How Easily Will It Sell? A Bayesian Two-Stage Proposal–Choice Ranking Model for Real Estate Properties

◯ Hao Wang[1], Koken Ozaki[2], Patrik Andersson[1], Kosuke Fukunaka[1]

[1] GA technologies Co., Ltd., Japan
[2] University of Tsukuba, Japan

Abstract

In the real estate buy-and-resell business, delayed sales are not always caused by low intrinsic attractiveness; limited proposal opportunities can also play an important role. Conventional approaches often link sales outcomes directly to listing attributes, leading to biased estimates of sale difficulty. This study proposes a two-stage framework that models listing sales as proposals followed by transactions. In Stage 1, a LightGBM model predicts whether a listing is proposed. In Stage 2, transactions are modeled using a Bayesian discrete choice model estimated via Markov Chain Monte Carlo. Results show that separating proposal and transaction processes provides interpretations that better align with operational workflows.

1 Introduction

Evaluating how easy a property is to sell is a major challenge in real estate research. In practice, buyers do not search through the entire market inventory; instead, they make decisions based on a small "consideration set" of properties [Han and Strange, 2015]. This is because the market is not transparent; it is a "high-friction market" filled with significant search costs and matching difficulties [Wheaton, 1990]. In such an environment, buyers' consideration sets are further shaped by real estate brokers. Leveraging significant information advantages, brokers influence which listings enter the buyer's consideration set.

Similar issues exist in the Japanese market. Previous studies on the Japanese housing market [Shimizu et al., 2004, Shimizu, 2010, Shimizu, 2011] have systematically revealed the significant information asymmetry in this market. On this basis, Saito [Saito, 2015] pointed out that in existing-home transactions, buyers' actual consideration sets are largely determined by brokers' information screening. Combined with the findings of Levitt and Syverson[Levitt and Syverson, 2008], it can be inferred that brokers are not neutral information conduits but instead adjust their proposal behavior based on internal incentives. From this, we infer that the set of properties finally presented to buyers is not randomly generated but is a filtered, non-random subset of the entire inventory. Therefore, whether a property ultimately sells depends not only on its own quality and attributes but also on whether it successfully enters the buyer's consideration set.

While previous research has identified such sample selection bias [Jud and Seaks, 1994], empirical studies are often constrained by a lack of process data. Accurate inference requires observing the complete decision path from the full inventory to the final transaction: (1) the *proposal stage*, where properties are selected from the inventory and proposed by brokers; and (2) the *sale stage*, where buyers make a choice from the consideration set. However, most existing studies only observe the final transaction outcome, treating the intermediate proposal stage as a black box [Do and Grudnitski, 2004]. This missing data situation makes it impossible to distinguish whether a property failed to sell due to a lack of attractiveness (rejected by the buyer) or because it never entered the buyer's consideration set (filtered by the broker). Failure to account for this "first-stage screening logic" is the root cause of estimation bias in traditional models, as parameters are estimated on a non-randomly selected sample [Heckman, 1979].

We address this gap using actual proposal-session data recorded by a large real estate resale company. We propose a two-stage Bayesian framework that structurally decomposes the sales process: Stage 1 models the broker's proposal probability, and Stage 2 models the probability of a transaction occurring within a given proposal set. Because buyers' latent consideration sets are not directly observable in operational data, the observed proposal records are treated as an empirical approximation of buyers' effective consideration sets.

The proposed framework is designed to mitigate estimation bias through the following mechanism. Traditional single-stage models attempt to explain sales

directly through property attributes, causing property-feature parameters to absorb broker-side proposal preferences. By explicitly modeling the proposal stage, the framework isolates broker-side proposal behavior as an independent stochastic process. In the second stage, buyer preferences are analyzed specifically within the proposed subsample. Based on conditional probability theory, this two-stage decomposition probabilistic disentangles broker-side screening effects from the estimation of buyer-side preference utility, thereby reducing contamination from broker-side selection effects.

In the present study, the primary objective is not to estimate exact sale probabilities but rather to prioritize listings according to their relative ease of sale and to diagnose factors causing sales stagnation. Empirical results show that although the proposed two-stage model yields a lower concordance index than the single-stage model, it provides a useful exploratory framework for diagnosing property stagnation. By decomposing sales obstacles into "proposal issues" and "preference issues," the framework contributes to a deeper understanding of the complex mechanisms underlying real estate transactions. This structurally decomposed modeling framework offers a potential reference for more scientific decision-making regarding trade-offs between, for example, advertising investment and price adjustments.

