# Automating Web Data Collection for Kenya's Residential Property Price Index

*Innovation, Leadership, and Lessons for Emerging Markets*

Lucas Sagire and Archibald Macharia
Kenya National Bureau of Statistics (KNBS)

International Conference on Real Estate Statistics
Tokyo, February 18–20, 2026

# The Challenge: Property Statistics in Emerging Markets

## Kenya's Context

- 47 counties with diverse property markets
- Urbanization rate >4% annually
- Major cities: Nairobi, Mombasa, Kisumu, Nakuru
- Property as primary investment vehicle

## Why Administrative Data Failed

- Paper-based deeds registration
- Fragmented across county offices
- Incomplete digitization
- Missing property characteristics data

## The Gap

**No official RPPI**

- Policymakers relied on fragmented private indices
- Inconsistent methodologies
- Limited geographic coverage (Nairobi-focused)
- Delayed traditional approach would take years

## KNBS Objective

*Create Kenya's first official RPPI using innovative and transparent digital data collection methods to support evidence-based decisions about housing policy, investment, macro-prudential analysis, credit risk, and economic planning in Kenya.*

# Data Collection Strategy: Hybrid Approach

## Web Scraping (Primary)

**Six Major Platforms:**
- Property24 Kenya
- Jiji
- Kenya Property Centre
- BuyRent Kenya
- PigiaMe
- PropertyPro

**Collection:** Twice monthly (bi-weekly)
**Consolidation:** Quarterly

## Agent Surveys (Validation)

**Coverage:**
Monthly targeted surveys of real-estate agents
**Purpose:**
- Validate web-scraped prices
- Extend coverage to secondary cities and rural areas
- Capture informal transactions absent from online listings
- Provide market context and ground truth

# Data Coverage

**5,000-6,000**
Listings per Quarter

**Over 60% Counties**
Coverage

## Geographic Distribution

- Nairobi City and environs: ~40% of listings (market concentration)
- Major cities (Mombasa, Kisumu, Nakuru): 30%
- Secondary/rural counties: 30%

## Data Completeness

Price: 95% | Location: 98% | Bedrooms: 92%
Property Type: 90% | Amenities: 70%
Floor Area: 40% (target for improvement)

# Methodology: Time-Dummy Hedonic Model

## How It Works

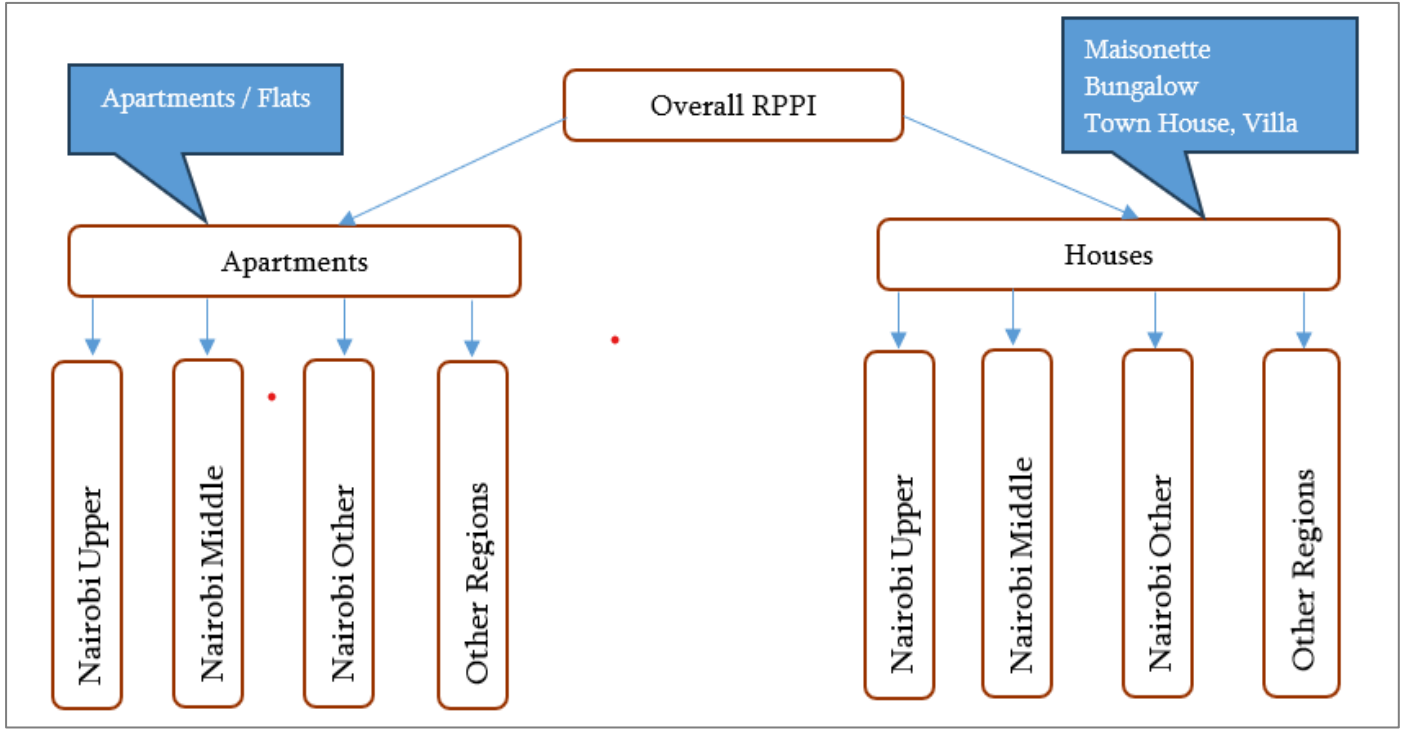
**Price Decomposition:**
Breaks down property price into bundles of characteristics (location, type, size-floor area, bedrooms, bathrooms- and amenities)
**Quality Adjustment:**
Controls for composition shifts as the mix of listed properties changes quarterly
**Pure Price Movement:**
Isolates actual price changes from changes in property mix

