Enhancing Real Estate Market Surveillance: A Machine Learning Approach to Developing a Monthly Housing Price Index

Extended Abstract

The housing market strongly affects the economy. It impacts prices, how people spend money, and the stability of banks. The Bangko Sentral ng Pilipinas (BSP) tracks housing prices with its Residential Real Estate Price Index (RREPI). This index comes out every three months. It shows price changes for different types of homes like houses, townhouses, and condos in Manila and other areas. The BSP derives this index using data from the BSP's Residential Real Estate Loan (RREL) reports on home loans. These reports include details about the property location, type, value, size, number of bedrooms, and age.

While the RREPI helps track housing prices, waiting three months between updates can miss short-term changes. This study proposes the development of a complementary monthly real estate price index utilizing machine learning (ML) techniques and alternative data sources. Initial analysis of the RREL dataset (2016–2023) reveals strong predictive power, with Random Forest models achieving an adjusted R^2 of 76.98% nationwide, and performance varying between 62% and 81% across different market segments. Key price determinants identified include house area, lot value, property age, number of bedrooms, and location characteristics.

This study proposes a two-stage data pipeline to develop a monthly housing price index:

1. **Key Indicator Identification**: ML models identify the main factors driving housing price changes across market segments.
2. **Monthly Index Construction**: The identified factors are used to predict housing prices, which are then aggregated into a monthly index. The index incorporates median prices, price per square meter, and weighted values based on amenity density. Google Trends data captures real-time consumer interest, while geospatial weights account for proximity to key amenities like schools, hospitals, and transportation hubs.

Literature Review:

Housing markets influence macroeconomic stability through multiple channels, as noted in the literature. Leamer (2007) identified housing cycles as leading indicators of economic downturns, while Case & Shiller (2003) linked speculative housing bubbles to systemic financial risks.

There are also studies that adopt advanced analytics to model housing market complexities. Ghysels et al. (2013) demonstrated improved forecasting accuracy using mixed-frequency models integrating high-frequency financial data with low-frequency macroeconomic indicators. Krasovytskyi et al. (2024) validated ML's superiority over traditional econometric models while Nneji et al. (2013) illustrated regime-dependent responses of housing markets to macroeconomic shocks.

Recent research by Hoxha (2024) analyzed ML models for predicting housing prices in Prishtina. Decision tree models produced the best performance. The study on tree-based ML algorithms by Yavuz Özalp et al. (2023) in Artvin City Center found XGBoost and Random Forest performed best. They achieved correlation coefficients of 0.705 and 0.701, respectively.

Wu & Brynjolfsson (2013) demonstrated that Google Trends data can predict housing market trends with 7.1% better accuracy for future home sales. Bency et al. (2017) showed that geospatial features significantly improve housing price prediction models. Their approach incorporated proximity to amenities through weighted distance metrics.

These studies underscore ML's potential to effectively capture complex non-linear relationships within housing data, particularly when leveraging temporal patterns and alternative data sources.

Methodology:

This study employs a structured analytical pipeline to develop a monthly housing price index, integrating ML techniques with geospatial amenity weighting. The methodology follows four key phases, supported by established practices in real estate analytics and econometric modeling.

Data Collection and Preparation

The analysis utilizes the BSP's RREL dataset (2016–2023), comprising 336,123 initial entries. Data cleaning removes inconsistent entries, including properties with zero total value, floor areas <6 sqm, missing house types, and negative ages or location codes, resulting in 302,797 validated observations. Spearman's rank correlation (≥0.7 threshold) eliminates multicollinearity among variables, discarding redundant features such as house segment, per-square-meter price, and province identifiers.

Feature Engineering and Geospatial Integration

Numerical variables undergo logarithmic transformation and MinMax scaling to normalize ranges. Categorical variables, such as year, quarter, city/municipality, property type, and strata, are encoded as dummy variables. With reference to the study by Alasia et al. (2021), location-based value drivers are approximated using geospatial amenity weights derived from regional/provincial amenity concentration. These weights are calculated as:

$$Amenity\ Score_i = \sum_{j=1}^{n} \frac{Count\ of\ Amenity\ Type\ j\ in\ Region\ i}{Area\ of\ Region\ i(km^2)}$$

Model Development and Validation

Seven ML models are trained to predict log-transformed acquisition costs: Linear Regression, Decision Tree, Random Forest, Extra Trees, Gradient Boosting, XGBoost, and AdaBoost. Independent variables include house area, lot value, bedrooms, property age, city class, and geospatial amenity

scores. Models are evaluated using adjusted R^2 and RMSE, with time-series cross-validation to ensure robustness.