2 Data

This study utilizes platform interaction records collected from March to May 2025.[1] The dataset has a longitudinal structure and records the sequence from property exposure to final transaction. Each observation corresponds to a listing observed under a particular market state on a given day as shown in Figure 1. Thus, each observation links a property, the contemporaneous market state, proposal information, and the final transaction outcome. This structure forms the basis of the proposed two-stage modeling framework.

2.1 Variable Categories

Following the data structure illustrated in Figure 1, explanatory variables are grouped into two categories: objective property characteristics and internal company factors.

[1]To comply with privacy requirements, all analyses are conducted using anonymized data, and no customer-level information is reported. Only aggregated distributions and summary statistics are presented throughout the paper.

Date	Property ID	Objective Property Characteristics	Internal Company Factors	Proposal Outcome	Sold
2025/03/01	P001	Age = Young Area = Large	Market State = (High Inventory, Medium Sales) Broker Incentive = High Fee Reduction = Low	Proposed to Customer C001 (Bin A)	No
2025/03/01	P002	Age = New Area = Medium	Market State = (High Inventory, Medium Sales) Broker Incentive = High Fee Reduction = High	Proposed to Customer C017 (Bin B)	Yes
2025/03/01	P003	Age = Old Area = Small	Market State = (High Inventory, Medium Sales) Broker Incentive = Low Fee Reduction = None	Not Proposed	No
2025/03/02	P004	Age = Middle-aged Area = Large	Market State = (Medium Inventory, Low Sales) Broker Incentive = Medium Fee Reduction = Medium	Proposed to Customer C023 (Bin A)	No

Figure 1: Illustrative example of the analysis dataset. Each row corresponds to a property observation on day t. The variables shown are simplified examples for illustration purposes.

2.1.1 Objective Property Characteristics

This category consists of property-specific attributes that describe the physical and financial characteristics of a listing. Specifically, it includes:

1. **Physical attributes**: including property age, floor level, floor area, and walking distance to the nearest station;

2. **Financial indicators**: including estimated-rent gap, management fees, and repair reserve funds. The *estimated-rent gap* denotes the difference between the company's internally estimated market rent and the current asking rent of the property.

2.1.2 Internal Company Factors

This category consists of non-market signals that influence broker and customer behavior. Specifically, it includes:

1. **Internal incentive mechanisms**: including broker performance rankings, business-specific bonuses;

2. **Market-state variables**: represented by inventory and sales activities within the platform, their construction is described in Section 2.3;

3. **Customer incentive strategies**: including fee reductions and point-based incentives.

Objective property characteristics are included because they capture the intrinsic attractiveness and marketability of listings, which may affect both brokers' proposal decisions and customers' purchase decisions. Internal company factors are incorporated to represent operational mechanisms within the platform, including

broker incentives, market conditions, and customer-oriented promotional strategies, all of which may influence the sales process.

2.2 Construction of Variables for Stage 1

The objective of Stage 1 is to model the conditional probability that a property is proposed to potential buyers under different market states. Proposal behavior depends not only on property characteristics but also on market state. To capture this effect, we define a market-state variable

$$\mathbf{m} = (\bar{I}_t, \bar{S}_t)$$

where $\bar{I}_t$ and $\bar{S}_t$ denote the rolling averages of inventory volume and sales volume over the previous three days ($t-3$ to $t-1$), respectively.

Using market-state variables instead of raw calendar dates also improves interpretability, because inventory volume and sales activity have direct operational meanings, whereas calendar dates mainly function as time identifiers and may capture period-specific effects that do not generalize beyond the observation period. To this end, we incorporate objective property characteristics, broker incentive information, and market-state variable reflecting market pressure as explanatory features. The binary target variable p_i indicates whether property i receives a proposal.

2.3 Construction of Variables for Stage 2

The objective of Stage 2 is to identify the probability that a proposed property is ultimately sold. We select objective property characteristics and customer incentive variables as explanatory features. Since the proposal process has already occurred at this stage, broker-side incentive rankings are excluded from the feature space. The binary target variable y_i indicates whether listing i is ultimately selected by the buyer from the proposal set.

2.4 Empirical Observations and Distributional Patterns

Empirical observations support the necessity of decomposing the sales process into two stages.

