ESTIMATION OF A HEDONIC HOUSING PRICE INDEX FOR ALBANIA

A data gathering and analysis project, employing Natural Language Processing

Orion Garo,
Research Department
Bank of Albania

Prepared for the "19th South-Eastern European Economic Research Workshop",
Bank of Albania,
November 6 – 7, 2025.

RESEARCH FOCUS

Research Motivation: - Trends of Albania's residential housing market: difficult to track/estimate;

- per-square-meter real estate prices, absent in official statistics.

Objective: - Estimate 3 monthly indices of the housing properties' per-square-meter prices, for the case of Albania;

- Intended to be achieved by:
  - use of online-listed prices for housing property sales in Albania's [1] capital, as well as its [2] coastal and [3] inland areas;
  - identifying, gathering, cleaning, organizing, and calculating website-residing data (on Albania's residential property sales), so as to synthetize the price indices we are after.

THEORETICAL BACKDROP

The **Hedonic Method:**
- a robust method developed for estimating prices of fixed assets (such as residential properties) [Court, 1939];
- decomposition of the characteristics of similar heterogeneous assets, and effective estimation of separate values;
- appropriate to address the values heterogeneous and illiquid assets, such as residential properties.

Natural language processing:
- residential property price information retrieval from online real estate sale ads;
- the method involves analyzing language-specific (and often unspaced) text entries, with the objective of assembling a dataset containing: the sale availability date, the property price, and the property characteristics detailed in each entry (observation).

DATA ASSESSMENT

Raw data sources:
njoftime.com *2 of Albania's largest websites*
merrjep.al *on UGC transactions*

Sample size: 150,000 records

Sale prices in Tirana: 86%
Sale prices in coastal areas: 9%
Sale prices in inland areas: 5%

Observation inclusion:
Apartments *in urban/semi-urban apt. buildings*
Single-storey houses/villas *in rural/semi-urban areas*
Multi-storey houses/villas *in rural/semi-urban areas*

Time frequency: Monthly
Data points: 96
Sample timespan: Jan, 2017 – Dec, 2024

Categorization of the available data:
Residential property sales web ads for:

capital city *Tirana;*
coastal *Albania's coastal cities, towns, and rural areas;*
inland *Albania's inland cities, towns, and rural areas.*

METHODOLOGICAL FRAMEWORK

Scrape user-generated content: Scan residential property sale listings from the online raw data sources, and retrieve specific data fields of each listing (such as: date, lister, price, currency, and property description);
- Tools used: [i] Python framework; [ii] Scrapy (a sub-component of the framework);
- Assemble & keep raw data locally, in an electronic file;

Process saved raw data: Extract & validate key residential property details (the data variables: price, area, apt. or villa, floor, no. of rooms, land plot area, etc.). Remove outliers and missing values; merge, organize, and save the output in tabular form, for further analysis.
- Tools used: [i] Python framework; [ii] libraries such as: pandas, regex, numpy, itertools, etc; [iii] MS Excel formulas; [iv] PowerPivot.

Processed data estimation: De-seasonalize and obtain trendlines of tabulated data; fit variables into separate OLS regressions, to obtain date dummy coefficients and estimate the indices.
- Tools used: [i] JDemetra+ (X13 approach, Henderson filter); [ii] Eviews.

