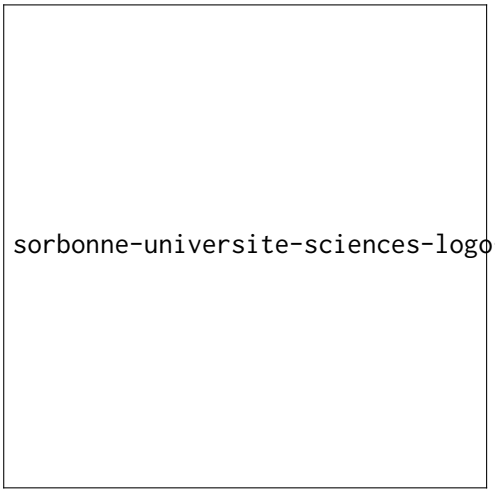

Predicting the Next Wave of Gentrification

A Data-Driven Analysis and Forecasting Model

Gabriele Argentieri & Marcel Alabart

DALAS - Sorbonne Université

November 6, 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 **price per square meter** 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. For each city, the process followed a consistent pattern, orchestrated using Python with the `pandas` and `geopandas` libraries.

Step 1: Geospatial Foundation as a Spatial Index

The first step was to establish a consistent geographical baseline for each city, serving as the spatial index for all subsequent data integration.

- We obtained official boundary files for the administrative neighborhoods of Paris, Milan, and Barcelona. For Barcelona, this involved extracting the specific **'Barri'** layer from the official multi-layer `BCN_UNITATS_ADM.zip` shapefile.
- These files were loaded into a `geopandas` GeoDataFrame and re-projected to a common CRS (EPSG:4326), creating a "base layer" with a unique ID, name, and polygon `geometry` for each neighborhood.

Step 2: Heterogeneous Data Sourcing

Data acquisition required a multi-faceted approach, combining direct downloads from open data portals with targeted web scraping.

- **Open Data Portals:** Socio-economic data for Barcelona, such as population demographics, income, and education levels, were sourced from the **Open Data BCN** portal. These datasets were provided at the fine-grained 'Seccio$_c ensal$'$(censustract)level, requiringsubsequentaggregat$
- **Web Scraping for Market Data:** For market-driven data in Paris and Milan, we utilized a resilient web scraping pipeline to acquire real estate and tourism data. This involved using Selenium with human-like behavior simulation to navigate commercial platforms and retrieve point-based data (e.g., property listings).
- **Third-Party Data Aggregators:** Tourism pressure was measured using data from **Inside Airbnb**, which provides periodic snapshots of listings data for our target cities.

Step 3: Source-Specific Processing and Spatial Joins

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

- **Point-based Data (Real Estate, Tourism):** For scraped listings with latitude/longitude coordinates, we performed a **spatial join** (`gpd.sjoin`) to assign each point to its correct neighborhood polygon.
- **Area-based Data (Socio-Economic):** Official data, like Barcelona's census data, was provided with official neighborhood codes ('Codi$_b arri$'). $We performed extensive cleaning, such as parsing n$ $tract level to the neighborhood level using the$ `median` $to ensure robustness to outliers. This aggregated data was t$

Step 4: Final Merging and Master Dataset Creation

With all data sources processed to the neighborhood level, we performed a final **inner join** on neighborhood ID and year. This crucial step ensures that our final master dataset contains only complete, consistent records for a defined analysis period, resolving issues of differing time coverages across datasets.

The Master Data Schema

The rigorous processing workflow resulted in the following unified data structure. Each row represents a single neighborhood in a given year, 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. Essential for comparative analysis.
	`neighborhood_id`	A unique, standardized ID for each neighborhood.
	`neighborhood_name`	The official name of the neighborhood.
	`geometry`	The geographic polygon shape of the neighborhood.
Target Variables	`median_price_per_m2`	**Normalized price metric**. For Barcelona, this is the median cadastral value per m². For Milan/Paris, it's the median market price per m². Used to compute the target variable.
	`median_household_income`	Median annual household income in euro.
	`gentrification_gap_index`	**TARGET VARIABLE**. Calculated as `Rank(Price) - Rank(Income)`.
Socio-Economic Features	`population_density`	Inhabitants per km². 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	`airbnb_density`	Number of active Airbnb listings per 1,000 residents.
	`pct_entire_home_airbnb`	Percentage of Airbnbs listed as "Entire home/apt," indicating commercial use.
Urban Environment Features	`is_in_renewal_zone`	A binary flag indicating if the neighborhood is in a designated urban renewal zone.