2.4.1 Extreme Sparsity of Proposals

By comparing the number of daily active listings, defined as properties that remained available for sale within the platform on a given day, with the number of properties actually proposed (Figure 2), a substantial discrepancy can be observed. Throughout the observation period, only approximately 35% of active listings were proposed to customers on a given day. This result indicates that a large proportion of active listings are not exposed to customers on a given day. Therefore, observed sales outcomes are influenced not only by customer preferences but also by broker-side proposal decisions, motivating the explicit modeling of the exposure process in Stage 1.

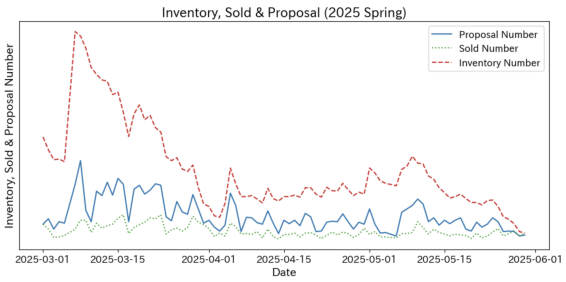


Figure 2: Daily numbers of active listings, proposed listings, and completed transactions.

2.4.2 Diversity of Property Attributes

Figure 3 illustrates the distributional patterns of physical and financial property characteristics. The substantial heterogeneity in inventory attributes directly affects both the probability of entering the proposal stage and the eventual probability of transaction.

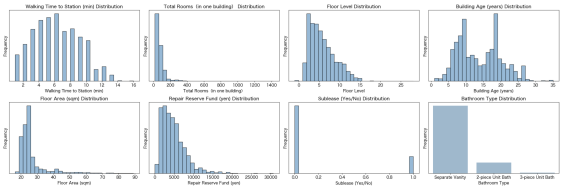


Figure 3: Distributions of representative property characteristics used in the analysis.

2.4.3 Constraints of the Decision Set

Figure 4 presents the average number of properties proposed to an individual customer and the corresponding number of completed transactions. The observed number of proposed properties is substantially smaller than the total inventory, reinforcing the necessity of modeling buyer choice within a proposal set rather than across the full inventory.

2.4.4 Customer Profile Distribution

Figure 5 presents the distributions of customer attributes used in Stage 2, including age, annual income, sex, and customer status (new versus existing

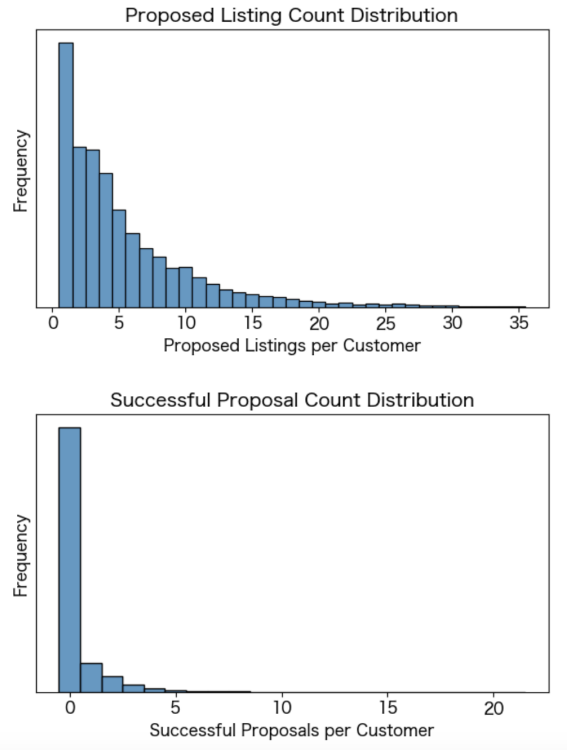


Figure 4: Numbers of proposed properties and completed transactions per customer.

customers). These variables are currently used in the company's customer segmentation and sales assignment process. Their distributions indicate that the customer population can be partitioned into a manageable number of demographic segments while maintaining sufficient observations within each segment.

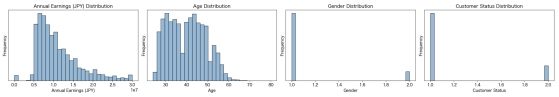


Figure 5: Distributions of customer characteristics.

3 Methods

3.1 Structural Decomposition of the Sales Process

To mitigate sample selection bias introduced by broker-side screening, we decompose the sales probability of property i, denoted as $P(\text{Sold}_i)$, into two sequential conditional probabilities.

