Supplementary Material

Machine Learning for Property Valuation During a Housing Boom: Predictive Evaluation, Extrapolation Bias and Explainability in the Netherlands

Contents

A Variable definitions **2**

B Construction of the neighborhood panel **5**

C Imputation of missing values **6**

D Model settings and additional results **7**

E Dashboard **10**

F Additional results **10**

This supplement documents the variables used, the construction of the neighborhood panel, the imputation of missing values, model settings, and the dashboard. Section, table and figure numbers prefixed with S refer to this document; all other references refer to the main text.

A Variable definitions

Table S1 lists the variables of the neighborhood panel, taken from the CBS *Kerncijfers wijken en buurten* releases 2014–2024 (Centraal Bureau voor de Statistiek, 2025b); only variables present in every release are included. Table S2 lists the variables of the Utrecht property dataset, taken from the *WOZ-waardeloket* (WOZ-waardeloket, 2025). Variable names follow the original sources.

Table S1. Variables of the neighborhood panel (CBS)

Variable	Description
Identifiers and target	
`regio`	Region name
`gm_naam`	Municipality name
`gwb_code`	Region code
`ind_wbi`	Indicator of code and boundary changes relative to the previous year
`year`	Year of observation
`woz`	Mean WOZ value of residential properties (€1,000); target variable
Population and households	
`a_inw`	Number of inhabitants
`a_man`	Male inhabitants
`a_vrouw`	Female inhabitants
`a_geb`	Number of births
`p_geb`	Birth rate
`a_ste`	Number of deaths
`p_ste`	Death rate
`a_hh`	Number of households
`g_hhgro`	Average household size
`bev_dich`	Population density
Housing stock	
`a_woning`	Number of dwellings
`p_1gezw`	Share of single-family dwellings
`p_mgezw`	Share of multi-family dwellings
`p_bewndw`	Share of occupied dwellings
`p_leegsw`	Share of vacant dwellings
`p_koopw`	Share of owner-occupied dwellings
`p_huurw`	Share of rental dwellings
`p_wcorpw`	Share of dwellings owned by housing associations
`p_ov_hw`	Share of dwellings owned by other landlords
Energy consumption	
`g_ele`	Average electricity consumption, all dwellings
`g_ele_vw`	Average electricity consumption, detached houses
`g_ele_hu`	Average electricity consumption, rental dwellings
`g_ele_ko`	Average electricity consumption, owner-occupied dwellings
`g_gas`	Average natural gas consumption, all dwellings
`g_gas_vw`	Average natural gas consumption, detached houses

Continued on next page

Table S1 continued

Variable	Description
g_gas_hu	Average natural gas consumption, rental dwellings
g_gas_ko	Average natural gas consumption, owner-occupied dwellings
Income	
a_inkont	Persons with personal income in private households
g_ink_po	Average income per income recipient
g_ink_pi	Average income per inhabitant
p_ink_li	Share of persons in the lowest 40% of the national income distribution
p_ink_hi	Share of persons in the highest 20% of the national income distribution
p_hh_li	Share of households in the lowest 40% of the national income distribution
p_hh_hi	Share of households in the highest 20% of the national income distribution
p_hh_lkk	Share of households with a low income
p_hh_osm	Share of households at or around the social minimum
Social security	
a_soz_wb	Recipients of social assistance benefits
a_soz_ao	Recipients of disability benefits
a_soz_ww	Recipients of unemployment benefits
a_soz_ow	Recipients of state old-age pensions
Business establishments	
a_bedv	Total number of business establishments
a_bed_a	Agriculture, forestry and fisheries
a_bed_bf	Industry and energy
a_bed_gi	Trade and hospitality
a_bed_hj	Transport, information and communication
a_bed_kl	Financial services and real estate
a_bed_mn	Business services
a_bed_ru	Culture, recreation and other services
Motor vehicles	
a_pau	Number of passenger cars
a_bst_b	Passenger cars running on petrol
a_bst_nb	Passenger cars running on other fuels
g_pau_hh	Passenger cars per household
g_pau_km	Passenger cars per km^2
a_m2w	Number of motorcycles
Facilities	
g_afs_hp	Average distance to a general practitioner
g_afs_gs	Average distance to a large supermarket
g_afs_kv	Average distance to a day-care center
g_afs_sc	Average distance to a school
g_3km_sc	Average number of schools within 3 km
Area, postal codes and urbanity	
a_opp_ha	Total area (ha)
a_lan_ha	Land area (ha)
a_wat_ha	Water area (ha)
pst_mvp	Most common postal code
pst_dekp	Coverage of the most common postal code
ste_mvs	Degree of urbanity, derived from address density
ste_oad	Address density: mean number of addresses per km^2 within a radius of 1 km

