Rental Price Analysis in Estonian Cities: A Geospatial Approach

Understanding Spatial Factors Affecting Rental Prices in Tartu and Tallinn

Mari Lee Lumberg • Karl Martin Puna • Daniel Henri Trump
Data Science Course • University of Tartu • Project: Real-Estate-Pricerange-Predictor

Project ID H9

1. Introduction & Motivation

Motivation: Urban rental markets are spatially structured - location and accessibility to services influence prices. This project investigates how network-distance to amenities and spatial autocorrelation shape rental prices in two Estonian cities.

Goal: Quantify spatial factors affecting rent per m² in Tartu and Tallinn, and build predictive models that account for spatial dependence.

2. Data & Sources

Rental listings: KV.ee (web-scraped)

- Tallinn: 1,359 → 1,345 final
- Tartu: 450 → 448 final

Street network & amenities: OpenStreetMap via OSMnx—amenity, shop, and leisure categories aggregated

Geospatial data of neighbourhoods and districts: Maa- ja Ruumiamet (Estonian Land and Spatial Development Board)

3. Data Cleaning

Rental listings cleaning: Started with 20 columns (e.g., id, url, price, latitude, longitude). Tallinn: 1,359 records; Tartu: 450 records.

- Removed whitespaces, replaced "?" with NaN
- Extracted price (€) and area (m²) from strings (e.g., '450 € 6.79 €/m²')
- Differentiated floor (e.g., 3) and total floors (e.g., 5) from '3/5'
- Converted values to numeric; removed properties missing critical fields (price, area_sqm, rooms)
- Removed columns with >20% missing data; final 10 columns for both cities

OSM data cleaning: Downloaded street networks and POIs via OSMnx, projected to EPSG:3301.

- Split geometries (points vs. polygons), converted polygons to centroids
- Filtered amenities/shops/leisure by relevance (e.g., kept education, health, food; removed irrelevant like 'access')
- Aggregated into categories (e.g., food_drink, parks_green); kept shop types appearing ≥3 times

4. Spatial Methodology

Network distances: Walking distance along street network using Dijkstra's algorithm (OSMnx), not straight-line

Spatial autocorrelation: Bivariate Moran's I tests between price and amenity proximity

Modeling: XGBoost classifier (3 classes by €/m² quantiles) with spatial cross-validation (5-fold spatial CV, 4km blocks)

5. Engineered Features

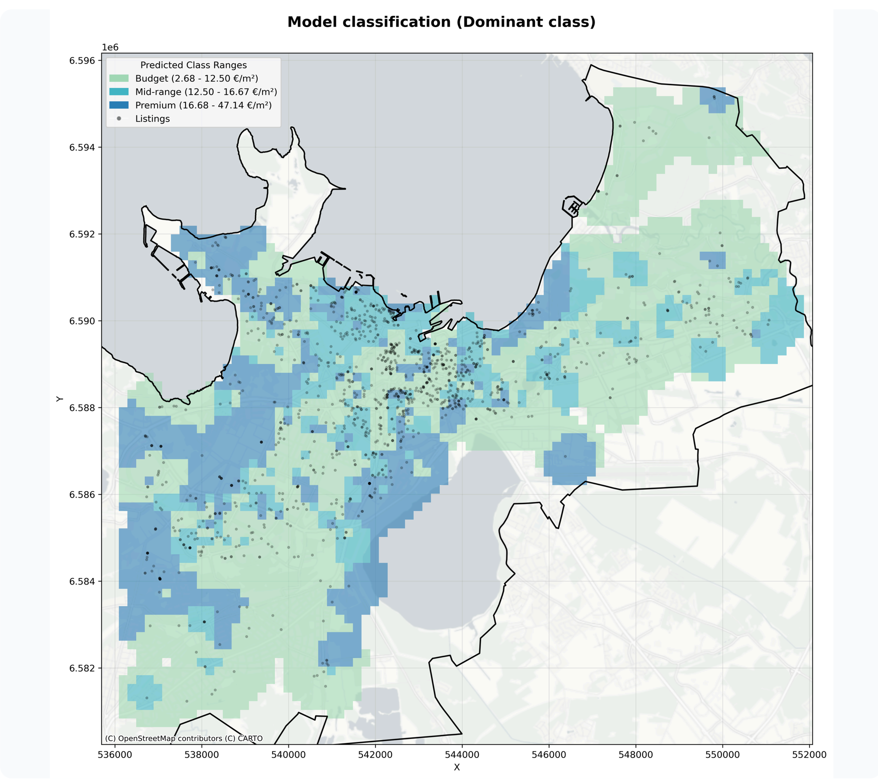
1. Price per sqm: price / area_sqm
2. Aggregated Services: Services grouped into categories
3. Distance to Services: Network distances to each category
4. Distance to Center: Direct distance to city center
5. Condition Mapping: Numerical condition value
6. Building Age: 2025 - build_year
7. Floor Ratio: floor / total_floors
8. Room Density: rooms / area_sqm
9. Condition-Age Interaction: condition_score × building_age
10. Floor-Total Interaction: floor_ratio × total_floors
11. Price Categories: Binned price ranges
12. Out-of-Fold Mean: Global mean from training folds
13. District OOF Mean: District-level mean price
14. Grid OOF Mean: 2km spatial grid mean
15. Amenity Density: Service count within 800m
16. Walkability Index: Composite (0-100) weighted amenity proximity
17. Amenity Diversity: Distinct categories within 800m