## Rolling Four-Quarter Window

Current quarter + 3 prior quarters
**Advantages:**
- Sufficient sample size per stratum
- Avoids over-fitting from sparse recent data
- Incorporates only recent market conditions
- Computationally efficient (~5,000 listings)

**Stratification:**
- Two (2) level stratification: property types (standalone houses, apartments/flats) and regional groupings
- Results into eight (8) strata/ market segments for index compilation

# Stratification



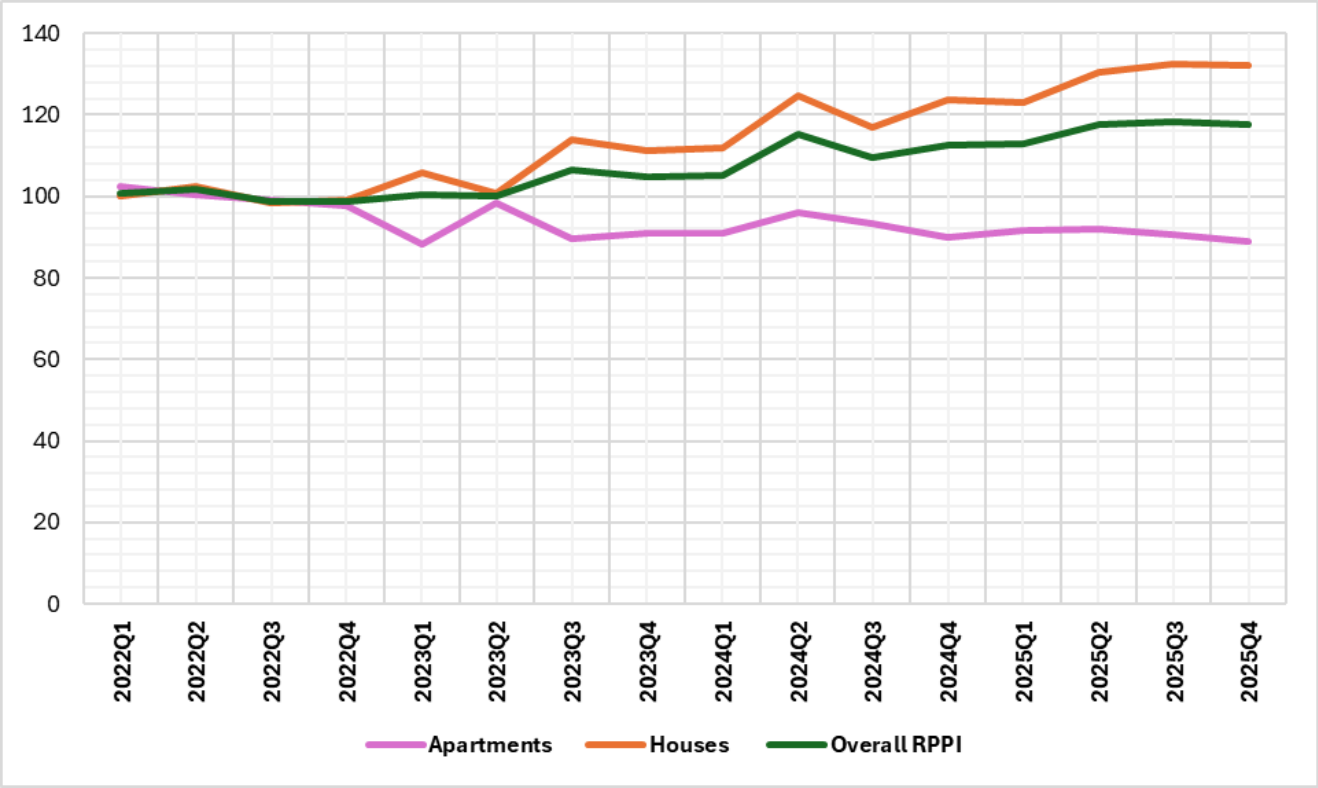
# Quality Assurance

- **Completeness check:** Implemented to ensure the dataset contains no gaps in the core property characteristics required for reliable index. Transactions with gaps were excluded.
- **Duplicates check:** A duplicate-removal procedure is applied before compiling the RPPI using a combination of key identifying fields namely property name, price, location, period, and number of bedrooms, the procedure systematically scanned the dataset for any repeated records that appeared to refer to the same property within the same period.