Table S2. Variables of the Utrecht property dataset (WOZ-waardeloket and BAG)

Variable	Description
`postcode`	Postal code (numerical part)
`huisnummer`	House number (identifier only; not used as predictor)
`straat`	Street name (label-encoded)
`buurt`	Neighborhood (label-encoded)
`bouwjaar`	Year of construction
`grondoppervlakte`	Lot size including garden; recoded to 1 if the dwelling has its own plot and 0 otherwise (mostly apartments)
`oppervlakte`	Floor area (m^2)
`woz_2014-woz_2022`	WOZ value in each year 2014 to 2022 (€)
`woz_2023`	WOZ value in 2023 (€); target variable
`latitude, longitude`	Geographic coordinates of the property

B Construction of the neighborhood panel

Administrative boundaries in the Netherlands change frequently through municipal mergers and redivisions of districts and neighborhoods (Figure S1). Neighborhoods are linked across releases only when the CBS indicator `ind_wbi` shows that their delineation is unchanged, so that values refer to the same territory in every year. Figure S2 shows the variables with at least 20% missing values in the merged panel before imputation.

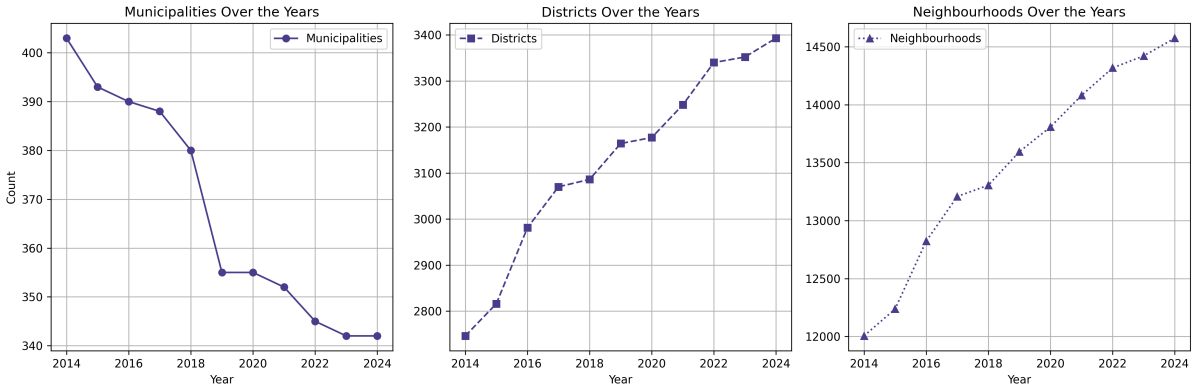

Figure S1. Number of municipalities, districts and neighborhoods in the CBS releases 2014–2024