Monthly Index Construction

Predicted prices are aggregated into a composite index using:

1. **Median Price per Square Meter**: Normalized by housing type and region.
2. **Weighted Geospatial Scores**: Amenity density metrics adjust values for location desirability.
3. **Lagged Macroeconomic Variables**: Three-month moving averages of GDP growth and inflation rates (Ghysels et al., 2013). The index is segmented by region (NCR, Metro Cebu, Metro Davao, Rest of PH) and housing type, with monthly updates via rolling 30-day window.

Preliminary Results:

Housing prices change differently across regions. Urban areas like Metro Cebu show 23% more price changes than rural areas. Random Forest models work best nationwide with 76.98% accuracy. XGBoost works better for condominiums in Metro Cebu with 79.02% accuracy. Analysis shows house area (0.62) and lot value (0.58) are the most important factors in pricing.

Property age affects price negatively (-0.34), meaning newer homes cost more, especially in cities. The number of bedrooms matters a lot for family homes. Each extra bedroom adds 8-12% to property value, depending on location and house type.

Timeline:

Period	Activity
Apr – May 2025	Finalize key indicator validation; optimize ML models
Jun – Jul 2025	Develop index aggregation methodology
Aug – Sep 2025	Integrate Google Trends data; compile geospatial amenity weights
Oct – Nov 2025	Refine methodology
Dec 2025 – Jan 2026	Final report drafting and submission

Contribution:

This study aims to improve housing market monitoring by developing a monthly index using machine learning techniques. The ML model provides more timely insights into market trends by integrating lagged macroeconomic data, Google Trends, and geospatial amenity weights. This monthly index complements the quarterly RREPI, offering policymakers a more frequent view of market conditions. In turn, this helps understand emerging housing market shifts.

Author Affiliation:

Carmelita G. Esclanda-Lo is a Bank Officer leading Big Data Analytics Studies at the Department of Economic Statistics of the Bangko Sentral ng Pilipinas.

References:

- Alasia, A., Newstead, N., Kuchar, J., & Radulescu, M. (2025). Measuring proximity to services and amenities: An experimental set of indicators for neighbourhoods and localities. BSP Reports on Special Business Projects. https://www150.statcan.gc.ca/n1/pub/18-001-x/18-001-x2020001-eng.htm?utm_source=chatgpt.com
- Bency, A.J., Rallapalli, S., Ganti, R.K., Srivatsa, M., & Manjunath, B.S. (2017). Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 320-329. https://ieeexplore.ieee.org/document/7926625
- Case, K.E., & Shiller, R.J. (2003). Is There a Bubble in the Housing Market? *Brookings Papers on Economic Activity*, 2, 299-362. https://www.brookings.edu/articles/is-there-a-bubble-in-the-housing-market/
- Ghysels, E., Plazzi, A., Torous, W.N., & Valkanov, R.I. (2013). Forecasting Real Estate Prices Using Mixed-Frequency Models. *Journal of Real Estate Finance & Economics*, 46(2), 240-256. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1b3efaaf72813b20ae1605db8674c419256e49ab
- Hoxha, V. (2024). Comparative Analysis of Machine Learning Models for Predicting Housing Prices in Prishtina. *International Journal of Housing Markets and Analysis*. https://www.emerald.com/insight/content/doi/10.1108/ijhma-09-2023-0120/full/html
- Krasovytskyi, D. and Stavytskyy, A. (2024) "Predicting Mortgage Loan Defaults Using Machine Learning Techniques", Ekonomika, 103(2), pp. 140–160. doi:10.15388/Ekon.2024.103.2.8.
- Leamer, E.E. (2007). Housing IS the Business Cycle. *Journal of Economic Literature*, 45(2), 301-345. https://www.nber.org/system/files/working_papers/w13428/w13428.pdf
- Nneji, O., Brooks, C., & Ward, C.W.R. (2013). House Price Dynamics & Macroeconomic Changes: A Markov Switching Approach. *Real Estate Economics*, 41(2), 278-312. https://centaur.reading.ac.uk/31493/1/31493.pdf
- Wu, L., & Brynjolfsson, E. (2013). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. *ICIS 2013 Proceedings*. https://www.nber.org/system/files/chapters/c12994/revisions/c12994.rev0.pdf
- Yavuz Özalp, A., & Akinci, H. (2023). Comparison of tree-based machine learning algorithms in price prediction of residential real estate. Journal of Housing Markets and Analysis. https://dergipark.org.tr/tr/download/article-file/3421312