



DUPLICATE ONLINE PROPERTY LISTINGS IN RESIDENTIAL PROPERTY PRICE INDEX COMPILATION: A METHODOLOGICAL FRAMEWORK FOR UZBEKISTAN

Dildora Akramovna Togayeva

Independent researcher, Institute for Advanced Training and
Statistical Research, National Statistics Committee of the Republic of
Uzbekistan, Tashkent, Uzbekistan

ORCID: 0009-0001-6989-2671 | Corresponding author:

dildar.akramovna@gmail.com

<https://doi.org/10.5281/zenodo.22768029>

ARTICLE INFO

Received: 10th September 2026

Accepted: 12th September 2026

Online: 14th September 2026

KEYWORDS

*residential property price index,
online listings, duplicate listings,
record linkage, deduplication,
hedonic regression, Uzbekistan*

ABSTRACT

Online real-estate listings provide timely and granular information for residential property price index compilation, but repeated advertisements for the same dwelling can inflate the number of observations, alter sample composition, and give some properties disproportionate influence in hedonic estimation. Methods: The study develops a property-quarter deduplication framework for Uzbekistan's RPPI pipeline. It combines data standardization, blocking, similarity-based record linkage, gold-standard validation with Precision, Recall and F1, and parallel RAW-versus-DEDUP index calculations. The approach is designed for the existing strata-based, four-quarter rolling time-dummy hedonic system. Results: The empirical compilation context covers 1,654,058 listings for 2024Q1-2026Q2, 28 strata, seven rolling windows and 196 hedonic regressions. A duplicate-rate measure and a Duplicate Impact Diagnostic are specified, together with rules for keeping the strata structure, regression specification, outlier treatment, chaining and aggregation weights identical across parallel calculations. Because the current database does not retain duplicate labels or matched RAW and DEDUP samples, the actual index effect of deduplication is not reported as an estimated result. Conclusion: Deduplication should be treated as a statistical quality-control component rather than a simple technical cleaning step. A validated property-quarter record-linkage module would allow official compilers to document matching quality and quantify whether repeated listings materially affect RPPI dynamics.

1. Introduction

Online housing advertisements can complement administrative sources by providing large volumes of timely and detailed information on asking prices and dwelling characteristics. Their value for market monitoring is especially strong where transaction-level data are incomplete or are available only with a lag. Loberto, Luciani, and Pangallo (2022) show that online listings can reveal housing-market dynamics at high frequency, while also demonstrating that repeated advertisements for the same underlying property can distort measured supply and sample composition.

For residential property price index (RPPI) compilation, source-data quality and preprocessing are integral to the index methodology. International guidance emphasizes the need to assess data sources, control for quality differences, and document the transformations applied before estimation (Eurostat et al., 2013; International Monetary Fund [IMF], 2020). In Uzbekistan, recent methodological work has further developed the compilation of housing-market price indices (National Statistics Committee of the Republic of Uzbekistan, 2026). Within this setting, duplicate listings are a specific data-quality issue because a single property may be reposted, advertised by several intermediaries, or re-entered with partially modified characteristics.

The current RPPI compilation database used in this study covers 2024Q1-2026Q2. After plausibility checks, regional price-per-square-metre controls and stratification, 1,654,058 advertisements enter the regression stage. Of these, 1,577,530 observations (95.37%) originate from OLX and 76,528 (4.63%) from Telegram; 1,405,897 observations refer to apartments and 248,161 to individual houses. The production design comprises 28 strata, seven four-quarter rolling windows and 196 hedonic regressions. However, the database does not retain a labelled duplicate/non-duplicate indicator or parallel pre- and post-deduplication samples. Consequently, the purpose of this short communication is methodological: to formalize a reproducible duplicate-detection and validation module and to define how its impact on the final RPPI should be measured in a subsequent empirical stage.

2. Methods

The proposed unit of deduplication is the property-quarter. Multiple advertisements judged to represent the same dwelling within a quarter are grouped into one property-quarter cluster. If the same property reappears in a later quarter, it is retained as a new time-period observation so that genuine intertemporal information is not removed from the rolling time-dummy model.