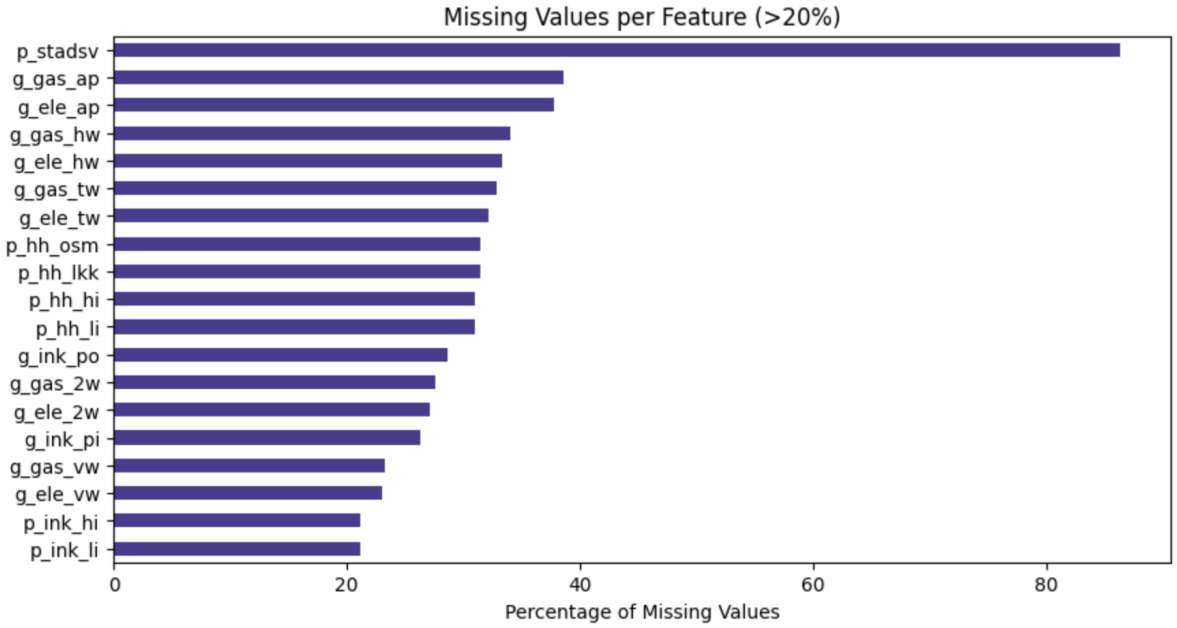

Figure S2. Variables of the neighborhood panel with at least 20% missing values before imputation

C Imputation of missing values

Figure S3 summarizes the rule-based imputation procedure described in Section 4.1 of the main text. Each variable is processed separately for each neighborhood, using its eleven yearly observations.