The joint probability of a transaction is defined as:

$$P(\text{Sold}_i) = P(i \in E \mid \mathbf{m}) \times P(y_i = 1 \mid i \in E) \quad (1)$$

where:

- **Proposal probability** $P(i \in E \mid \mathbf{m})$ represents the probability that property i is selected by the broker and enters the proposal set E (Stage 1).
- **Selection probability** $P(y_i = 1 \mid i \in E)$ represents the conditional probability that the buyer selects property i from the proposal set, given that it has already been proposed (Stage 2).

This analytical framework enables us to mathematically disentangle broker-controlled visibility from buyer-driven preference utility.

3.2 Stage 1: Proposal Model (LightGBM)

The objective of Stage 1 is to estimate the probability that property i receives a proposal under market state $\mathbf{m}$. Although customer preferences may influence individual proposal decisions, the objective of Stage 1 is to estimate the intrinsic exposure propensity of properties. Incorporating customer-specific attributes would make proposal probabilities highly dependent on the temporal composition of customers and potentially reduce generalizability. Therefore, Stage 1 models the proposal probability using property characteristics, market state, and operational factors only.

We formulate this task as a binary classification problem:

$$\hat{q}_i = \Pr\left(p_i = 1 \mid \mathbf{x}_i^{obj}, \mathbf{z}_i^{int}, \mathbf{m}\right) \quad (2)$$

where $\mathbf{x}_i^{obj}$ and $\mathbf{z}_i^{int}$ represent the objective property characteristics and internal company factors associated with property i, respectively. We employ a LightGBM classifier to capture nonlinear interactions between property characteristics and market state, predicting a proposal probability $\hat{q}_i$ for each property.

3.3 Stage 2: Bin-Based Conditional Discrete Choice Model

3.3.1 Bin Construction and Candidate Set

Direct estimation of customer-level preference models is challenging because most customers were proposed with only a small number of properties, resulting in extremely sparse transaction observations. To obtain sufficiently stable estimates of buyer preferences, customers are grouped into demographic bins, and preference parameters are estimated at the bin level rather than at the individual level.

The bins are constructed using customer attributes currently employed in the company's operational segmentation process, including age, annual earnings, sex, and customer status (new versus existing customers). These variables are routinely used within the platform

to categorize customers and support sales assignment and customer management workflows.

The binning granularity was selected to balance two practical considerations: maintaining sufficient observations within each bin and avoiding excessive fragmentation of the data. Under this design, fewer than 100 bins were generated, providing adequate sample sizes for stable Bayesian estimation. To further mitigate data sparsity at the individual proposal-session level, all proposed properties associated with customers in bin b are aggregated into a common candidate set $\mathcal{C}_b$.

3.3.2 Latent Utility

Given the candidate set $\mathcal{C}_b$, the latent utility of property i, denoted as $u_{i,b}$, is modeled as follows:

$$u_{i,b} = (\mathbf{x}_i^{obj})^\top \mathbf{w} + \alpha \cdot \text{Incentive}_{i,b}^{cust} \tag{3}$$

where:

- $\mathbf{x}_i^{obj}$: objective characteristics of listing i;
- $\text{Incentive}_{i,b}^{cust}$: customer incentive variables (e.g., fee reductions and point-based incentives) associated with listing i for customers in bin b.
- $\mathbf{w}$: coefficient vector associated with objective property characteristics;
- α: coefficient associated with customer incentive variables.

The conditional selection probability $P(i \mid b)$ is computed using the Softmax function

$$P(i \mid b) = \frac{\exp(u_{i,b})}{\sum_{j \in \mathcal{C}_b} \exp(u_{j,b})} \tag{4}$$

3.3.3 Likelihood Function

We aggregate transaction events within each bin and estimate the model using a multinomial likelihood:

$$(y_{i_1,b}, \ldots, y_{i_{|\mathcal{C}_b|},b}) \sim \text{MLL}\left(d_b, \{P(i \mid b)\}_{i \in \mathcal{C}_b}\right) \tag{5}$$

where d_b denotes the total number of transactions observed within bin b.

3.3.4 Bayesian Inference

We estimate the posterior distribution of $\mathbf{w}$ and α using the No-U-Turn Sampler (NUTS) algorithm implemented in NumPyro. A weakly informative prior is assigned to the model parameters:

$$\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \qquad \alpha \sim \mathcal{N}(0, \sigma_\alpha^2). \tag{6}$$

All continuous features are standardized or rescaled to the range [-1,1] prior to estimation to improve numerical stability and sampling efficiency. Model convergence is assessed using the Gelman–Rubin statistic $\hat{R} < 1.01$ and effective sample size (ESS).

