 17.97 |
| Atlanta | 15,828 | 3.33 | 372.71 | 3.69 | 43.72 | 17.27 | 24.37 | 14.64 |
| Houston | 12,937 | 2.72 | 303.41 | 3.01 | 42.57 | 14.15 | 28.35 | 14.92 |
| **Boston** | 8,268 | 1.74 | 303.20 | 3.00 | 20.36 | 12.95 | 57.92 | 8.78 |
| **Seattle** | 10,744 | 2.26 | 279.32 | 2.77 | 34.78 | 14.64 | 39.27 | 11.30 |
| Phoenix | 13,512 | 2.84 | 277.81 | 2.75 | 46.14 | 16.21 | 22.14 | 15.51 |
| **San Francisco** | 7,561 | 1.59 | 242.48 | 2.40 | 21.18 | 8.48 | 60.49 | 9.85 |
| **DC VA burbs** | 5,051 | 1.06 | 236.12 | 2.34 | 36.26 | 10.96 | 42.38 | 10.40 |
| Northern NJ | 10,114 | 2.12 | 205.36 | 2.03 | 24.83 | 28.42 | 32.81 | 13.94 |
| **San Diego** | 9,332 | 1.96 | 199.01 | 1.97 | 31.34 | 19.69 | 33.75 | 15.22 |
| **San Jose** | 6,280 | 1.32 | 197.36 | 1.96 | 17.02 | 23.26 | 50.56 | 9.15 |
| **Washington DC** | 2,395 | 0.50 | 147.88 | 1.47 | 16.20 | 1.18 | 78.14 | 4.48 |
| **Miami** | 7,239 | 1.52 | 142.94 | 1.42 | 30.99 | 19.78 | 27.49 | 21.74 |
| All Others | 285,472 | 59.97 | 5,019.61 | 49.73 | 36.85 | 20.13 | 21.08 | 21.94 |

*Notes:* This table shows the number of transactions, the % of transactions,volume and share of volume for transactions for the largest 16 markets, ranked by volume, as well as in the remaining 44 markets (last row). The last four columns shows the percentage breakdown of volume by sector (Apartments, Industrial, Office, and Retail) within each market. Volume is expressed in billions of 2023 U.S. dollars. The cities in bold are 11 of the 16 superstar cities. The other five superstar cities that are not explicitly listed in the top of this table are: East Bay, DC MD burbs, Orange Co, NYC Boroughs, and Austin.

Table A.6: Transactions By Transaction Type

| Type | Subtype | # Trans | % Trans | Vol | % Vol |
|---|---|---|---|---|---|
| Conventional Sale | Ind. Asset | 354,712 | 74.53 | 7068.52 | 70.06 |
| | Portfolio | 89,256 | 18.75 | 2087.41 | 20.69 |
| Entity Sale | Private Merger | 1,852 | 0.39 | 61.10 | 0.61 |
| | Privatization | 5,832 | 1.23 | 289.44 | 2.87 |
| | Public Merger | 7,549 | 1.59 | 274.93 | 2.73 |
| | Publicization | 2,098 | 0.44 | 61.28 | 0.61 |
| Foreclosure | Ind. Asset | 7,895 | 1.66 | 128.55 | 1.27 |
| | Portfolio | 1,867 | 0.39 | 22.48 | 0.22 |
| Debtor/Trustee Sale | Ind. Asset | 3,486 | 0.73 | 74.64 | 0.74 |
| | Portfolio | 1,415 | 0.30 | 20.50 | 0.20 |

*Notes:* This table shows the number of transactions, the share of transactions,volume and share of volume by transaction type and subtype. Volume is expressed in billions of 2023 U.S. dollars.

Table A.7: Mapping Investor Types

| RCA Group | RCA Type | Investor Type | %Trans Buy | % Vol Buy | %For Buy | %Trans Sell | % Vol Sell | %For Sell |
|---|---|---|---|---|---|---|---|---|
| CMBS | CMBS | Institutional | 0.90 | 0.71 | 0.68 | 0.93 | 0.62 | 0.43 |
| Equity Fund | Equity Fund | REPE | 6.79 | 12.54 | 2.03 | 5.34 | 10.43 | 1.89 |
| Institutional | Bank | Institutional | 1.25 | 1.83 | 26.65 | 1.75 | 2.15 | 25.37 |
| | Endowment | Institutional | 0.09 | 0.12 | 0.00 | 0.09 | 0.12 | 0.00 |
| | Finance | Institutional | 0.33 | 0.32 | 22.93 | 0.32 | 0.30 | 13.16 |
| | Insurance | Institutional | 0.84 | 1.88 | 25.77 | 0.97 | 2.13 | 14.98 |
| | Investment Manager | Institutional | 5.90 | 11.36 | 19.58 | 5.00 | 8.85 | 17.31 |
| | Open-Ended Fund | Institutional | 0.01 | 0.08 | 98.74 | 0.00 | 0.02 | 94.23 |
| | Pension Fund | Institutional | 1.12 | 2.73 | 33.09 | 1.01 | 2.29 | 21.31 |
| | Sovereign Wealth Fund | Foreign | 0.57 | 0.89 | 100 | 0.22 | 0.30 | 100 |
| Private | Developer/Owner/Operator | Foreign | 0.67 | 0.98 | 100 | 0.45 | 0.68 | 100 |
| | | $DOO_L$ | 29.40 | 12.96 | 0.02 | 32.62 | 16.33 | 0.02 |
| | | $DOO_N$ | 32.07 | 30.43 | 0.00 | 28.24 | 30.44 | 0.00 |
| | High Net Worth | Individual | 1.81 | 1.20 | 22.71 | 2.97 | 1.41 | 7.14 |
| | Non Traded REIT | Individual | 2.37 | 3.19 | 0.06 | 1.22 | 1.64 | 5.09 |
| Public | Listed Funds | REITS | 0.01 | 0.02 | 89.01 | 0.01 | 0.03 | 78.36 |
| | REIT | REITS | 7.09 | 12.27 | 6.66 | 7.34 | 12.93 | 5.60 |
| | REOC | REITS | 0.61 | 0.94 | 72.49 | 0.49 | 1.05 | 64.98 |
| Unknown | Unknown | Unknown | 2.39 | 1.22 | 0.83 | 4.12 | 2.44 | 0.27 |
| User/Other | Cooperative | User | 0.01 | 0.01 | 3.68 | 0.03 | 0.04 | 0.20 |
| | Corporate | User | 4.17 | 3.04 | 6.49 | 5.84 | 4.85 | 6.59 |
| | Educational | User | 0.31 | 0.34 | 0.10 | 0.12 | 0.16 | 0.13 |
| | Government | User | 0.62 | 0.49 | 1.12 | 0.38 | 0.34 | 0.73 |
| | Non-Profit | User | 0.48 | 0.31 | 1.69 | 0.39 | 0.30 | 7.96 |
| | Other | User | 0.06 | 0.06 | 4.33 | 0.03 | 0.03 | 0.00 |
| | Religious | User | 0.14 | 0.07 | 2.10 | 0.11 | 0.10 | 2.54 |

*Notes:* The top table shows the mapping from RCA investor group (column 1) and RCA investor type (column 2) into our investor type (column 3). We show the share of buy transactions and buy volume for each type in columns 4 and 5 and the share of sell transactions and sell volume for each type in columns 7 and 8. Columns 6 and 9 show the share of buy and sell transactions that include a foreign entity. Volume is expressed in billions of 2023 U.S. dollars.

Table A.8: Transactions by Number of Investors

| | 1 Seller | 2 Sellers | 3 Sellers | 4 Sellers |
|---|---|---|---|---|
| 1 Buyer | 87.30 | 5.723 | 0.584 | 0.154 |
| 2 Buyers | 4.73 | 0.654 | 0.099 | 0.020 |
| 3 Buyers | 0.50 | 0.076 | 0.004 | 0.003 |
| 4 Buyers | 0.14 | 0.022 | 0.001 | 0.001 |

| | 1 Seller | 2 Sellers | 3 Sellers | 4 Sellers |
|---|---|---|---|---|
| 1 Buyer | 73.34 | 10.033 | 1.270 | 0.367 |
| 2 Buyers | 10.31 | 1.914 | 0.422 | 0.118 |
| 3 Buyers | 1.38 | 0.424 | 0.026 | 0.014 |
| 4 Buyers | 0.29 | 0.071 | 0.003 | 0.018 |

*Notes:* The left panel splits the number of transactions into sixteen groups, by the number of buyers and the number of sellers involved in each transactions. The numbers in the table add up to 100%. The right panel does the same based on transaction volume, expressed in billions of 2023 U.S. dollars.