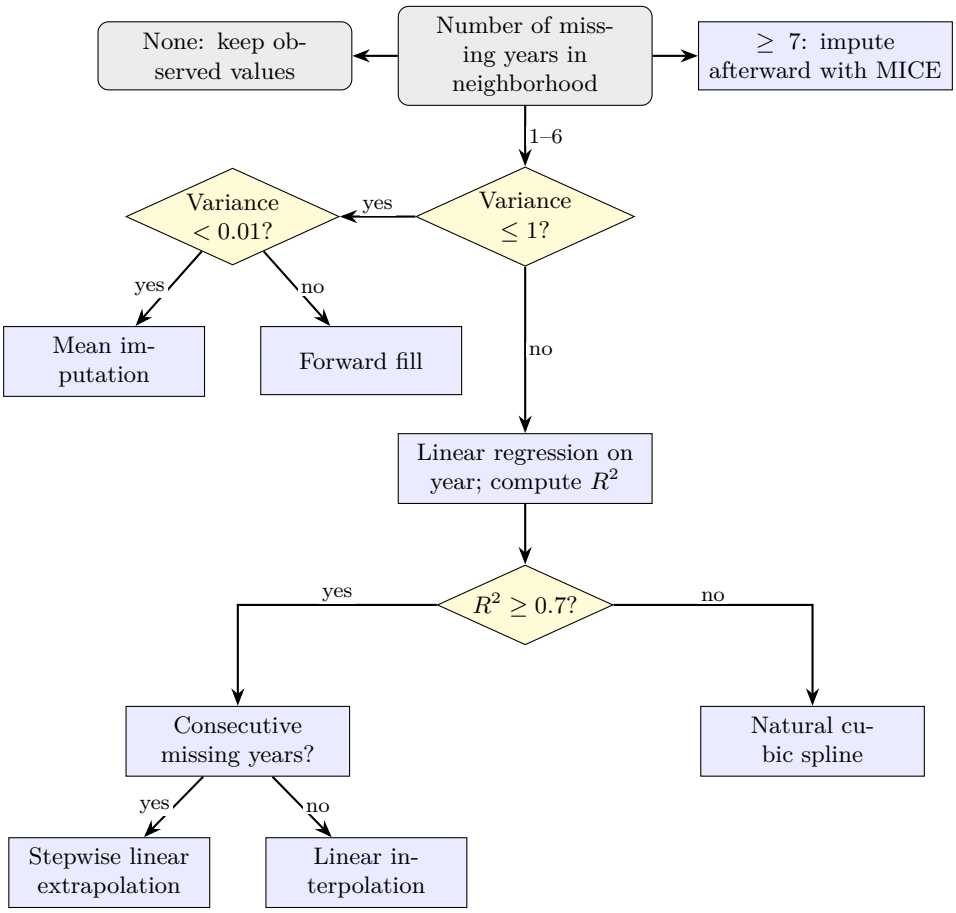

Figure S3. Decision procedure for imputing missing values in the neighborhood panel, applied to each variable and neighborhood. Values left missing, mainly where seven or more years are missing, are completed afterward with multivariate imputation by chained equations (MICE; scikit-learn iterative imputer on standardized values of all numerical variables, at most 50 iterations)

D Model settings and additional results

Table S3 reports the model configurations, taken from the analysis code. Table S4 lists the predictors retained by recursive feature elimination (RFE). Table S5 reports the five most important predictors of the 2023 valuation in Utrecht. Table S6 and Figures S4 and S6 report the results of the original analysis run, which the main text reproduces with current library versions. Figure S5 shows the impurity-based importance scores of the trained models.

Table S3. Model configurations

Model and use	Configuration
XGBoost grid search (neighborhoods)	5-fold cross-validation, R^2 criterion; 200/300/500 trees × learning rate 0.01/0.05/0.1 × maximum depth 3/5/7
XGBoost, evaluation (both datasets)	700 trees, learning rate 0.1, maximum depth 5
Random forest, evaluation (both datasets)	200 trees, $\sqrt{p}$ predictors per split, maximum depth 40, minimum 5 observations to split, minimum 2 per leaf, bootstrap samples
RFE (neighborhoods)	Random forest with 20 trees; 5 predictors removed per step; at least 10 retained; 3-fold cross-validation on 2014–2023
RFE (Utrecht)	Random forest with 50 trees; 5 predictors removed per step; at least 5 retained; 3-fold cross-validation on the training sample
XGBoost, forecasting (neighborhoods)	500 trees, learning rate 0.1, maximum depth 5; 40 predictors plus WOZ lags of 1–3 years and the 3-year rolling mean
XGBoost, forecasting (Utrecht)	200 trees, learning rate 0.1, maximum depth 7
Linear regression (benchmark)	Ordinary least squares on standardized predictors (same predictors as the ensembles)
Naive benchmarks	Persistence: previous year's value of the same unit; drift: previous value × mean growth factor of the last observed year
XGBoost robustness checks (neighborhoods)	Evaluation configuration; (i) without `gm_naam`, `gwb_code`, `pst_mvp`; (ii) log WOZ target, positive training values only; (iii) plus WOZ lags of 1–3 years and the rolling 3-year mean, missing lags handled natively
SHAP values	Exact tree SHAP of the XGBoost evaluation model (`pred_contribs`), 2024 test set
Software	Python; XGBoost 3.4.1 and scikit-learn 1.9.1 for the results in the main text

Table S4. Predictors retained by recursive feature elimination

Dataset	Retained predictors
Neighborhoods (18 of 68)	`gm_naam`, `gwb_code`, `year`, `bev_dich`, `a_woning`, `p_1gezw`, `p_koopw`, `g_ele`, `g_ele_hu`, `g_ele_ko`, `g_gas_ko`, `g_ink_po`, `g_ink_pi`, `p_ink_li`, `p_hh_hi`, `p_hh_lkk`, `pst_mvp`, `ste_oad`
Utrecht (10 of 17)	`woz_2022`, `oppervlakte`, `woz_2021`, `woz_2020`, `woz_2018`, `woz_2019`, `latitude`, `longitude`, `woz_2014`, `woz_2017`

Note: For Utrecht, cross-validation retained all 17 predictors; the reduced set consists of the ten with the highest importance.