METHODOLOGICAL FRAMEWORK

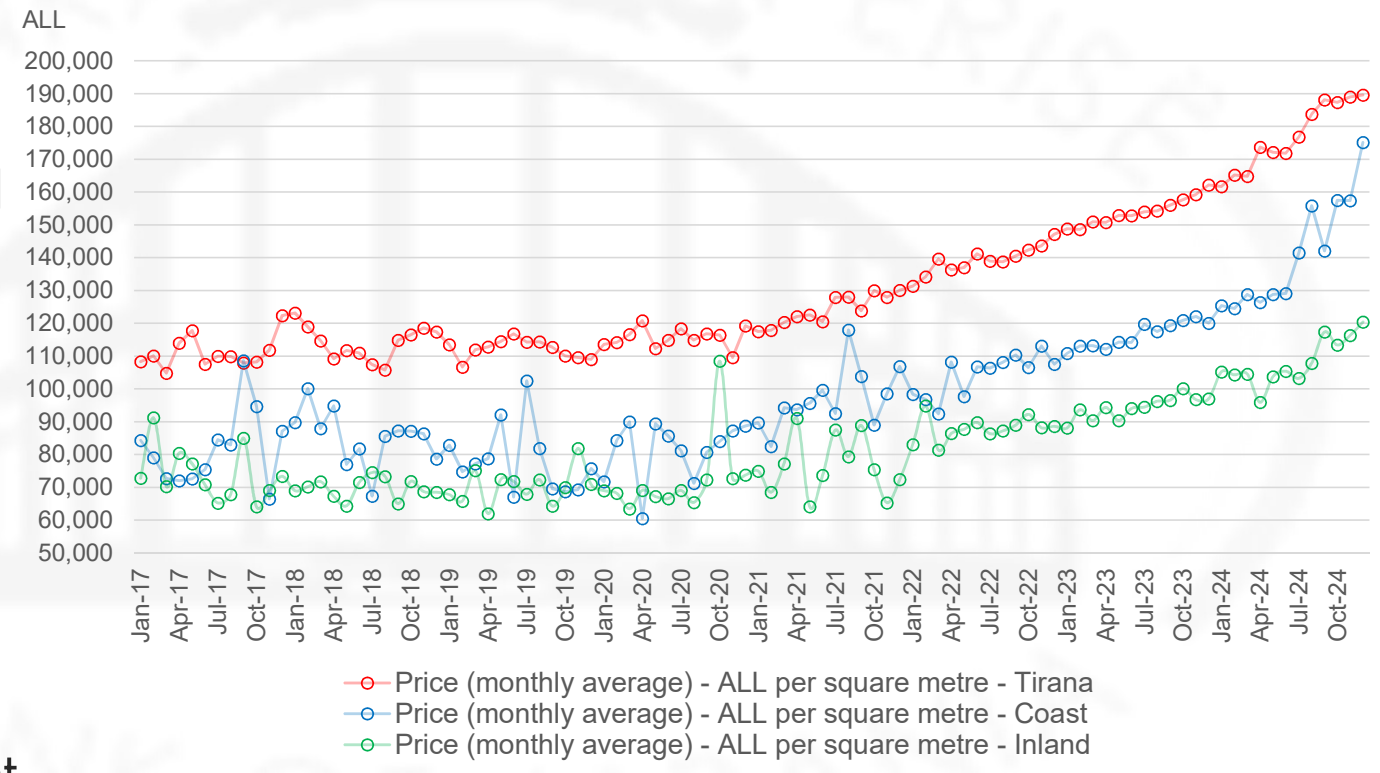
Processing challenges:

1. Related to the state of the text after data retrieval: very often unformatted, run-on text (scripto continua), where spaces between words are irregular or missing; extensive cleanup of Unicode escape sequences (such as: \\u00eb, \\u00e7...) necessary.
2. Distinguishing numeric contexts: setting the processing rules to correctly differentiate between: prices (Euro vs. Lek amounts), surface areas, counts (rooms, floors), phone and address numbers, etc.
3. Validating plausible real estate metrics: The realistic property metrics should be retrieved, avoiding errors and irrelevant numbers.
4. Distinguishing between words and phrases: meaningful compound phrases (such as: “offered for rent”, “land for sale” vs. “apartment for sale”, etc.) are necessary to isolate, so as to understand context, and process the record properly.
5. Free-form text from multiple sources: inconsistent formats, abbreviations, and unstandardized description structure.

METHODOLOGICAL FRAMEWORK

DATA VISUALIZATION

- **Monthly means** of residential housing listed prices (seasonally adjusted series).
- **Data divided** into **three** geographical areas.
- **Notable upward trends** for all three geographical areas, specifically the coast.



METHODOLOGICAL FRAMEWORK

- The econometric instrument: based on Kristo and Bollano's (2012) methodological groundwork on hedonic prices;
- Specifics of index estimation obtained from NasserEddine's (2017) hedonic regression approach;
- Semi-logarithmic form of the OLS specification, as detailed besides:
- Component X_{iq} of the estimation can allow for as many qualities of the property as raw data can provide ([i] square meters, and [ii] studio apartment or not, used for this estimation).
- Once model fitted and date dummy coefficients obtained, index is built by taking the average of three initial months, using it as reference period, and taking the antilogs of all other months' coefficients (NasserEddine, 2017).