6. The Education Paradox

Observation: Greater distance from primary/secondary schools correlates with higher rents in both cities. Schools in peripheral neighborhoods (suburban).

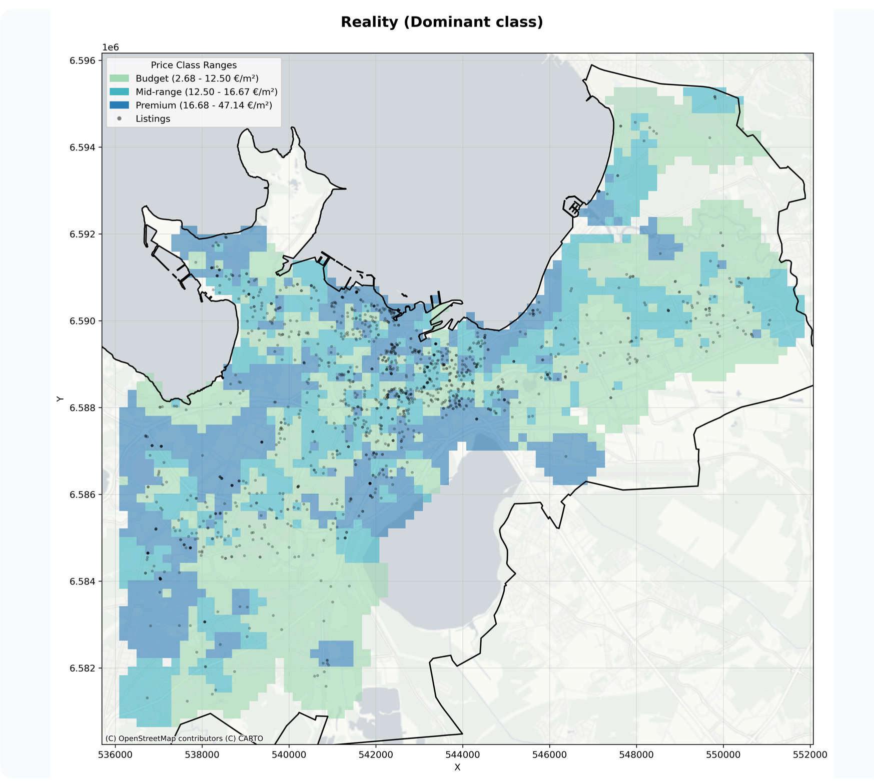
Possible explanations:

- Schools in peripheral neighborhoods
- Older housing near schools
- Families buy vs. rent
- Traffic/noise externalities



