Predicting the Next Wave of Gentrification

A Data-Driven Analysis and Forecasting Model

Gabriele Argentieri & Marcel Alabart

DALAS - Sorbonne Université

November 3, 2025

Contents

Introduction and Problem Recap

This report details the data assessment and processing phase of our project, which aims to identify the drivers of gentrification and predict future hotspots in Paris, Barcelona, and Milan. This phase is critical as it involves transforming a wide array of raw, heterogeneous data into a unified, analysis-ready dataset.

Problem Recap

Our objective is to train a regression model to predict a "Gentrification Gap" index, defined for each neighborhood i as:

$$G_i = \text{PercentileRank}(P_i) - \text{PercentileRank}(I_i)$$

where P_i is the median property price and I_i is the median household income. The feature importances derived from this model will allow us to understand the key drivers of this phenomenon.

Data Processing and Integration Workflow

To construct our master dataset, we executed a multi-stage workflow designed to handle the diverse formats and structures of our raw data sources. The entire process was orchestrated using Python, primarily leveraging `pandas`, `geopandas`, and a sophisticated, custom-built web scraping framework.

Step 1: Geospatial Foundation as a Spatial Index

The first and most critical step was to establish a consistent geographical baseline for each city, which serves as the spatial index for all subsequent data integration.

- We obtained official GeoJSON boundary files for the administrative neighborhoods of Paris (*arrondissements*), Barcelona (*barris*), and Milan (*Nuclei di Identità Locale*).
- These files were loaded into a `geopandas` GeoDataFrame, creating a "base layer" containing a unique ID, name, and polygon `geometry` for each neighborhood.

Step 2: Advanced Web Scraping for Real Estate Data

Acquiring real estate data from commercial platforms like `Idealista` and `SeLoger` was the most significant technical challenge due to sophisticated anti-bot measures.

- **Methodology:** We developed a resilient scraping pipeline using **Selenium** in conjunction with **`undetected-chromedriver`**. This combination launches a modified version of the Chromium browser that is significantly harder for bot detection systems to identify as automated.
- **Multi-Layered Evasion Strategy:** To ensure stable, long-term scraping sessions without being blocked, we implemented a comprehensive strategy to mimic human behavior, including browser fingerprint masking with `selenium-stealth`, user-agent rotation, simulated human-like scrolling, and session management with randomized delays between batch requests.
- **Two-Phase Data Extraction:** Our process was divided into two distinct, resumable stages: (1) an initial scraper collected and saved unique property URLs, and (2) a second

process read from this file, visited each URL in batches, and scraped the detailed data incrementally to ensure data integrity.

Step 3: Source-Specific Processing and Spatial Joins

Each raw dataset was independently cleaned and then mapped to our geospatial base layer.

- **Real Estate and Tourism Data:** Both scraped property listings and Inside Airbnb data provided latitude and longitude coordinates. We converted these into `geopandas` GeoDataFrames of `Point` geometries. The core integration step was a **spatial join** (`gpd.sjoin`) using the "within" predicate to efficiently assign each property or listing to its correct neighborhood polygon.
- **Socio-Economic Data:** Official data from INSEE, INE, and ISTAT was joined to our base layer using standard `pandas.merge` on standardized neighborhood names or official codes.

Step 4: Final Merging and Feature Engineering

With all data sources joined to the base layer, we performed a final aggregation by neighborhood. This step produced our final, clean master table, upon which we engineered our target variable, the `gentrification_gap_index`.

The Master Data Schema

The rigorous processing workflow resulted in the following unified data structure. Each row represents a single neighborhood, allowing for direct comparison across all three cities.

Table 1: Final Master Schema for the Analysis-Ready Dataset

Category	Final Column Name	Description & Justification
Identifiers	`city`	The city name (e.g., "Paris"). Essential for comparative analysis.
	`neighborhood_id`	A unique, standardized ID for each neighborhood (e.g., `PAR_75003`).
	`neighborhood_name`	The official name of the neighborhood.
	`geometry`	The geographic polygon shape of the neighborhood. Used for all spatial joins.
Target Variable	`median_property_price_eur`	Median price of properties. Used to compute the target variable.
	`median_household_income_eur`	Median annual household income. Used to compute the target variable.
	`gentrification_gap_index`	**TARGET VARIABLE**. Calculated as `Rank(Price) - Rank(Income)`.
Socio-Economic Features	`population_density`	Inhabitants per km^2. Measures urban density.
	`delta_population_density`	Percentage change in population density over the analysis period.
	`pct_young_adults`	Percentage of population aged 25-39, a key gentrifier demographic.
	`pct_higher_education`	Percentage of residents with a university degree.

Table 1 – continued from previous page

Category	Final Column Name	Description & Justification
	`delta_higher_education`	Change in the percentage of residents with a university degree.
Housing & Tourism Features	`avg_price_per_m2`	Average price per square meter from real estate listings.
	`delta_price_per_m2`	Percentage change in price per m^2 over the analysis period.
	`airbnb_density`	Number of active Airbnb listings per 1,000 residents. Measures tourism pressure.
	`pct_entire_home_airbnb`	Percentage of Airbnbs listed as "Entire home/apt," indicating commercial use.
Urban Environment & Policy Features	`commercial_density_gentrify`	Number of "gentrifying" businesses (cafes, art galleries) per km^2.
	`distance_to_new_transport`	Distance (km) from the neighborhood's center to the nearest major new transport hub.
	`is_in_renewal_zone`	A binary flag (1/0) indicating if the neighborhood is in a designated urban renewal zone.

Current Data Assessment

- **Missing Values:** Missing data was a significant issue, particularly in the scraped real estate datasets where price or size were sometimes omitted. We handled this by dropping rows where the price (a critical field) was missing. For less critical numerical features, we used median imputation. For categorical features, missing values were filled with a distinct "Unknown" category.

- **Outliers:** We identified outliers primarily in property prices. These high-value properties were considered legitimate data points representing the high-end market rather than errors. To mitigate their influence on our analysis, we prioritized the use of robust statistical measures like the median over the mean.

- **Cleaning and Pre-processing:** This was the most time-intensive task. It involved extensive string manipulation to parse numerical values from text (e.g., "1.200.000 €"), standardizing addresses across different formats, and creating consistent geographical identifiers to enable the merging of disparate datasets.

- **Bias:** We have identified two primary sources of potential bias. First, **source bias** exists as our real estate data comes from major online portals, which may not capture private sales and could over-represent certain types of properties. Second, **measurement bias** is inherent in our target variable; while our "Gentrification Gap" is a robust and data-driven proxy, it remains a simplification of a deeply complex social phenomenon. We will acknowledge these limitations in our final report.