3.4 Stage Integration: Overall Sales Probability

The final sales probability $P(\text{Sold}_i)$ is obtained by integrating the two stages through the law of total probability. We marginalize the conditional selection probability across customer bins using their population shares π_b:

$$P(\text{Sold}_i) = \hat{q}_i \times \sum_b P(i \mid b)\pi_b \tag{7}$$

where $\hat{q}_i$ is the proposal probability estimated in Stage 1, and π_b denotes the empirical population share of customer bin b.

This score represents the expected probability that property i is both proposed by a broker and selected by a customer.

3.5 Baseline Models

We compare the proposed framework against the following two baselines:

- **LightGBM LambdaRank:** An industry-standard learning-to-rank algorithm that directly optimizes ranking performance but does not distinguish between exposure and customer choice.
- **Single-Stage Bayesian Softmax:** A Bayesian discrete-choice model that directly models transaction outcomes over the full inventory. All active listings are assumed to be equally available for customer consideration, and feature coefficients are estimated via MCMC using the same prior specification as Stage 2.

4 Results and Discussion

Consistent with the objective of this study, we evaluate model performance based on ranking quality rather than absolute probability accuracy. Operational decisions such as inventory allocation, pricing review, and promotional resource assignment are typically made based on relative ranking rather than absolute probabilities. Therefore, ranking performance is more relevant than classification accuracy or probability calibration.

4.1 Evaluation Metric: Concordance Index (C-index)

To evaluate the ranking performance of the proposed models in predicting property sales probability, we adopt the Concordance Index (C-index) [Harrell et al., 1996]. Under this metric, we consider all comparable pairs of properties with observed sales durations. For each pair, a prediction is considered concordant if the property assigned a higher predicted sales probability $P(\text{Sold}_i)$ is observed to sell in a shorter period than the comparison property. The C-index is defined as the proportion of concordant pairs among all comparable pairs, where 0.5 indicates random ranking and 1.0 indicates perfect concordance between predicted sale probability and observed selling speed.

4.2 Ranking Performance Comparison

Table 1 reports the ranking performance evaluated using the overall sales probability $P(\text{Sold}_i)$. The empirical results indicate that Bayesian models outperform the LightGBM baseline. While the single-stage model achieves the highest C-index (0.640), the proposed two-stage model attains a competitive performance of 0.610.

Table 1: C-index Performance Across Models

Model	C-index
LightGBM LambdaRank	0.538
Single-Stage Bayesian Softmax	0.640
Two-Stage Bayesian Model	0.610

4.3 Posterior Interpretation and Feature Analysis

One advantage of the Bayesian choice model is that it provides posterior distributions rather than point estimates. Figure 6 presents posterior distributions of selected coefficients in the Stage 2 choice model. Due to space limitations, not all coefficients are shown. The displayed features were selected because they are either directly relevant to business decision-making or exhibit substantial posterior effects. Positive coefficients indicate attributes associated with a higher probability of being selected, whereas negative coefficients indicate reduced attractiveness. The spread of each posterior distribution reflects the uncertainty of the corresponding parameter estimate.

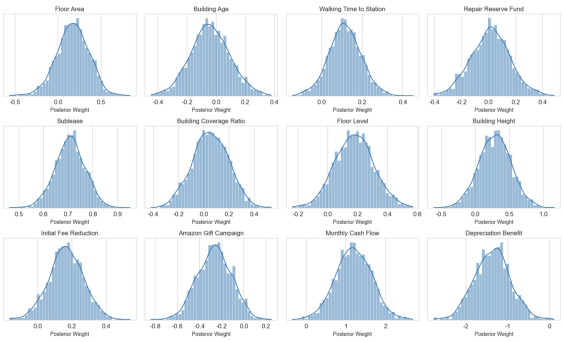


Figure 6: Posterior distributions of Stage 2 coefficients estimated by MCMC.

4.4 Discussion: The Trade-off Between Goodness-of-Fit and Debiasing

The 0.03 performance gap between the single-stage and two-stage models does not imply failure of the proposed framework. In the single-stage model defined in Section 3.5, we assume complete visibility of the inventory, allowing the parameter vector $\mathbf{w}$ to improve historical fit by implicitly absorbing broker-side proposal preferences.

However, this improved goodness-of-fit may be misleading when proposal opportunities are not randomly assigned across listings. According to Steck [Steck, 2010], when models selectively ignore sample selection bias, offline evaluation metrics (e.g., C-index) may appear superior, while the estimated parameters lose interpretability due to contamination from observational bias.