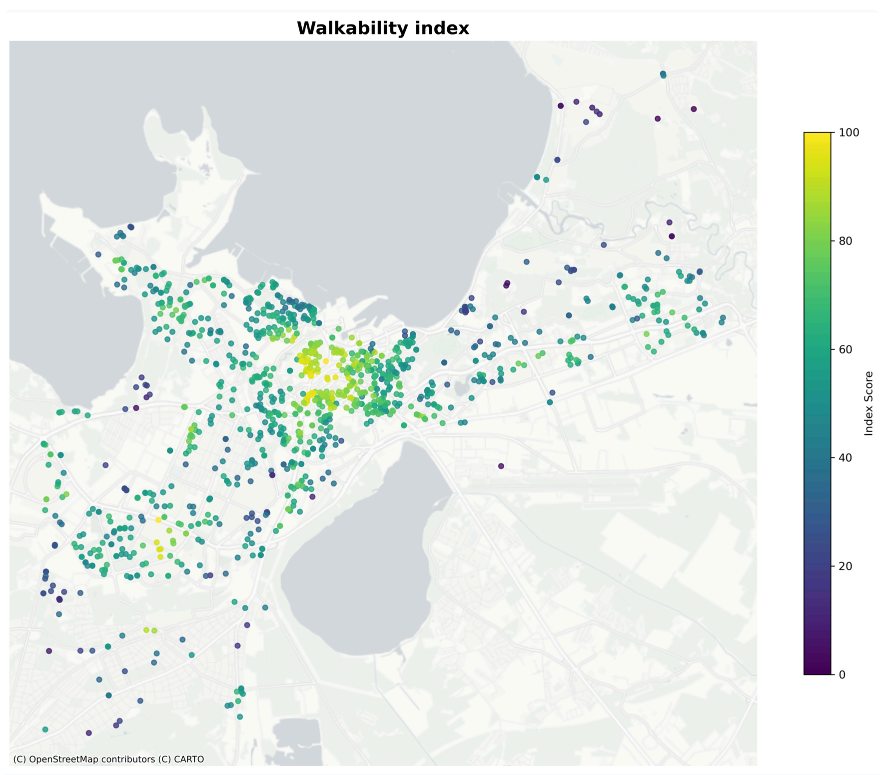
Tallinn Model Classification

Predicted dominant price class per 200m grid cell using inverse distance weighting. Green = Budget (€2.68-12.50/m²), Blue = Mid-range (€12.50-16.67/m²), Purple = Premium (€16.68-47.14/m²). Black dots = listings.



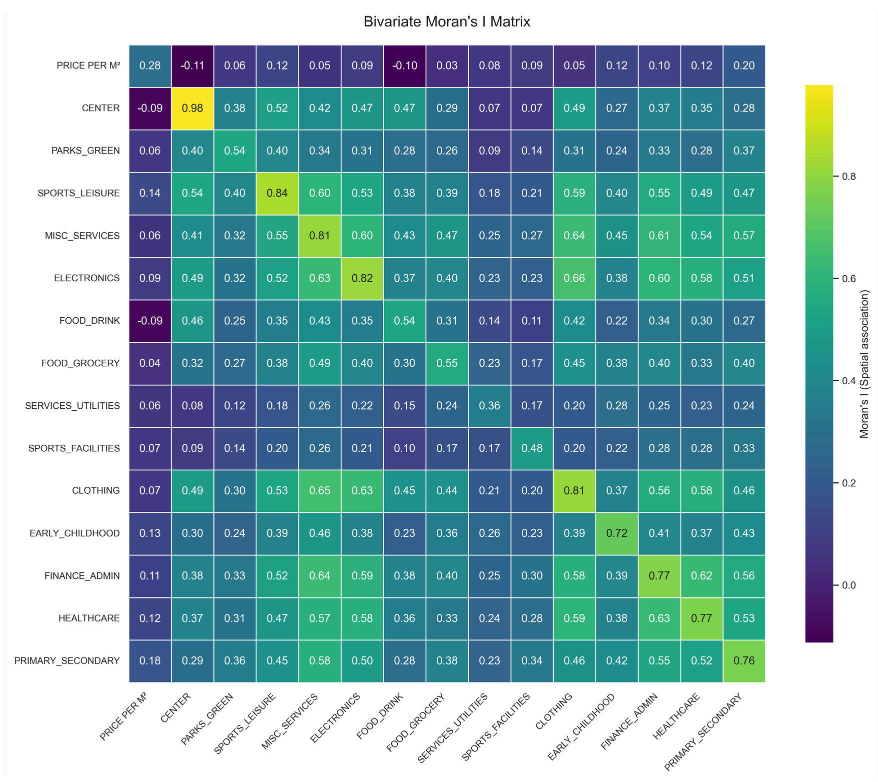
Tallinn Reality Map

Actual dominant price class based on listings, interpolated to 200m grid. Shows real spatial patterns; black dots = property listings.