Table A.9: Joint Ventures

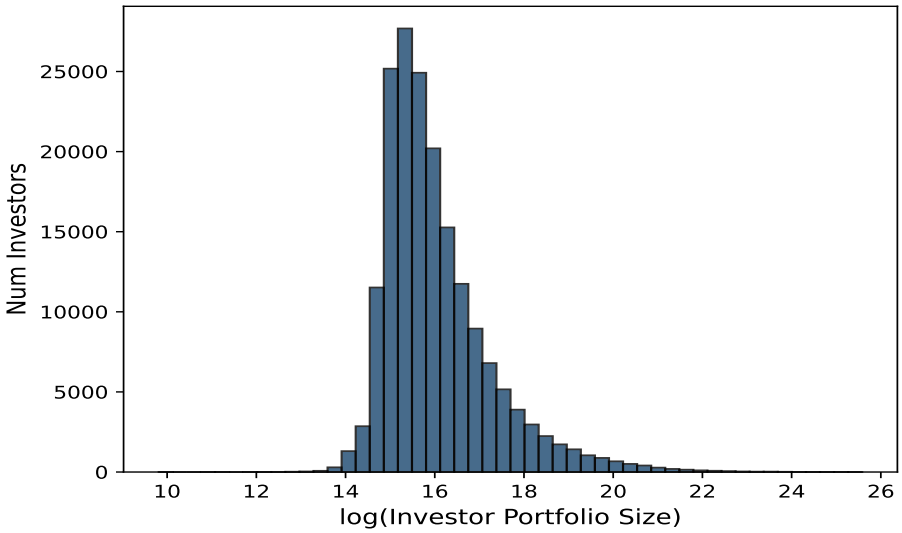
| | REPE | Institutional | $OOD_L$ | $OOD_N$ | Individual | REITS | Foreign | User |
|---|---|---|---|---|---|---|---|---|
| REPE | 2.35 | | | | | | | |
| Institutional | 5.03 | 2.51 | | | | | | |
| $OOD_L$ | 1.22 | 1.12 | 5.2 | | | | | |
| $OOD_N$ | 13.31 | 14.06 | 7.45 | 13.63 | | | | |
| Individual | 0.66 | 1.07 | 1.0 | 1.96 | 0.46 | | | |
| REITS | 1.46 | 2.75 | 0.23 | 2.37 | 1.09 | 0.07 | | |
| Foreign | 1.62 | 3.41 | 0.64 | 6.26 | 0.72 | 1.83 | 4.49 | |
| User | 0.09 | 0.28 | 0.35 | 1.01 | 0.04 | 0.04 | 0.04 | 0.15 |

| | REPE | Institutional | $OOD_L$ | $OOD_N$ | Individual | REITS | Foreign | User |
|---|---|---|---|---|---|---|---|---|
| REPE | 9.12 | 16.63 | 7.10 | 22.16 | 9.40 | 14.84 | 8.54 | 4.68 |
| Institutional | 19.53 | 8.29 | 6.52 | 23.41 | 15.31 | 27.94 | 17.96 | 14.16 |
| $OOD_L$ | 4.74 | 3.71 | 30.24 | 12.41 | 14.32 | 2.28 | 3.36 | 17.36 |
| $OOD_N$ | 51.70 | 46.50 | 43.29 | 22.70 | 28.07 | 24.10 | 32.94 | 50.25 |
| Individual | 2.55 | 3.54 | 5.82 | 3.27 | 6.54 | 11.07 | 3.77 | 1.85 |
| REITS | 5.68 | 9.10 | 1.31 | 3.95 | 15.59 | 0.73 | 9.62 | 2.22 |
| Foreign | 6.31 | 11.29 | 3.71 | 10.43 | 10.25 | 18.58 | 23.59 | 1.97 |
| User | 0.36 | 0.94 | 2.03 | 1.68 | 0.53 | 0.45 | 0.21 | 7.51 |

| | REPE | Institutional | $OOD_L$ | $OOD_N$ | Individual | REITS | Foreign | User |
|---|---|---|---|---|---|---|---|---|
| % Total Buys | 27.34 | 24.77 | 4.43 | 13.99 | 11.66 | 10.73 | 34.16 | 2.42 |

*Notes:* This first panel breaks down the total number of buy transactions involving multiple buyers into who forms a JV with whom. The second panel shows this number as a share of JV property purchases by each investor type; every column sums to 100%. The third panel table shows the share of properties bought in JV with every other investor type, as a share of each investor's total purchases.

Figure A.3: Investor Size Distribution




*Notes:* Figure shows a histogram of log(investor portfolio size) at the end of 2023 for investors that own atleast one property. Log investor size varies from 11.8 ($\sim$ \$135K) to 25.5 ($\sim$ \$119 bi). Not that 16.2 = \$10 mi, 17.7 = \$50 mi, 20.7 = \$1 bi.

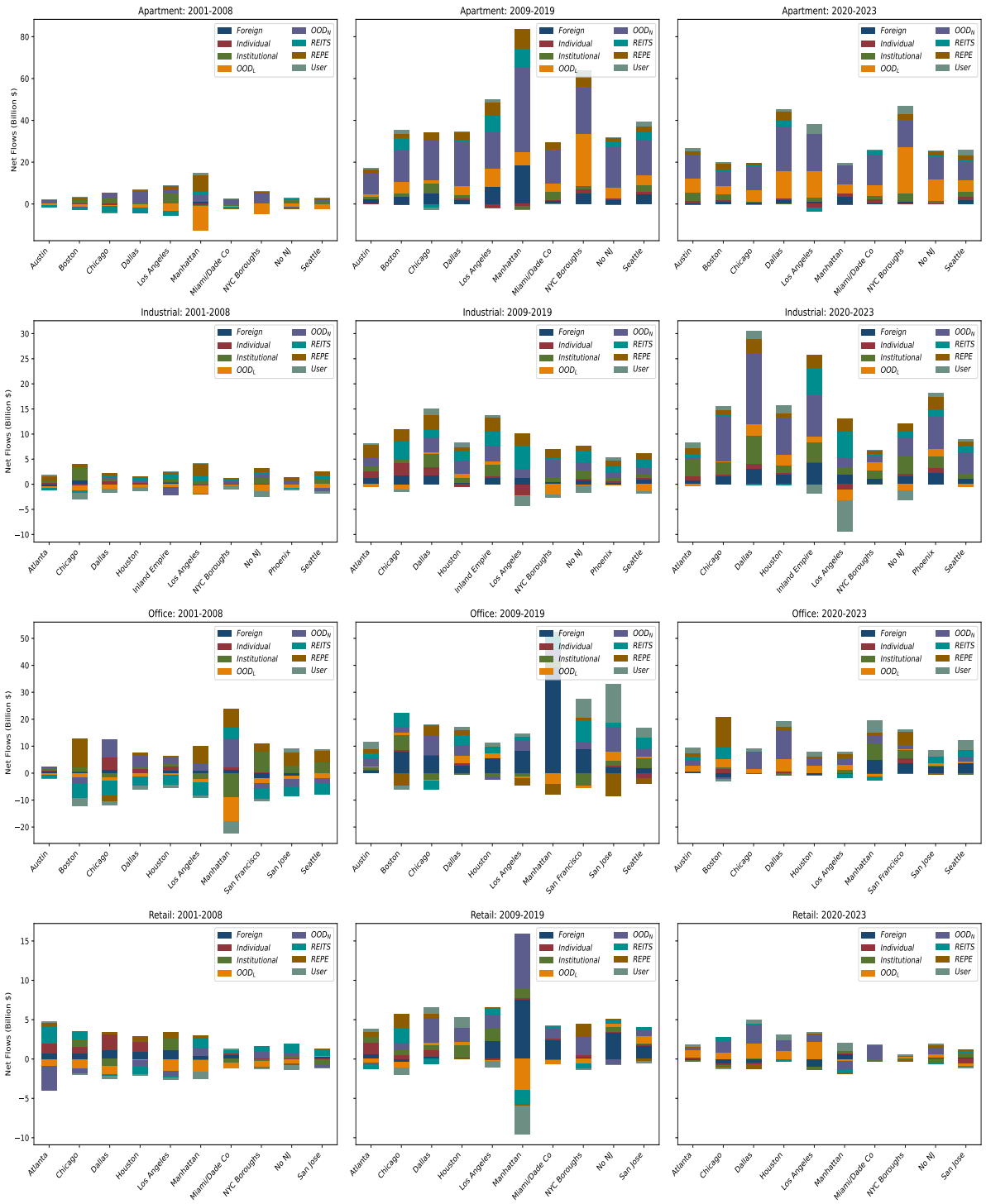
Table A.10: Trade Networks

| | REPE | Institutional | $OOD_L$ | $OOD_N$ | Individual | REITS | Foreign | User | Unknown |
|---|---|---|---|---|---|---|---|---|---|
| REPE | 205.87 | 241.76 | 55.18 | 334.05 | 44.70 | 132.03 | 159.68 | 28.33 | 4.71 |
| Institutional | 243.07 | 321.74 | 82.39 | 461.55 | 56.27 | 192.81 | 176.05 | 42.65 | 7.44 |
| $OOD_L$ | 102.32 | 137.32 | 590.45 | 610.10 | 52.38 | 83.67 | 55.94 | 96.05 | 27.51 |
| $OOD_N$ | 368.63 | 536.94 | 395.71 | 1263.38 | 152.55 | 309.66 | 235.54 | 111.02 | 29.29 |
| Individual | 27.06 | 36.34 | 52.15 | 82.97 | 17.98 | 62.32 | 31.71 | 9.40 | 2.23 |
| REITS | 318.93 | 243.37 | 39.15 | 278.62 | 51.98 | 324.63 | 186.06 | 30.75 | 8.51 |
| Foreign | 145.44 | 131.80 | 29.09 | 164.04 | 32.83 | 71.97 | 123.05 | 25.37 | 1.93 |
| User | 68.04 | 81.48 | 88.76 | 182.84 | 22.79 | 80.48 | 34.79 | 74.60 | 3.90 |
| Unknown | 9.67 | 23.89 | 48.21 | 76.17 | 9.65 | 27.93 | 8.43 | 12.82 | 37.86 |