In contrast, the proposed two-stage framework explicitly separates broker-side proposal logic $P(i \in E \mid \mathbf{m})$ from buyer-side selection logic $P(y_i = 1 \mid i \in E)$ through the structural decomposition introduced in Section 3.1. Consequently, the utility parameters estimated in the second stage are less contaminated by broker-side screening effects.

This strategy of sacrificing a portion of predictive fit in exchange for behaviorally meaningful parameter estimates allows the model to better capture buyers' genuine preferences while mitigating spurious fitting effects.

4.5 Business Value: A Diagnostic Framework for Property Stagnation

Based on the probabilistic integration described in Section 3.4, the proposed model provides a two-dimensional diagnostic framework for analyzing the causes of property stagnation. A low first-stage proposal probability $P(i \in E \mid \mathbf{m})$ suggests that a property is unlikely to be presented to potential buyers un-

der the current broker-side screening process. In such cases, operational interventions may focus on increasing exposure through revised proposal rules, modified sales assignment strategies, or additional promotional activities.

In contrast, a low second-stage selection probability, $P(y_i = 1 \mid i \in E)$ indicates that the property is frequently proposed but rarely chosen. This pattern suggests that visibility is not the primary bottleneck and that interventions such as price adjustments, incentive redesign, or customer-targeting refinement may be more effective.

By distinguishing between exposure-related and preference-related causes of stagnation, the proposed framework supports more efficient allocation of marketing resources and inventory management efforts. Such mechanism-aware diagnostics cannot be obtained from a traditional single-stage ranking score, which only provides a single aggregated measure of sale probability.

5 Conclusion and Limitation

This study proposes a two-stage probabilistic framework based on Bayesian inference to address sample selection bias in broker-mediated real estate markets. By decomposing the sales process into a broker-side proposal stage (Stage 1) and a buyer-side selection stage (Stage 2), the proposed framework facilitates the separation of broker-side proposal effects from buyer-side preference utility.

Empirical results demonstrate that the proposed framework maintains relatively strong ranking performance while providing a mechanism-aware diagnostic tool. Beyond identifying potentially high-value properties, the model enables practitioners to distinguish between proposal issues and preference issues, thereby offering a more scientific basis for adjusting promotional incentives and pricing strategies.

One limitation of the present study is the use of demographic-based customer bins. Although this aggregation improves statistical stability and enables Bayesian estimation under sparse proposal data, customers within the same bin may still exhibit heterogeneous preference structures.

Future research will explore hierarchical Bayesian structures in which customer segments partially share information through common prior distributions. Such partial pooling may improve parameter stability in sparse bins while preserving segment-specific preference patterns and reducing the limitations associated with demographic-based aggregation.

References

[Do and Grudnitski, 2004] Do, A. Q. and Grudnitski, G. (2004). A critical review of real estate durations. *Journal of Real Estate Literature*, 12(1):41–66.

[Han and Strange, 2015] Han, L. and Strange, W. C. (2015). The microstructure of housing markets. In *Handbook of Regional and Urban Economics*, volume 5, pages 813–886. Elsevier.

[Harrell et al., 1996] Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387.

[Heckman, 1979] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

[Jud and Seaks, 1994] Jud, G. D. and Seaks, T. G. (1994). Sample selection bias in estimating housing sales and prices. *Journal of Real Estate Research*, 9(3):289–298.

[Levitt and Syverson, 2008] Levitt, S. D. and Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *Review of Economics and Statistics*, 90(4):599–611.

[Saito, 2015] Saito, H. (2015). *A Study on Information and the Role of Professionals in Existing Home Transactions*. PhD thesis, Meikai University.

[Shimizu, 2010] Shimizu, C. (2010). Removing sample selection bias in repeat-sales house price index estimation. *CSIS Discussion Paper*.

[Shimizu, 2011] Shimizu, C. (2011). House prices at different stages in buying/selling process. Ottawa Group Meeting, United Nations Statistical Division.

[Shimizu et al., 2004] Shimizu, C., Nishimura, K. G., and Asami, Y. (2004). Search and vacancy costs in the tokyo housing market: Attempt to measure social costs of imperfect information. *Review of Urban and Regional Development Studies*, 16(3):210–230.

[Steck, 2010] Steck, H. (2010). Training and testing of recommender systems on data missing not at random. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*, pages 167–174.

[Wheaton, 1990] Wheaton, W. C. (1990). Vacancy, search, and prices in a housing market matching model. *Journal of Political Economy*, 98(6):1270–1292.