Walkability Index

Composite score (0-100) based on proximity to weighted amenities (e.g., food/drink high, parking low). Higher values indicate better walkability and access to essential services.



Bivariate Moran's I Matrix

Spatial correlations in Tallinn. Positive values (≥0.5): High-high matches. Negative values (≤-0.1): High-low matches. -0.1 to 0.1: Weak/no relation.

Reading the Correlation Matrix

Logic for bivariate Moran's I: Measures spatial correlation between two variables, accounting for spatial dependence.

Key insights: Top row shows price per m² vs. distances to amenities—high price associates with lower distance to center and higher distance to primary education (schools in suburban areas). Tallinn is monocentric (center correlates with most amenities). Sports/leisure, misc services, electronics cluster (multicollinearity). Price vs. price shows weak segregation (renovations blur rich/poor areas).

7. Model Performance

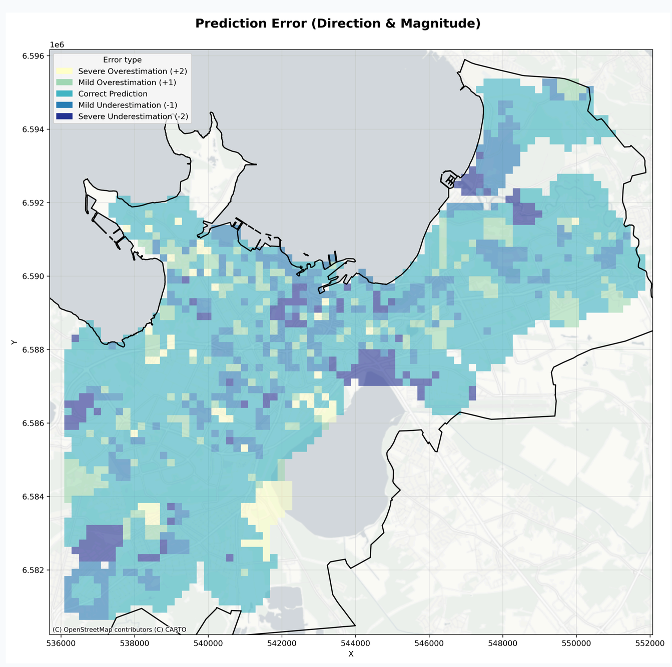
Classification: 3 classes (Budget / Mid-range / Premium by €/m² quantiles)

Performance (Tallinn):

71.4% Accuracy
0.552 Kappa
0.277 R^2

Key features: Target encoding (districts, 2km grids), walkability index, amenity diversity

Challenge: Strong spatial autocorrelation—nearby properties influence each other



Understanding Prediction Error

Color-coded error:

- **Dark Blue = Underestimation:** Model predicts lower class than actual (-2 to -1 class difference).
- **Blue = Correct:** No error (0 difference).
- **White = Overestimation:** Model predicts higher class than actual (+1 to +2 class difference).

Insight: Reveals spatial biases (e.g., overestimation in suburban areas, underestimation near amenities).

8. Conclusions

Spatial autocorrelation dominates: Neighborhood effects strong (ρ≈0.48 in Tallinn)

Consistent patterns: Proximity to food/drink amenities increases rent

Education paradox: Distance from primary/secondary schools associates with higher rent—likely urban structure, not causal.

Tartu limitations: With only 448 listings, reliable predictions or assumptions couldn't be made due to insufficient data size

Data Sources: KV.ee • OpenStreetMap • OSMnx • XGBoost • Maa- ja Ruumiamet (Estonian Land and Spatial Development Board)

Repository: Real-Estate-Pricerange-Predictor

Full code and figures available in the project repository

Scan for GitHub

For practitioners: Include spatial terms, use blocked CV, export maps as vectors