| | REPE | Institutional | $OOD_L$ | $OOD_N$ | Individual | REITS | Foreign | User | Unknown |
|---|---|---|---|---|---|---|---|---|---|
| REPE | 1.81 | 2.13 | 0.49 | 2.94 | 0.39 | 1.16 | 1.40 | 0.25 | 0.04 |
| Institutional | 2.14 | 2.83 | 0.72 | 4.06 | 0.49 | 1.70 | 1.55 | 0.38 | 0.07 |
| $OOD_L$ | 0.90 | 1.21 | 5.19 | 5.37 | 0.46 | 0.74 | 0.49 | 0.84 | 0.24 |
| $OOD_N$ | 3.24 | 4.72 | 3.48 | 11.11 | 1.34 | 2.72 | 2.07 | 0.98 | 0.26 |
| Individual | 0.24 | 0.32 | 0.46 | 0.73 | 0.16 | 0.55 | 0.28 | 0.08 | 0.02 |
| REITS | 2.80 | 2.14 | 0.34 | 2.45 | 0.46 | 2.85 | 1.64 | 0.27 | 0.07 |
| Foreign | 1.28 | 1.16 | 0.26 | 1.44 | 0.29 | 0.63 | 1.08 | 0.22 | 0.02 |
| User | 0.60 | 0.72 | 0.78 | 1.61 | 0.20 | 0.71 | 0.31 | 0.66 | 0.03 |
| Unknown | 0.09 | 0.21 | 0.42 | 0.67 | 0.08 | 0.25 | 0.07 | 0.11 | 0.33 |

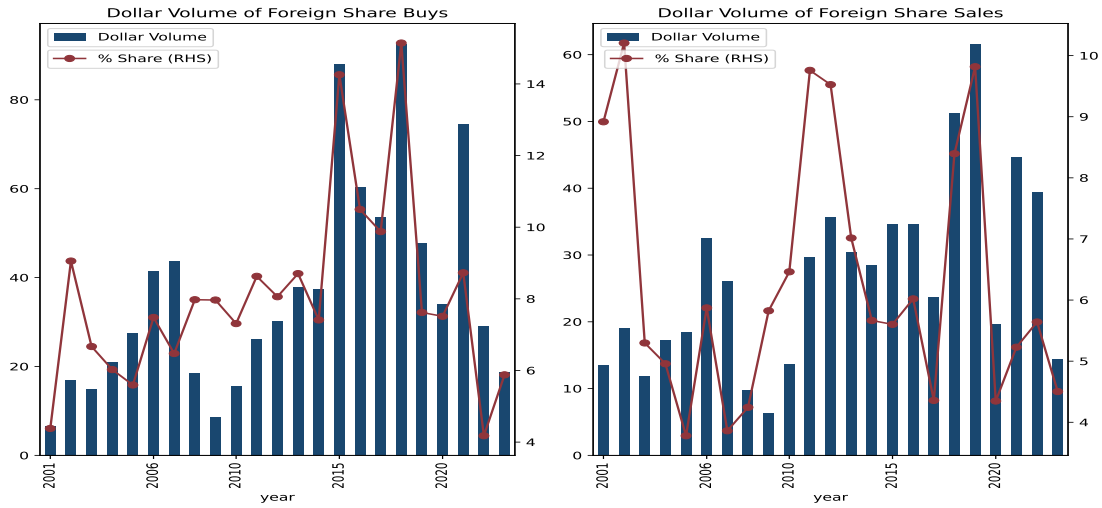
*Notes:* This table summarizes the trade networks among various investor types. We show buyers along the columns and sellers along the rows. Top Panel reports the \$ Volume of transaction between different investor types, expressed in billions of 2023 U.S. dollars. If a transaction involves multiple buyers (sellers) of different types, transaction value is split among them proportionately. The bottom table reports the share of overall transaction \$ Volume. Values in the entire table sum up to 100%

Figure A.4: Net Flows by Investor Types




*Notes:* This figure shows the net acquisitions of properites by investor types and sector for during 2001-2008, 2009-2019 and 2020-2023, calculated as the dollar value of purchases minus sales (in 2023 dollars) for each sector.

Figure A.5: Foreign Investor Share




*Notes:* Figures show foreign investor activity in US Commercial Real Estate Ecosystem during 2001-2023. The blue bars show the dollar value (in Billions, 2023 USD) traded, aggregated by property type while the red lines show the share of foreign investors among all investors. The left panel summarizes foreign investor buys while the right panel summarizes foreign investor sales.

# B Valuation Model Appendix

## B.1 Linear Model with Investor Characteristics

Table B.1: Linear Model with Investor Characteristics

| | Apartment | | Industrial | | Office | | Retail | |
|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| CBD Indicator | 0.120* | 0.098* | 0.248*** | 0.185*** | 0.126*** | 0.094*** | 0.204*** | 0.170*** |
| | (0.058) | (0.053) | (0.054) | (0.044) | (0.038) | (0.033) | (0.042) | (0.039) |
| Age | -0.065*** | -0.059*** | 0.004 | 0.007 | -0.020*** | -0.011** | -0.005 | -0.005 |
| | (0.008) | (0.007) | (0.006) | (0.004) | (0.006) | (0.005) | (0.006) | (0.004) |
| Renovation Adj. Age | -0.026*** | -0.021*** | -0.085*** | -0.068*** | -0.067*** | -0.058*** | -0.087*** | -0.061*** |
| | (0.008) | (0.007) | (0.008) | (0.005) | (0.008) | (0.006) | (0.008) | (0.005) |
| Property Size | -0.233*** | -1.690*** | -0.369*** | -1.898*** | -0.386*** | -2.112*** | -0.456*** | -2.244*** |
| | (0.018) | (0.094) | (0.017) | (0.080) | (0.024) | (0.120) | (0.020) | (0.058) |
| Property Subtype | 0.099*** | 0.135*** | 0.106*** | 0.099*** | | | 0.031 | 0.004 |
| | (0.022) | (0.020) | (0.018) | (0.016) | | | (0.022) | (0.021) |
| No. of Floors | 0.122*** | 0.110*** | 0.047** | 0.064*** | 0.087*** | 0.048*** | 0.050** | 0.035* |
| | (0.014) | (0.012) | (0.019) | (0.013) | (0.008) | (0.006) | (0.020) | (0.017) |
| Occupancy Rate | 0.542*** | 0.449*** | 0.241*** | 0.240*** | 0.699*** | 0.607*** | 0.459*** | 0.409*** |
| | (0.102) | (0.094) | (0.034) | (0.028) | (0.044) | (0.034) | (0.054) | (0.036) |
| Entity Sale | -0.071 | -0.088 | -0.172* | -0.161* | -0.237** | -0.204** | -0.277** | -0.189** |
| | (0.079) | (0.078) | (0.095) | (0.086) | (0.088) | (0.072) | (0.107) | (0.070) |
| Distressed | -0.218*** | -0.175*** | -0.136*** | -0.112*** | -0.164*** | -0.144*** | -0.142*** | -0.100*** |
| | (0.031) | (0.024) | (0.032) | (0.027) | (0.026) | (0.026) | (0.037) | (0.031) |
| Market Occupancy | 0.043 | 0.025 | -0.149* | -0.112* | 0.126 | 0.081 | -0.066 | -0.045 |
| | (0.281) | (0.247) | (0.079) | (0.061) | (0.091) | (0.082) | (0.099) | (0.084) |
| NOI Growth | 0.047 | 0.026 | -0.011 | 0.019 | 0.044 | 0.056* | -0.057 | -0.033 |
| | (0.079) | (0.069) | (0.041) | (0.022) | (0.027) | (0.028) | (0.038) | (0.031) |
| Personal Income | 0.501*** | 0.450*** | 0.226*** | 0.175*** | 0.287*** | 0.217*** | 0.387*** | 0.285*** |
| | (0.067) | (0.059) | (0.054) | (0.043) | (0.049) | (0.042) | (0.037) | (0.029) |
| Population/Employment | 0.011 | -0.001 | 0.004 | 0.004 | -0.019* | -0.021** | 0.036*** | 0.025*** |
| | (0.012) | (0.011) | (0.010) | (0.008) | (0.010) | (0.008) | (0.007) | (0.006) |
| NER | 0.038*** | 0.001 | 0.204*** | -0.628*** | 0.363*** | -0.323** | 0.144*** | -0.403*** |
| | (0.005) | (0.023) | (0.027) | (0.125) | (0.049) | (0.150) | (0.024) | (0.086) |
| Investor Size | 0.146*** | -0.367*** | 0.145*** | -0.394*** | 0.212*** | -0.365*** | 0.187*** | -0.304*** |
| | (0.012) | (0.024) | (0.012) | (0.030) | (0.011) | (0.038) | (0.015) | (0.026) |
| Same Market Share | 0.345*** | -2.159*** | 0.465*** | -4.463*** | 0.506*** | -4.021*** | 0.759*** | -5.297*** |
| | (0.063) | (0.324) | (0.044) | (0.521) | (0.030) | (0.490) | (0.052) | (0.494) |
| Same Sector Share | -0.234*** | -1.644*** | -0.211*** | 0.791* | 0.428*** | -0.286 | 0.056 | 1.336*** |
| | (0.075) | (0.287) | (0.038) | (0.415) | (0.040) | (0.390) | (0.058) | (0.370) |
| Superstarcity Share | 0.117** | -0.045 | 0.048 | -0.035 | -0.033 | -0.041* | 0.127*** | -0.005 |
| | (0.047) | (0.046) | (0.032) | (0.022) | (0.039) | (0.023) | (0.041) | (0.030) |
| JV | 0.085** | 0.113*** | -0.049 | -0.030 | 0.166*** | 0.154*** | 0.148** | 0.094* |
| | (0.031) | (0.023) | (0.054) | (0.044) | (0.025) | (0.023) | (0.068) | (0.046) |
| Investor Size × Property Size | | 0.082*** | | 0.084*** | | 0.087*** | | 0.093*** |
| | | (0.005) | | (0.004) | | (0.004) | | (0.003) |
| Investor Size × Same Market Share | | 0.141*** | | 0.273*** | | 0.244*** | | 0.333*** |
| | | (0.018) | | (0.031) | | (0.028) | | (0.028) |
| Investor Size × Same Sector Share | | 0.071*** | | -0.043** | | 0.036* | | -0.061*** |
| | | (0.013) | | (0.020) | | (0.020) | | (0.020) |
| Investor Size × NER | | 0.004** | | 0.052*** | | 0.030*** | | 0.030*** |
| | | (0.002) | | (0.008) | | (0.009) | | (0.006) |
| Property Size × NER | | -0.008*** | | -0.035*** | | 0.013 | | -0.012 |
| | | (0.002) | | (0.012) | | (0.018) | | (0.010) |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Market FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Investor Type FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 141,135 | 141,135 | 116,737 | 116,737 | 96,139 | 96,139 | 114,223 | 114,223 |
| Adj. $R^2$ | 65.1 | 68.9 | 64.1 | 71.3 | 57.7 | 66.0 | 64.3 | 73.0 |