✓ Any duplicated property transactions identified are automatically dropped to prevent double-counting, which could otherwise distort price levels and bias index estimates.

- **Outliers Correction:** Cook's Distance used to clear outliers in the data before running the model.

# RPPI


140
120
100
80
60
40
20
0
2022Q1
2022Q2
2022Q3
2022Q4
2023Q1
2023Q2
2023Q3
2023Q4
2024Q1
2024Q2
2024Q3
2024Q4
2025Q1
2025Q2
2025Q3
2025Q4
Apartments
Houses
Overall RPPI

# Kenya's Innovation: Unified Web-Scraping Dashboard

**Core Innovation:** Automated system that harvests property listings from six major platforms twice monthly, supplemented by agent surveys, producing over 5,000 listings quarterly for official RPPI compilation.

## Design Principles

### Automation & Scale
Collect thousands of listings routinely with minimal manual work

### Robustness
Handle platform changes, anti-scraping defenses, and missing data gracefully

### Transparency
Log all data collection, transformations, and decisions for reproducibility

# Technical Innovations

## 1. Floor Area Data Integration with NLP

Extract floor area from unstructured text using regex patterns + ML-based imputation (RandomForest) → Increases coverage from 40% to 85%, improving hedonic model $R^2$ by 10-15%

## 2. Machine Learning Data Validation

GradientBoosting classifier trained on 2 years of manual reviews (15,000+ properties) automatically categorizes listings as valid/flagged/excluded → Reduces analyst QA time by 60-80%

## 3. Nationwide Agent Survey Integration

Extend scheduler framework to treat agent surveys as configurable data sources alongside web scrapers → Unified tracking of all 47 counties, ~240 transactions monthly for validation

# Challenges Encountered & Solutions

## Scalability Bottleneck

**Challenge:** System designed for ~100 listings but needed to handle 1,000+
**Solution:** Distributed cloud infrastructure, optimized code, smart validation → Doubled capacity

## Hedonic Model Performance

**Challenge:** Luxury segment (Nairobi Upper) $R^2$ ~0.50 due to heterogeneity
**Solution:** Separated market segments, published with confidence intervals and metadata flags

## Platform Disruptions

**Challenge:** Property24 changed website structure, breaking scraper
**Solution:** Formalized platform monitoring, maintained spare modules, rapid response protocol

## Data Quality Heterogeneity

**Challenge:** Different platforms have varying data quality standards
**Solution:** Automated validation, cross-validation with agent surveys, transparent documentation

**Key Learning:** Continuous maintenance (10-15% of capacity), platform diversification, and transparent quality reporting are essential for sustained operation

# Kenya's Approach in International Context

## How Kenya Differs

**Data Source Innovation:**
• Primary: Online property listings (asked prices)
• Traditional: Transaction data from deeds/mortgages
**Hybrid Approach:**
• Few NSOs combine web scraping with agent surveys systematically at Kenya's scale
**Granularity:**
• 47 county-level indices (target) vs. UK 12 regions, US states
• Reflects devolution and policy demand

## Leadership Position

**Regional First:**
• Among the first in sub-Saharan Africa using systematic digital collection
**Replicability for Peers:**
• Nigeria, South Africa, Uganda, Tanzania face similar constraints
• Lower-cost, faster path than building traditional infrastructure
**Integration with SDGs:**
• SDG 11: Housing affordability, regional equity
• Well-being indicators: Property as wealth proxy

# Technical Lessons for Other NSOs

## What Worked

- **Platform diversification** — rely on 6 platforms to mitigate bias
- **Respectful automation** — minimize legal risk with proper request rates
- **Platform-specific adaptation** — modular architecture essential
- **Continuous maintenance** — 10-15% capacity prevents breakdowns
- **Hybrid data sources** — surveys mitigate bias and provide validation

## Quality Assurance

Weekly automated dashboards + quarterly technical/statistical audits + transparent documentation for every release

## Resource Requirements

**Skill Mix Needed:**

- Web developers (scraping, DevOps)
- Data engineers (ETL, validation)
- Statisticians (hedonic modeling)
- Subject-matter experts (real estate context)

**Investment:**

- Development: ~3-4 FTE over 2-3 years (2022-2025)
- Ongoing: ~2-3 FTE for quarterly releases

**Institutional Partnerships:**
Central bank, ministries, industry cooperation essential

# Conclusion: A Replicable Model for Emerging Markets

## Key Achievements

✓ Overcame infrastructure gaps without waiting for traditional transaction registries
✓ Achieved comprehensive nationwide coverage (47 counties) with quarterly releases
✓ Demonstrated technical innovation: scalable, auditable, reproducible systems
✓ Positioned Kenya as regional leader in digital data transformation for official statistics
✓ Provided replicable blueprint for Nigeria, Uganda, Tanzania, and other emerging markets

**Contact:** directorgeneral@knbs.or.ke | *First Official RPPI Release: March 2026*