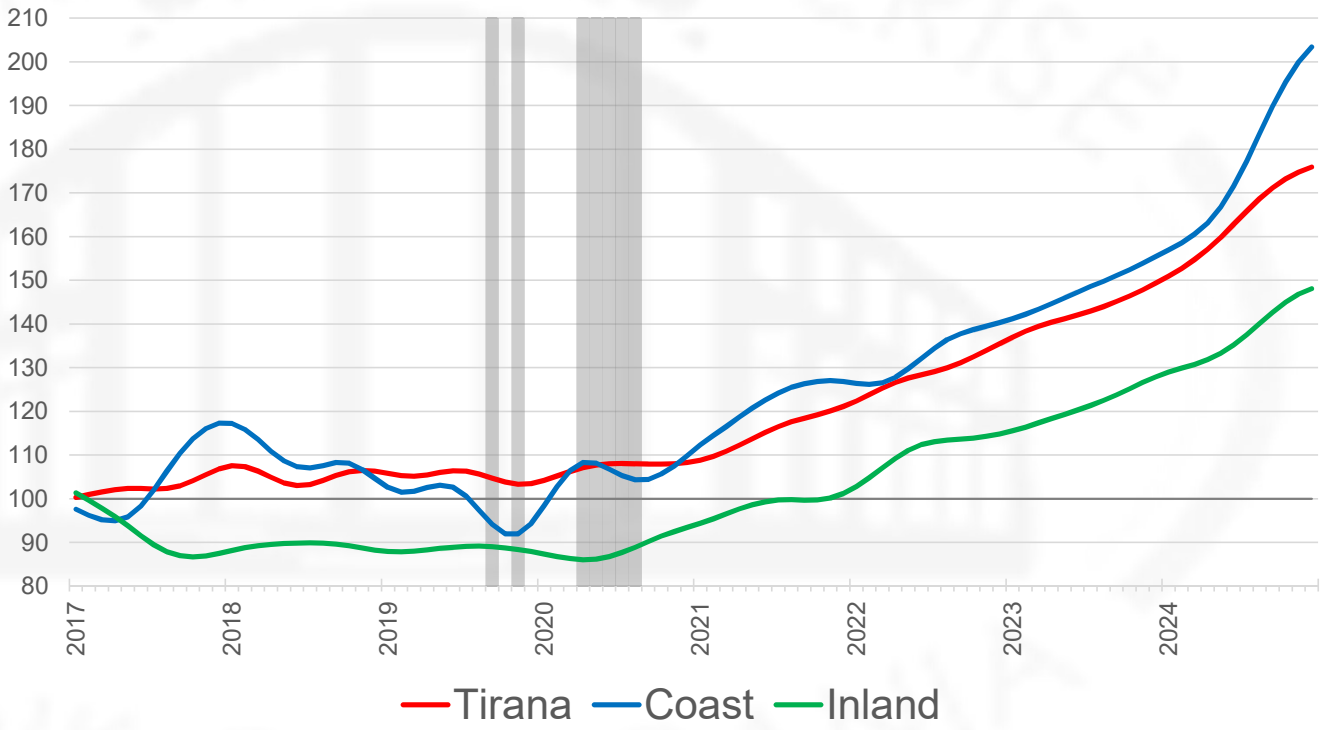
$$\text{Log } p_{it} = \alpha + \sum_{q=1}^{n} \beta_q X_{iq} + \sum_{t=1}^{m} \delta_t D_{it} + \varepsilon_{it}$$

where:

- p_{it} -> price of property i at month t;
- X_{iq} -> estimated value of the quality q of property i;
- D_{it} -> set of time dummy regressors pertaining to the listed price of property i;
- ε_{it} -> white noise residual of property i at month t.

OBTAINED RESULTS

- Country-wide data grouped by geographical area; each group 1 index.
- Constant and steady increase of house prices in Tirana
- House prices in coastal areas catch up Tirana's growth trend in 2020, and show a steeper trend thereafter.
- House prices in inland areas; more inert, but display growth after 2022.



Note: Henderson filter applied, to smooth series.