*Notes:* This table shows the full-sample coefficient estimates and goodness of fit statistics for a linear valuation model with investor characteristics and select interactions. Standard errors in bracket are clustered by RCA Market and Year.

## B.2 Light GBM Hyperparameter tuning and Model Complexity

LightGBM (Guolin Ke, 2017) is a gradient-boosted decision tree algorithm designed for efficiency and scalability. Unlike traditional tree-based models, it uses histogram-based learning to speed up training and leaf-wise tree growth to capture complex patterns in the data. It can capture complex, non-linear relationships without requiring explicit interaction terms, automatically select relevant features and interactions, improving model interpretability and regularize effectively to prevent overfitting.

We fine-tune LightGBM hyperparameters for each model specification using Optuna (Takuya Akiba and Koyam, 2019), a Bayesian optimization framework. Hyperparameters are selected via 10-fold cross-validation, minimizing mean squared error (MSE) on the validation folds. Each model undergoes 50 Optuna trials, and the configuration with the lowest average validation MSE is retained for out-of-sample evaluation. The search space is deliberately bounded to avoid overly complex models that may overfit. We vary the number of estimators between 500 and 1,000 in increments of 10, and the maximum tree depth between 2 and 5 for our baseline models. The number of leaves is constrained by the chosen depth, ensuring trees remain interpretable and regularized. Minimum child samples vary from 1% to 10% of the number of observations, in increments of 50, to control leaf-level granularity. The feature and bagging fractions range from 0.5 to 1.0 in steps of 0.01, encouraging subsampling for robustness. L1 and L2 regularization coefficients are searched between 0.0 and 5.0 in increments of 0.1, while the minimum split gain is restricted to 0.2–1.0 to discourage splits with marginal predictive value, thus keeping the maximum depth under control. The learning rate is tuned between 0.01 and 0.1 on a log scale.

Restricting tree depth to 2–5 in the baseline search produces trees that capture main effects and low-order interactions without memorizing idiosyncratic patterns in the training data. In the "deep" LightGBM variant, we relax this bound and allow max depth to be selected by cross-validation without restriction, resulting in depths of 7–11. While deeper trees can model more complex nonlinearities and higher-order interactions, they carry a greater risk of overfitting, especially in settings with persistent residual correlation. Our design mitigates this by combining depth constraints (in the baseline), strong regularization, feature and row subsampling, and a relatively high min child samples threshold, all of which reduce variance and improve stability across folds. Together, these choices ensure that any gains from increased flexibility in the deep model reflect genuine predictive signal rather than noise-fitting.

Tables B.2 and B.3 present the hyperparameter search spaces and optimized values for the final LightGBM models, i.e. those with investor characteristics.

We evaluate the effect of increasing model complexity on valuation performance by gradually expanding the set of predictors and allowing for progressively richer functional forms. Figure B.1 shows the in-sample $R^2$ as we move from a linear baseline hedonic model with only asset and transaction characteristics to specifications that incorporate macroeconomic variables, time and market fixed effects, investor characteristics, non-linear transformations of continuous predictors, and select interaction terms. This staged build-up illustrates how much explanatory power is attributable to each successive modeling enhancement.

This shows that moving from a linear model to then allowing for more flexible functional forms—first through targeted transformations and selected interactions, then via a fully non-linear LightGBM specification, substantially improves the fit. The final step in our complexity ladder is the "deep LightGBM" model, in which the maximum tree depth is determined by cross-validation without restriction.

Increasing model complexity comes with important trade-offs. More complex models, particularly those with deeper trees, tend to achieve higher in-sample $R^2$ but at the risk of fitting patterns that do not generalize out-of-sample. In our max-depth analysis (unreported), the marginal gains in in-sample $R^2$ beyond depths

Table B.2: LightGBM Hyperparameters: Buyer Model

| | Search Space | Apartment | Industrial | Office | Retail |
|---|---|---|---|---|---|
| Number of Estimators | 500 - 1000 | 750 | 860 | 750 | 600 |
| Maximum Tree Depth | 2 - 5 | 5 | 5 | 4 | 4 |
| Number of Leaves | 3 - $2^{MaxDepth}$ | 27 | 27 | 13 | 11 |
| Minimum Child Samples | 1% Obs - 10% Obs | 1410 | 1170 | 1360 | 1140 |
| Feature Fraction | 0.5 - 1.0 | 0.76 | 0.74 | 0.87 | 0.56 |
| Bagging Fraction | 0.5 - 1.0 | 0.89 | 0.62 | 0.95 | 0.54 |
| L1 Regularization | 0.0 - 5.0 | 3.60 | 0.80 | 0.30 | 0.80 |
| L2 Regularization | 0.0 - 5.0 | 2.10 | 0.50 | 2.80 | 1.80 |
| Minimum Split Gain | 0.2 - 1.0 | 0.44 | 0.37 | 0.29 | 0.45 |
| Learning Rate | 0.01 - 0.1 | 0.08 | 0.06 | 0.06 | 0.06 |
| Number of Obs | | 141,135 | 116,737 | 96,139 | 114,223 |

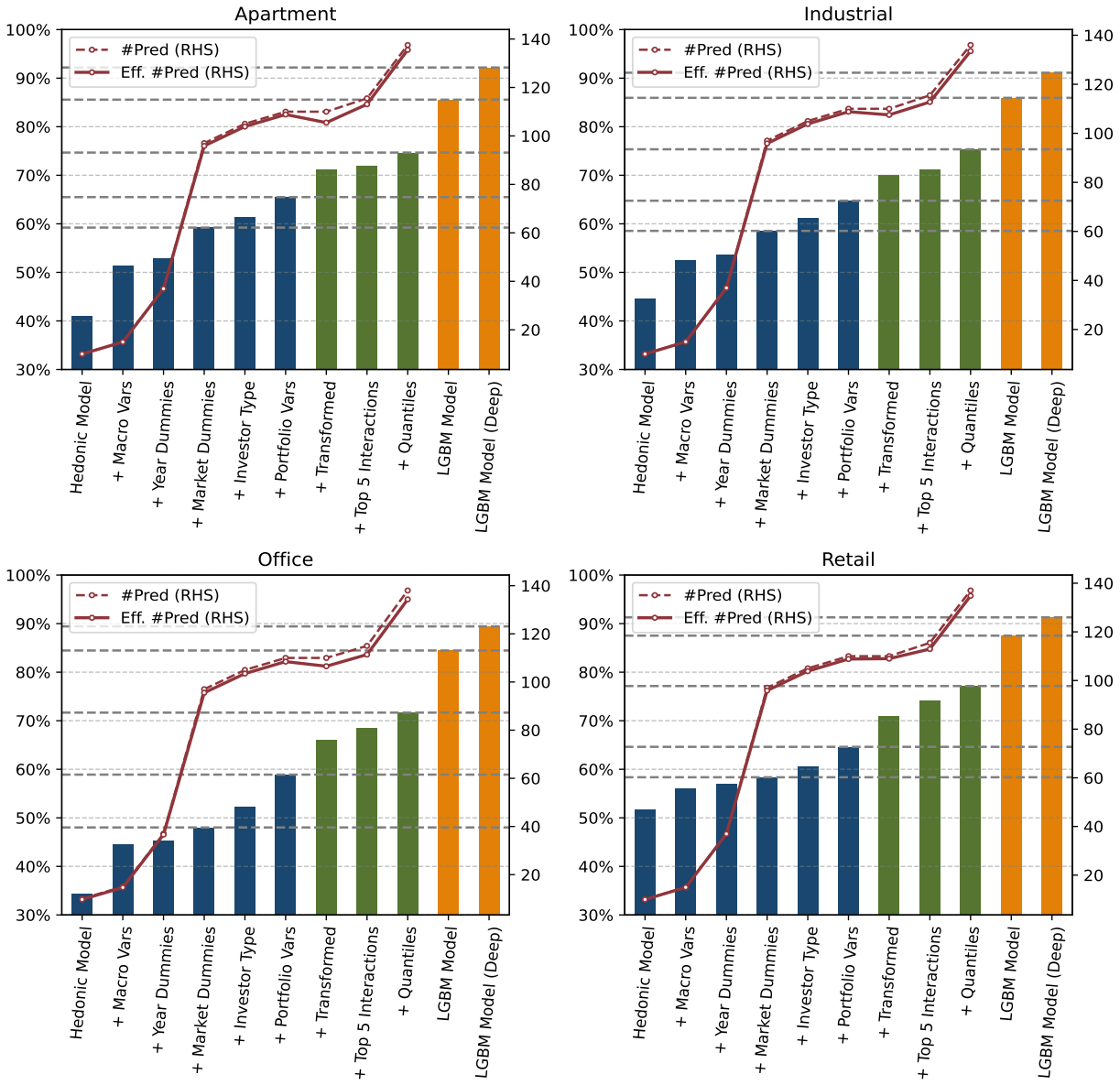
Table B.3: LightGBM Hyperparameters: Seller Model