Let N_t denote the number of raw records in quarter t and U_t the number of unique property-quarter observations remaining after deduplication. The share of redundant records is defined as:

$$DR_t = ((N_t - U_t) / N_t) \times 100 \quad (1)$$

DR_t measures the proportion of excess records created by repeated listings; it is not the proportion of properties that were duplicated. It should therefore be reported by quarter, region, market segment and stratum rather than only at national level.

The existing RPPI approach links the logarithm of property price to structural and geographic characteristics within a rolling time-dummy hedonic specification:

$$\ln(P_{ist}) = \alpha_s + X'_{ist} \beta_s + \sum_q \delta_{sq} D_{q,ist} + \gamma'_s G_{is} + \varepsilon_{ist} \quad (2)$$

For apartments, the current specification includes total area, number of rooms, floor, number of storeys, quarter dummies and relevant geographic fixed effects. For individual houses, plot size, number of rooms, quarter dummies and necessary geographic controls are used. Each stratum is first estimated by ordinary least squares. Observations with Cook's Distance $D_i \geq 4/n$ are then excluded and the model is re-estimated.

Duplicate listings are not necessarily econometric outliers. Exact or near-exact repetitions can remain close to the fitted regression surface while effectively increasing the relative influence of one property. For an exact duplicate appearing m_{jt} times within quarter t , its simplified implicit weight can be represented as $w_{jt} = m_{jt}$. For reposted listings whose price or descriptions change, this is not an exact algebraic weight, but it still captures the mechanism by which repeated observations can increase the property's influence on estimated coefficients and time effects.

When no unique property identifier is available, the problem is treated as record linkage or entity resolution. Following the classical framework of Fellegi and Sunter (1969) and later data-matching practice (Christen, 2012), the proposed procedure has two main stages. First, blocking reduces the number of candidate pairs. A practical blocking vector is: quarter, region, district, market type, number of rooms and area group. Including quarter prevents advertisements from different quarters from being automatically merged, while the remaining fields restrict comparisons to plausible matches.

Second, candidate pairs within a block receive a similarity score:

$$S_{ij} = \sum_{k=1}^K \omega_k s_{ijk}, \quad \omega_k \geq 0, \quad \sum_k \omega_k = 1 \quad (3)$$

Table 1. Candidate attributes for similarity-based duplicate matching

Attribute	Suggested comparison	Purpose
Region/district	Exact match	Strong geographic restriction
Total area	Small relative difference	Allows minor entry errors
Rooms	Exact or ± 1	Soft tolerance for miscoding
Floor/storeys	Exact or close	Apartment-specific matching
Price	Logarithmic difference	Allows reposted price changes
Listing date	Time interval	Separates repeated postings

A candidate pair is classified as a potential match when $S_{ij} \geq \tau$. Matched pairs are then consolidated into property-quarter clusters. The threshold τ and attribute weights ω_k should be calibrated on manually reviewed duplicate/non-duplicate pairs. Classification quality should be evaluated with Precision, Recall and F1 because false positives remove genuinely distinct dwellings, whereas false negatives leave repeated observations in the database.

$$\text{Precision} = TP / (TP + FP) \quad (4)$$

$$\text{Recall} = TP / (TP + FN) \quad (5)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (6)$$

3. Results

The main output of this study is an operational diagnostic framework that can be inserted before RPPI estimation without changing the downstream index methodology. The proposed sequence is: raw advertisements → standardization and plausibility controls → property-quarter blocking → similarity matching → gold-standard validation → unique property-quarter database → stratum assignment → initial OLS → Cook's Distance control → re-estimated OLS → rolling time-dummy RPPI.

To isolate the effect of deduplication, two index pipelines should be calculated from the same source data. I_t^{RAW} denotes the RPPI produced without the new module, whereas I_t^{DEDUP} denotes the RPPI produced after validated deduplication. Their relative difference is summarized by the Duplicate Impact Diagnostic:

$$DI_t = ((I_t^{RAW} / I_t^{DEDUP}) - 1) \times 100 \quad (7)$$

The term “diagnostic” is used deliberately rather than “bias” because the deduplicated series cannot be regarded as an error-free benchmark until the matching model has been validated. A value of DI_t close to zero would indicate a small difference between the two pipelines. Persistent positive or negative values for selected strata, regions or periods would indicate that duplicate intensity is systematically associated with the measured index result.