Current Data Assessment

Our initial data exploration and validation have yielded the following key insights and resolutions:

- **Data Comparability (Cadastral vs. Market):** Our most significant challenge is the use of different value types: official cadastral values for Barcelona and market values for Milan/Paris. We address this by: **(1)** Normalizing all price data to **price per square meter**, and **(2)** Using a **rank-based target variable**, which is robust to differences in absolute scales and focuses on the relative ordering of neighborhoods within each city.
- **Missing Values:** Differing time periods across raw datasets initially created many missing values. This was resolved by using an **inner merge** to define a core analysis period (e.g., 2018-2023) for which all key data points are available, resulting in a complete and consistent final dataset.
- **Outliers:** We detected outliers in population density and income, corresponding to real-world phenomena (e.g., industrial zones or extremely wealthy enclaves). By consistently using the **median** for aggregation and analysis, we have ensured our metrics are robust to these extreme values.
- **Bias:** We acknowledge two primary biases. **Source bias** exists as market data from portals may not capture all property sales. **Measurement bias** is inherent in our target variable, as the "Gentrification Gap" is a proxy for a complex social phenomenon. These limitations will be explicitly discussed in our final analysis.

Data Processing and Modeling Pipeline

With a clean and validated master dataset, our project now enters the modeling phase. The following pipeline outlines our plan to move from data to actionable insights.

Modeling Flight Plan

Objective: Predict the 'gentrification$_g ap_i ndex$' *attheendofouranalysisperiodandidentifythemostimportan*

Step 1: Final Feature Engineering:

- **Define Analysis Period:** We will establish a consistent analysis period across all cities based on our data availability, likely 2018 to 2023.
- **Create 'delta**$\cdot Features$: $We willcalculatethepercentagechangeforkeytime-varyingfeatures(e.g.,$'$population_density$', '$pct_higher_education$'$)betweenthestartandendofthisp$

- **Assemble Modeling Matrix: We will create a final cross-sectional dataset where each row represents one neighborhood. The features (X) will be the socio-economic indicators from the start year (2018) plus the calculated 'delta**$\cdot features.Thetargetvariable(\mathbf{y})willbethe$'$gentrification_g ap_i ndex$'$fromtheendyear$

Step 2: Algorithm and Model Selection:

- **Algorithm:** We have selected the **XGBoost Regressor** as our primary model. Its gradient boosting framework is highly effective for tabular data, robust to outliers, and ideal for capturing complex, non-linear interactions between urban features.
- **Task:** The problem is framed as a **regression task** to predict the continuous 'gentrification$_g ap_i ndex$'.

Step 3: Data Preprocessing for Modeling:

- **Data Splitting:** The combined dataset from all three cities will be split into an 80% training set and a 20% hold-out test set to ensure an unbiased evaluation of our final model.
- **Scaling:** Although XGBoost is insensitive to feature scaling, we will apply 'StandardScaler' to the feature matrix as a best practice, facilitating potential comparisons with other models like regularized linear regression.

Step 4: Model Evaluation Metric:

- **Metric:** Model performance will be assessed using **R-squared (R^2)** on the test set. This metric will quantify the proportion of variance in the gentrification index that our model can explain, providing a clear measure of its predictive power. We will also report the **Mean Absolute Error (MAE)** to understand the average prediction error in terms of index points.

Step 5: Feature Importance Analysis:

- **Methodology:** The core goal of our analysis is to identify the drivers of gentrification. Upon validating our model, we will use **SHAP (SHapley Additive exPlanations)** to interpret the trained XGBoost model.

- **Deliverable:** The SHAP analysis will produce summary plots ranking the features by their overall impact on the model's predictions. This will provide data-driven insights into which factors—tourism pressure, demographic shifts, or urban policy—are the most significant predictors of gentrification across our three European cities.