| | Search Space | Apartment | Industrial | Office | Retail |
|---|---|---|---|---|---|
| Number of Estimators | 500 - 1000 | 930 | 670 | 600 | 870 |
| Maximum Tree Depth | 2 - 5 | 5 | 5 | 4 | 5 |
| Number of Leaves | 3 - $2^{MaxDepth}$ | 20 | 29 | 15 | 23 |
| Minimum Child Samples | 1% Obs - 10% Obs | 2360 | 1170 | 1280 | 1140 |
| Feature Fraction | 0.5 - 1.0 | 0.73 | 0.64 | 0.89 | 0.85 |
| Bagging Fraction | 0.5 - 1.0 | 0.66 | 0.65 | 0.72 | 0.73 |
| L1 Regularization | 0.0 - 5.0 | 1.90 | 1.70 | 0.20 | 3.50 |
| L2 Regularization | 0.0 - 5.0 | 1.70 | 1.60 | 1.40 | 0.80 |
| Minimum Split Gain | 0.2 - 1.0 | 0.26 | 0.88 | 0.34 | 0.27 |
| Learning Rate | 0.01 - 0.1 | 0.05 | 0.03 | 0.07 | 0.06 |
| Number of Obs | | 141,135 | 116,737 | 96,139 | 114,223 |

of 4–5 are modest, while the gap between in-sample and out-of-sample $R^2$ widens. This widening gap reflects the classic bias-variance trade-off: deeper models reduce bias by capturing more intricate patterns but increase variance by fitting idiosyncrasies in the training data.

In our setting, where both asset characteristics and investor attributes can exhibit persistence and cross-sectional correlation, these safeguards are important. A model that is too shallow may fail to capture economically important interaction effects, but a model that is too deep may overfit to transient, investor-specific shocks embedded in the training sample. The optimal balance depends on the objective of the model, forecasting horizon and the stability of the underlying relationships: for short-horizon prediction, some residual persistence can legitimately be exploited; for structural interpretation, more conservative complexity controls are required.

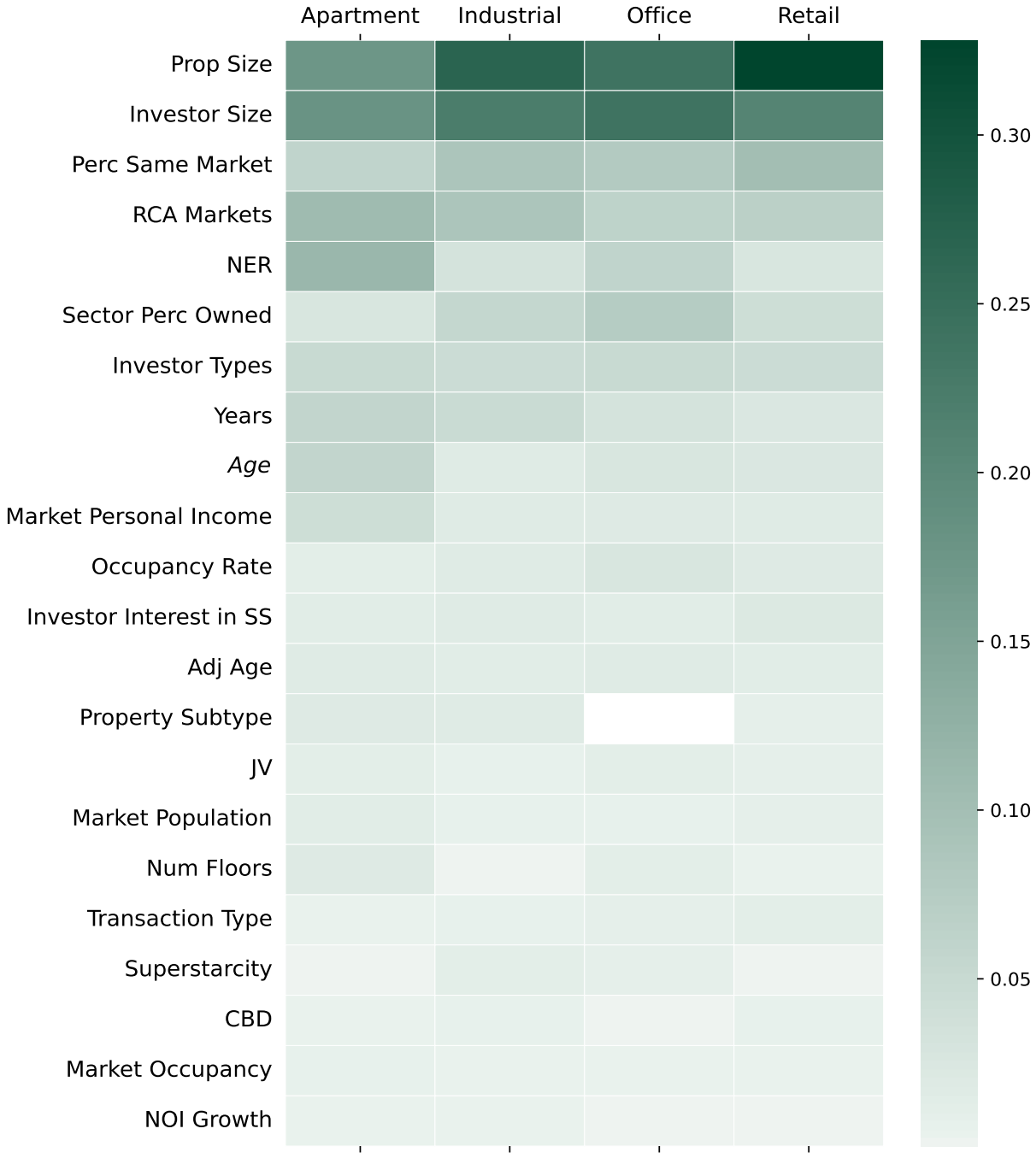
Figure B.1: Incremental $R^2$ from Sequential Model Enhancements




*Notes:* Figure shows the : In-sample $R^2$ for all sectors as we sequentially expand the set of predictors, starting from the baseline hedonic model containing only asset and transaction characteristics (the first blue bar). The next five blue bars add variables step-by-step: macro variables, year and market fixed effects, investor type variables, and investor portfolio variables (see sections 5.2–5.3), but retain the linear setting. The "+ Transformed" bar replaces continuous predictors with their best-fit polynomial transformations, as implied by the LightGBM SHAP dependence plots in Figure 6. The next green bar ("+ Top 5 Interactions") adds the five most important interaction terms identified in Figure 7. Given the prominence of investor size, property size, and their interaction, the following green bar ("+ Quantiles") includes quintile dummies for investor size and property size and their interaction, while dropping terms eliminated by a simple LASSO selection. The final two yellow bars report $R^2$ for our full LightGBM non-linear model and for the "deep LightGBM" specification as described in section B.2 where the maximum tree depth is selected via cross-validation without restriction. All $R^2$ values are obtained from ridge regressions to ensure stable coefficient estimates in the presence of multicollinearity. The solid red line plots the effective number of predictors-calculated from the singular values of the standardized design matrix as $\sum s^2/\left(s^2+\alpha\right)$-while the dotted red line shows the actual number of predictors included at each stage.