A valid parallel comparison requires the two pipelines to keep all other methodological components fixed: the same strata, the same rolling window, identical hedonic specifications, the same Cook's Distance rule, the same chaining procedure and the same aggregation weights. Additional diagnostics should compare time-dummy coefficients, adjusted R-squared, observation counts, the share of Cook-filtered records and distributions of key property characteristics.

The current database provides the empirical scale on which this module would operate but does not contain the labelled duplicate/non-duplicate gold standard required to estimate τ and ω_k , and it does not store parallel RAW and DEDUP index results. Therefore, no numerical estimate of deduplication's impact on the national RPPI is claimed in this paper. This limitation is substantive and defines the next empirical step rather than being replaced by an unsupported result.

4. Discussion

The framework has two practical implications for official statistics. First, duplicate detection should be separated conceptually from outlier detection. Cook's Distance identifies influential observations relative to the fitted regression, while duplicates can be entirely plausible and still over-represent a property. Treating both issues as the same cleaning problem can therefore leave a hidden source of unequal observational influence.

Second, the proposed property-quarter unit preserves the time structure needed for rolling hedonic indices. Removing all reappearances of a property across the full database would risk discarding legitimate information about later periods. Restricting consolidation to a quarter offers a practical compromise: it controls repeated exposure within the estimation period while retaining inter-quarter observations.

The framework is also auditable. Blocking rules, similarity variables, threshold calibration and classification metrics can be documented and reproduced. This is particularly important for official-statistics production, where a data-cleaning algorithm should be

explainable and its impact on published indicators should be measurable. The proposed RAW-versus-DEDUP design creates such a measurement boundary.

The principal limitation is that the present study is not yet an empirical validation of the duplicate-detection algorithm. Without manually labelled pairs, Precision, Recall and F1 cannot be calculated, and without parallel index runs, DI_t cannot be estimated. The next stage should therefore construct a representative gold-standard sample across market types and regions, calibrate the matching threshold, and rerun the RPPI under identical methodological settings with and without deduplication.

5. Conclusion

Online listing data are valuable for timely RPPI compilation, but repeated advertisements can change the effective composition of the sample and the relative influence of individual properties in hedonic estimation. This study proposes a property-quarter deduplication framework that combines blocking, similarity-based record linkage, gold-standard validation and a controlled RAW-versus-DEDUP index comparison.

The contribution is not a new record-linkage theory. It is the integration of established linkage principles into Uzbekistan's strata-based, rolling time-dummy RPPI pipeline as a documented statistical quality-control module. The framework provides explicit measures for redundant-record intensity, matching quality and index sensitivity. Its empirical implementation requires a labelled validation sample and parallel RPPI calculations; until that stage is completed, the effect of deduplication should be described as a testable methodological diagnostic rather than an observed index result.

Funding

Funding information was not specified in the source manuscript and should be confirmed by the author before submission.

Conflict of Interest

Conflict-of-interest information was not specified in the source manuscript and should be confirmed by the author before submission.

AI Tool Use Disclosure

Generative AI was used for English-language editing, structural adaptation, and formatting assistance. The author remains responsible for the study design, data, analysis, interpretation, citations, and final submitted text.

References:

- Loberto, M., Luciani, A., & Pangallo, M. (2022). What do online listings tell us about the housing market? *International Journal of Central Banking*, 18(4), 325-377.
- Eurostat, International Labour Organization, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations Economic Commission for Europe, & World Bank. (2013). *Handbook on residential property price indices*. Publications Office of the European Union.
- International Monetary Fund. (2020). *Residential property price index (RPPI) practical compilation guide*. International Monetary Fund.
- National Statistics Committee of the Republic of Uzbekistan. (2026, July 13). O'zbekiston Respublikasida uy-joy bozori narx indekslarini hisoblash metodologiyasidagi o'zgarishlar

to'g'risida [On changes in the methodology for calculating housing market price indices in the Republic of Uzbekistan].

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

<https://doi.org/10.1080/01621459.1969.10501049>

Lehti, P., & Fankhauser, P. (2005). A precise blocking method for record linkage. In A. C. M. Fong, et al. (Eds.), *Data warehousing and knowledge discovery* (LNCS 3589, pp. 210-220). Springer. https://doi.org/10.1007/11546849_21

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. <https://doi.org/10.1007/978-3-642-31164-2>