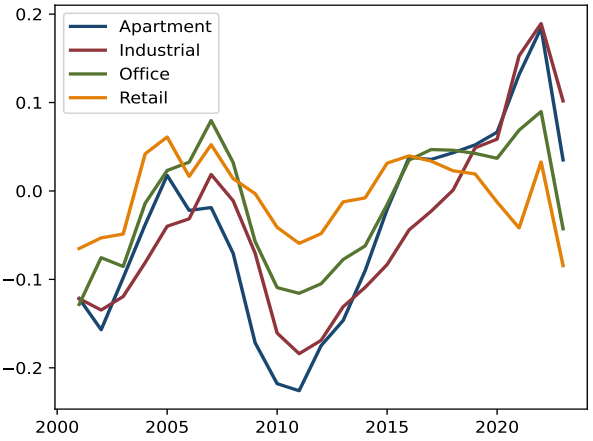
## B.3 Valuation Model: Variable Importance

Figure B.2: LightGBM Feature Importance



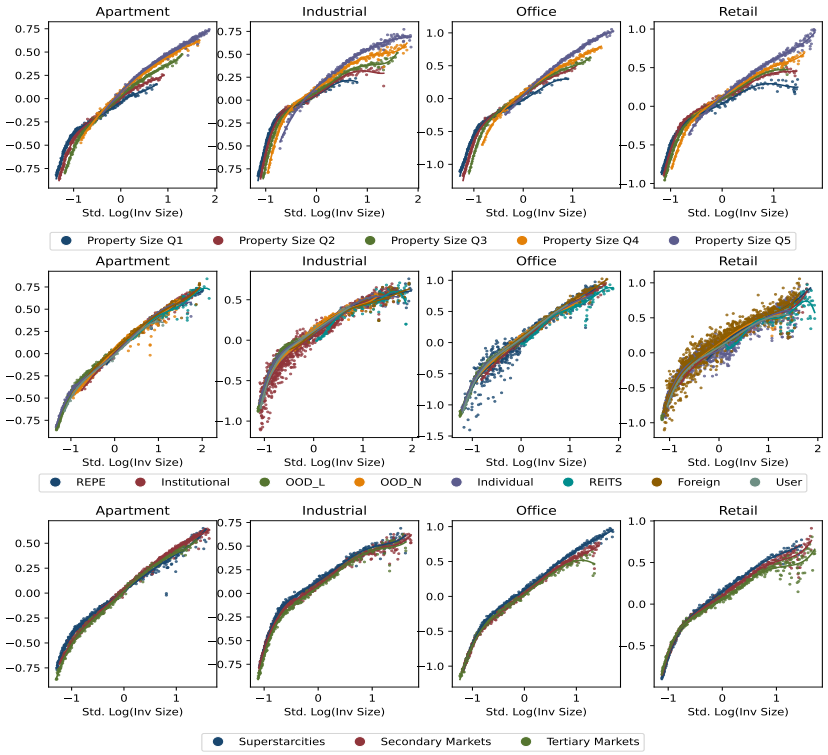

*Notes:* Figure shows the importance of features selected by the LightGBM Model. Feature importance is defined as the absolute SHAP (Shapley Additive Explanations) values for each feature summed across all observations and is normalized such that the importance for each feature sums to 1. Features are sorted by the average importance across sectors. Investor types are treated as a single group.

Figure B.3: Year Fixed Effects in the Valuation Model



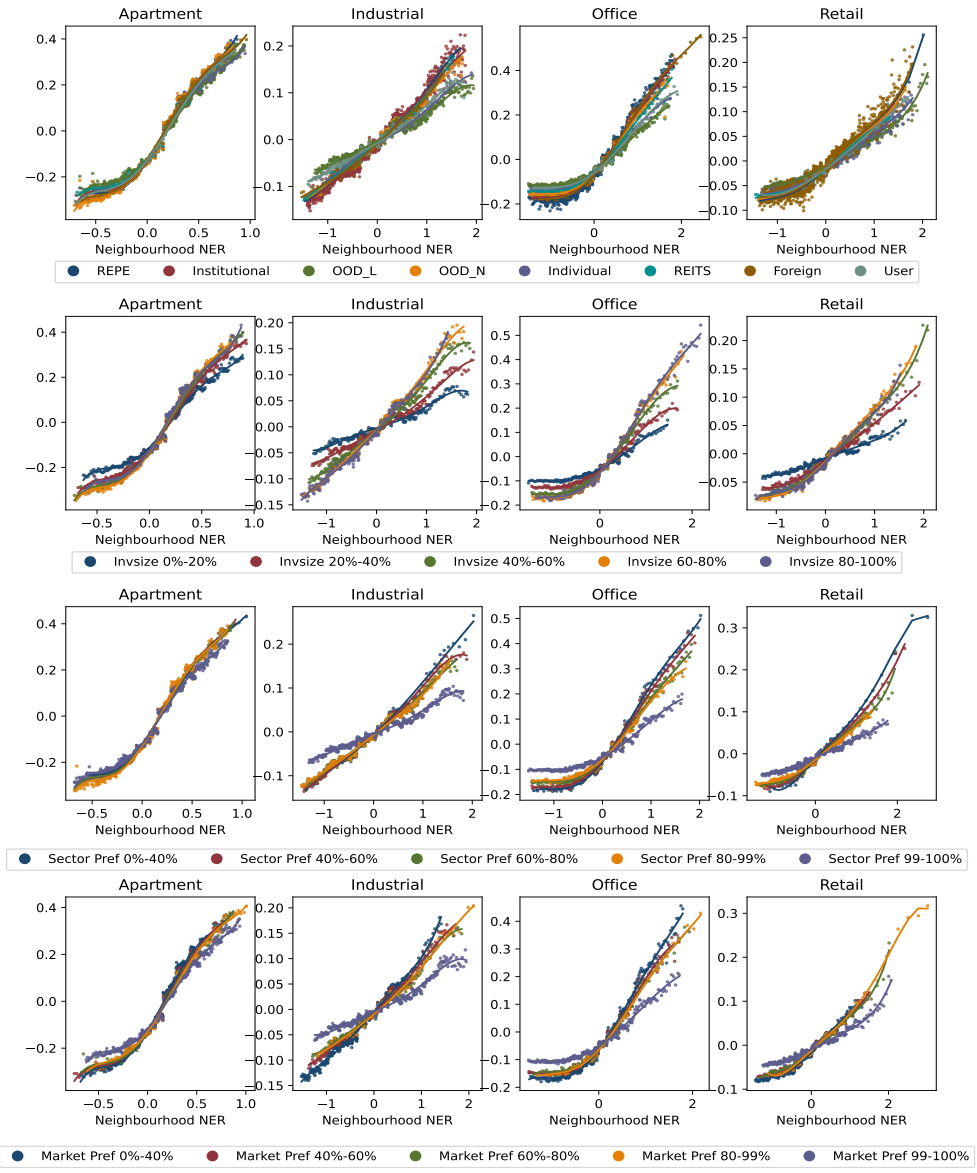

*Notes:* This figure This figure plots the average SHAP values for the year categorical variable. SHAP values are averaged across buyer and seller models and standardized to ensure comparability across sectors.

Figure B.4: SHAP Dependence: Interactions effects with Investor Size




*Notes:* This figure shows how the effect of investor size on property values depends on property size (top row), investor type (second row), or market (third row).

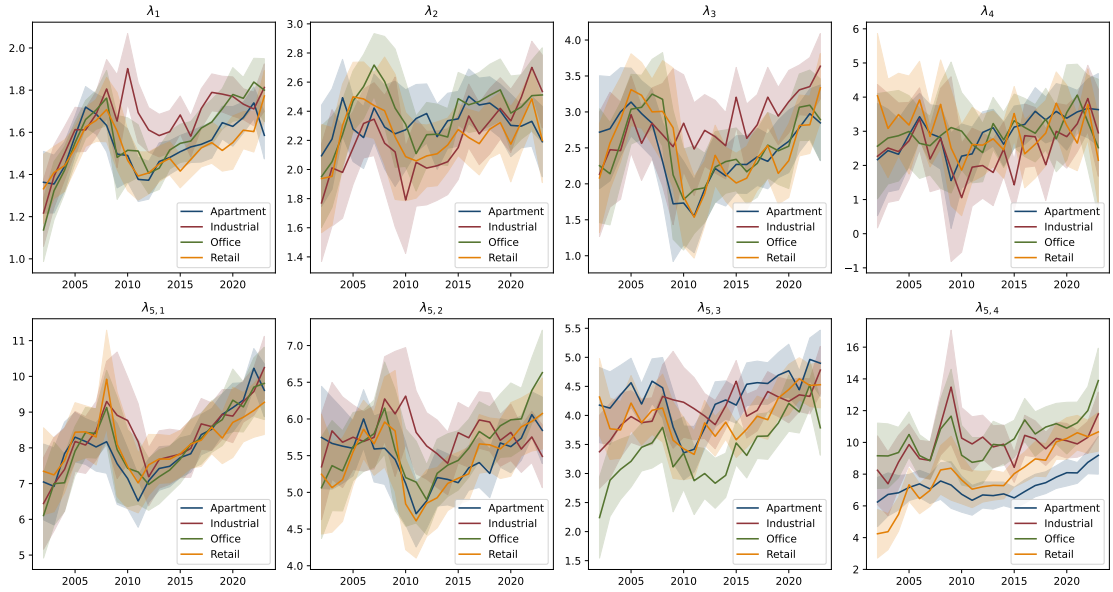
Figure B.5: SHAP Dependence: Interaction effects with Net Effective Rent




*Notes:* This figure shows how the effect of NER depends on investor type (top row) or investor size (second row), investor sector concentration (third row) and investor market concentration (fourth row).

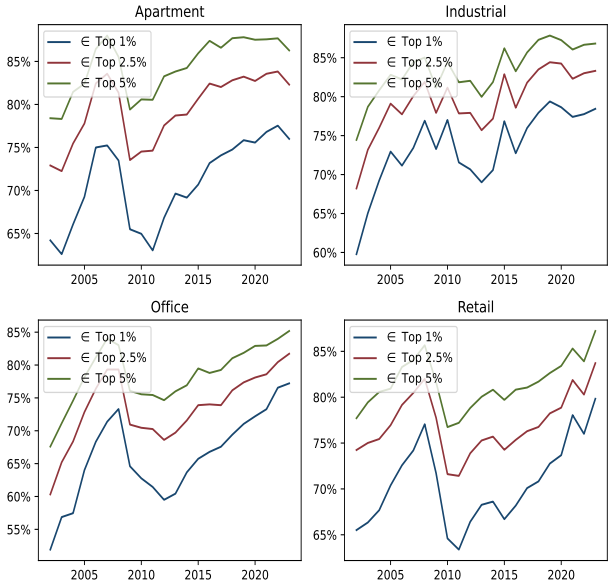
# C Meeting Model Appendix

Figure C.1: Meeting Model Parameters by year




*Notes:* This figure shows, the estimated values of the annual meeting model parameters (shaded area shows their confidence intervals). The point estimates $\hat{\lambda}_t$ are obtained by maximum likelihood for each year and standard errors are calculated using bootstrap (with 1,000 samples) for each year. The maximum likelihood estimation minimizes the loss function in (7). $\lambda_1$ is the coefficient of investor size, $\lambda_2$ is the coefficient of the indicator that equals one when the buyer owns more than 2 assets, $\lambda_3$ is the coefficient on buyer-seller size difference. $\lambda_4$ is the coefficient of the indicator that equals one when the buyer and seller have a prior transactional relationship. $\lambda_{5,1}$ through $\lambda_{5,4}$ are the coefficients of asset, market, sector and NER alignment.

Figure C.2: Meeting Probabilities: True Buyer in Top Predicted Matches

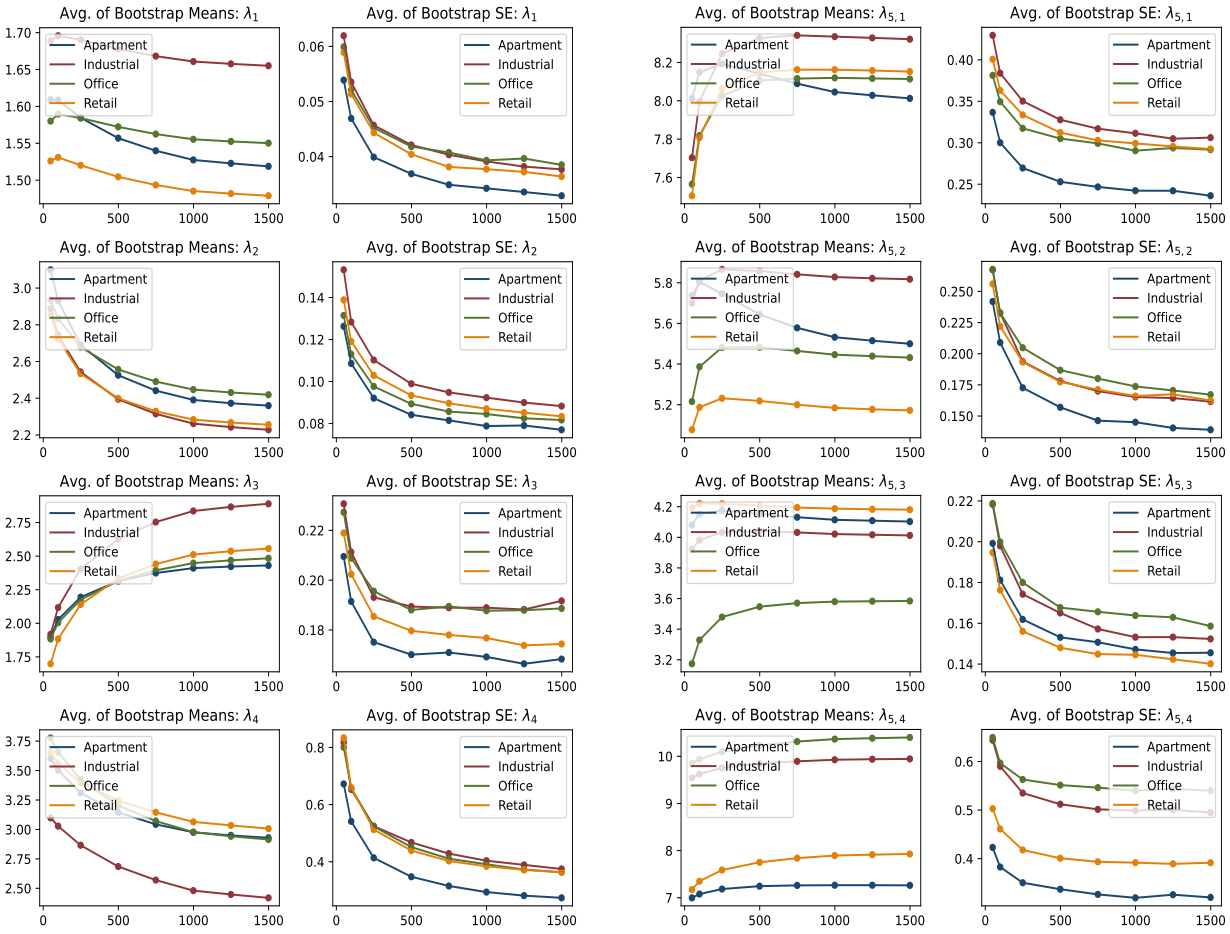



*Notes:* This figure reports the probability that the true buyer ranks among the top 1%, 2.5%, or 5% of the 1,000 negative sample buyers—i.e., among the top 10, 25, or 50 buyers respectively—based on model-implied meeting probabilities.

## C.1 Sensitivity to the Size of the Negative Sample

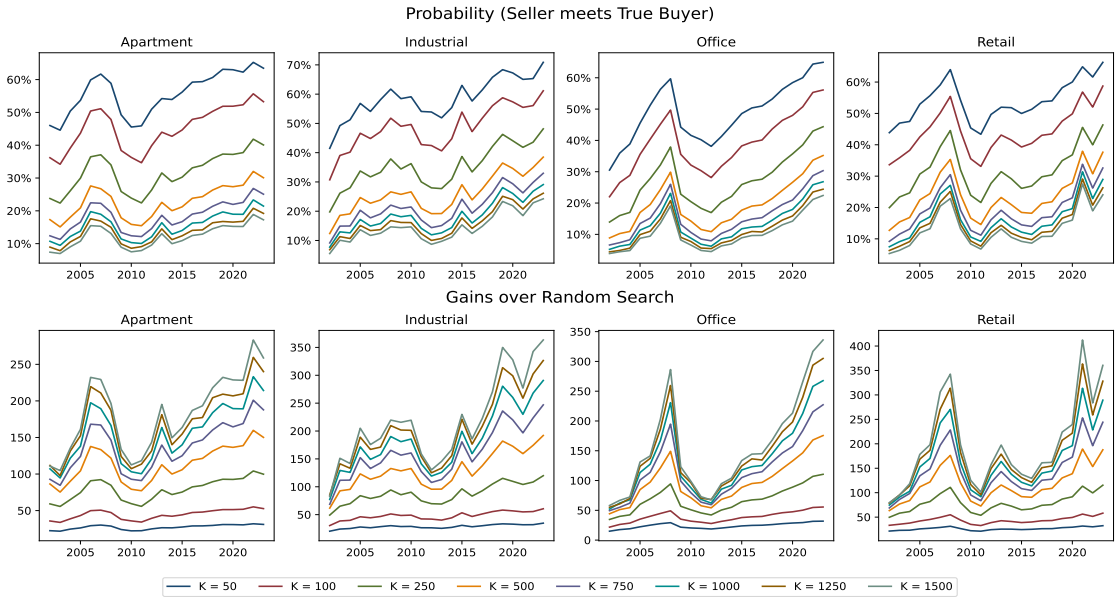
As discussed in Section 3.2, rather than including all potential buyers in the likelihood estimation, we use an important insight from the NLP literature showing it is sufficient to consider a negative set with small cardinality $K$. In this section, we verify that the parameter values and standard errors stabilize for K = 1,000. We test this by using bootstrap with 1,000 samples for various values of K between 50 and 1,500.

Figure C.3: Bootstrap Parameter Means and Standard Errors



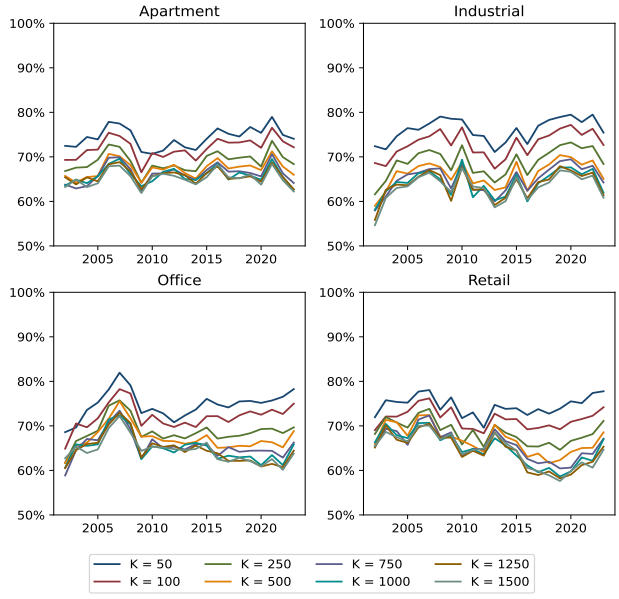

*Notes:* This figure shows, for each parameter in the matching model, the point estimates (left column) and the standard errors (right column) for varying values of K, the size of the negative sample. The point estimates $\hat{\lambda}_t$ and standard errors are calculated using bootstrap (with 1,000 samples) for each year, and averaged across years. The MLE minimizes the loss function in (7). $\lambda_1$ and $\lambda_2$ are coefficients of investor size and buyer-seller size difference respectively. $\lambda_{3,1}$ through $\lambda_{3,4}$ are the coefficients of asset, market, sector and NER alignment an $\lambda_3$ is the coefficient of the indicator that equals one when the buyer owns $> 2$ assets.

Figure C.4: Probability of Meeting the True Buyer by K




*Notes:* This figure shows, the likelihood that the model identifies the true buyer from among the set of K potential buyers and the true buyer. The raw probabilities for each sector and K are shown in the top Panel. Since the probability of meeting the true buyer under random search varies with the size of K, in the bottom panel, we normalize the raw probability by the probability under random matching. The corresponding ratio thus gives the gains from directed search for various sizes of the negative sample. The incremental gains from increasing K beyond 1,000 are small.

Figure C.5: Probability of Transaction (Conditional on Listing) by K




*Notes:* This figure shows, the model-implied probability of a transaction occurring with any buyer as we vary the size of the negative sample. This stabilizes around K = 1,000.

## C.2 Robustness: Meeting Model with All Investors

While calibrating the meeting model in Section 6, we restrict the negative sample to those investors who have transacted at least once in the previous five years to contain the number if inactive investors. In this robustness check, we relax that assumption.

Table C.1 presents the updated parameter estimates. The coefficients retain the same signs and exhibit comparable magnitudes as in Table 4, reinforcing the robustness of our results to the negative sample restriction.

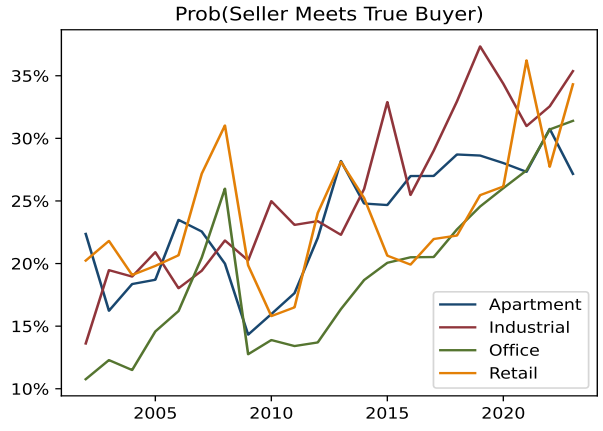
Table C.1: Meeting Model Parameter Estimates

| | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_{5,1}$ | $\lambda_{5,2}$ | $\lambda_{5,3}$ | $\lambda_{5,4}$ |
|---|---|---|---|---|---|---|---|---|
| Apartment | 1.8 | 2.86 | 2.53 | 3.38 | 9.03 | 5.85 | 4.69 | 7.87 |
| | (0.03) | (0.08) | (0.17) | (0.31) | (0.24) | (0.15) | (0.15) | (0.33) |
| Industrial | 1.92 | 2.71 | 2.88 | 2.96 | 9.33 | 6.17 | 4.29 | 10.8 |
| | (0.04) | (0.09) | (0.19) | (0.42) | (0.32) | (0.16) | (0.16) | (0.5) |
| Office | 1.82 | 2.99 | 2.59 | 3.48 | 8.95 | 6.01 | 3.6 | 11.12 |
| | (0.04) | (0.09) | (0.19) | (0.41) | (0.3) | (0.18) | (0.16) | (0.54) |
| Retail | 1.75 | 2.8 | 2.55 | 3.52 | 8.98 | 5.78 | 4.23 | 8.54 |
| | (0.04) | (0.09) | (0.18) | (0.4) | (0.3) | (0.17) | (0.15) | (0.4) |

*Notes:* This table reports the average estimated values of the meeting model parameters along with their standard errors in bracket. The maximum likelihood estimation minimizes the loss function in (7). The parameter interpretation is as in Table 4

Figure C.6 illustrates the model's ability to identify the true buyer from the set of potential buyers under this alternative specification. Since the relaxed negative sample includes a larger share of infrequent traders—who are less likely to transact—the model performs better at matching sellers with actual buyers. The probability of correctly identifying the buyer rises to 25-30%, compared to a little over 20% in the base model, as shown in figure 9).

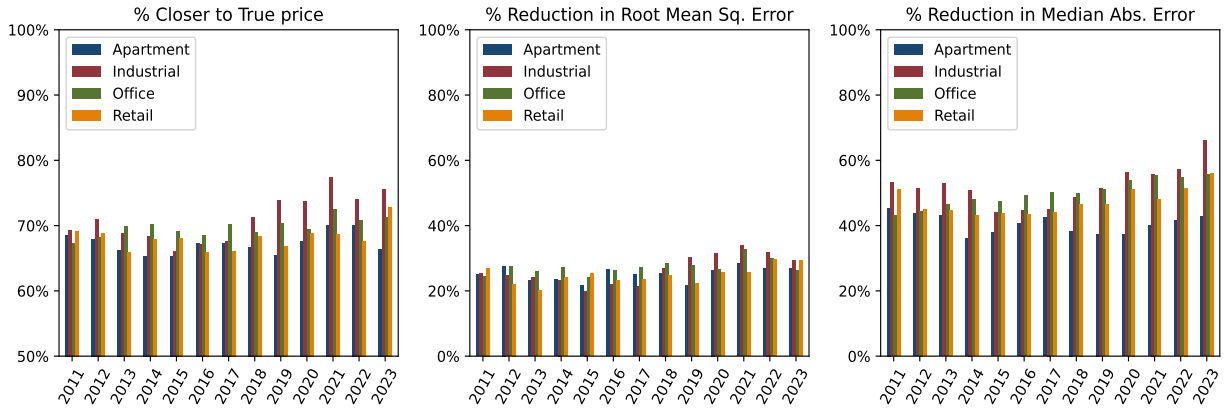
Figure C.6: Meeting Probabilities




*Notes:* This figure shows the probability of meeting the actual buyer from among a set of $K = 1,000$ potential buyers (negative set) and the actual buyer, as implied by the meeting model described in Section 6 with the change that the negative sample now includes all investors and not just those that transacted in the previous five years.

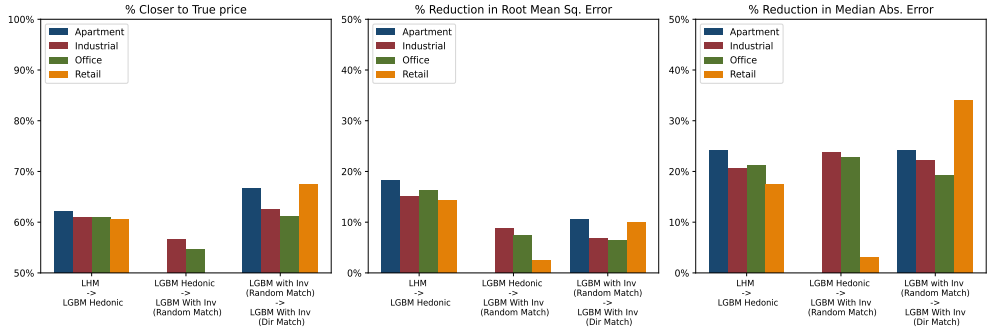
# D Out of Sample Prediction Appendix

Figure D.1: Out of Sample performance




*Notes:* This figure shows the relative out-of-sample performance of the LGBM model relative to the linear hedonic model (LHM). The left chart reports the fraction of transactions for which the LGBM prediction is closer to the true transaction price than the LHM prediction. The middle chart reports the percent reduction in root mean squared prediction error of the LGBM relative to the LHM model. The right chart reports the percent reduction in median absolute prediction error of the LGBM relative to the LHM model. In each case, the prediction model is always estimated with information up until time $t$ in order to predict the price at $t+1$. Further, it assumes that $(x_{t+1}, z_{t+1})$ must both be predicted, and we use our directed matching model to identify the potential buyers.

Figure D.2: LGBM Predictive performance improvement breakdown




*Notes:* Figure 13 compares the out-of-sample performance of the LGBM model to the linear hedonic model (LHM), breaking the gains into three components. The left bars show the improvement from moving from the LHM to a non-linear hedonic LGBM. The middle bars capture the additional gain from adding investor characteristics to the LGBM, with potential buyers drawn via random matching. The right bars show the further improvement when buyers are identified using our directed matching model. All models are estimated with information available at time $t$ to predict prices at $t+1$