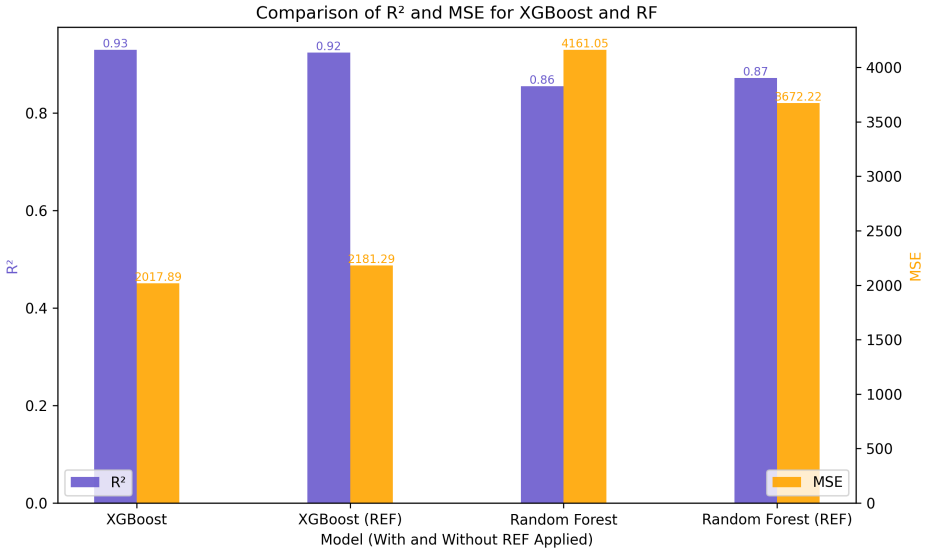

Figure S4. Coefficient of determination (R^2) and mean squared error (MSE, in (€1,000)^2) of XGBoost and random forest on the 2024 neighborhood test set, with and without recursive feature elimination (labeled REF in the figure), original analysis run

Table S5. Five most important predictors of the 2023 WOZ value in Utrecht (random forest importance after RFE)

Predictor	Importance
`woz_2022`	0.963
`oppervlakte`	0.004
`woz_2021`	0.003
`woz_2020`	0.003
`woz_2018`	0.003

Table S6. Predictive accuracy in the original analysis run

Dataset	Model	R^2	MSE	Replicated MSE
Neighborhoods, 2024 (€1,000)	XGBoost	0.93	2,017.89	1,989.53
	XGBoost (RFE)	0.92	2,181.29	2,159.35
	Random forest	0.86	4,161.05	4,148.53
	Random forest (RFE)	0.87	3,672.22	3,660.37
Utrecht properties, 2023 (€)	XGBoost	0.9833	485,765,319	484,996,703
	XGBoost (RFE)	0.9812	545,186,111	–
	Random forest	0.9833	486,371,382	486,371,382
	Random forest (RFE)	0.9821	519,576,823	–

Note: MSE in the squared unit of the target. The replicated values were obtained with XGBoost 3.4.1 and scikit-learn 1.9.1 and are those reported in the main text; the Utrecht RFE models were not re-run.

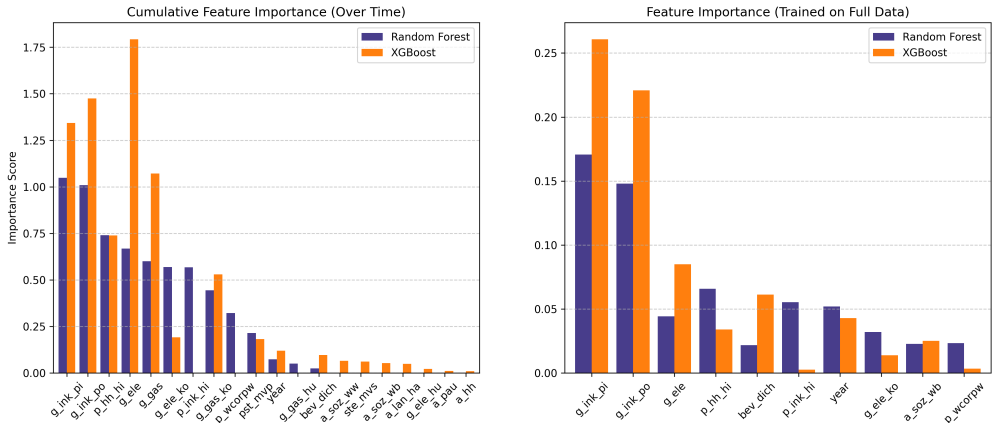

Figure S5. Impurity-based feature importance of random forest and XGBoost for the neighborhood panel: cumulative importance over models trained on expanding windows 2014–2024 (left) and importance of models trained on the full panel (right)

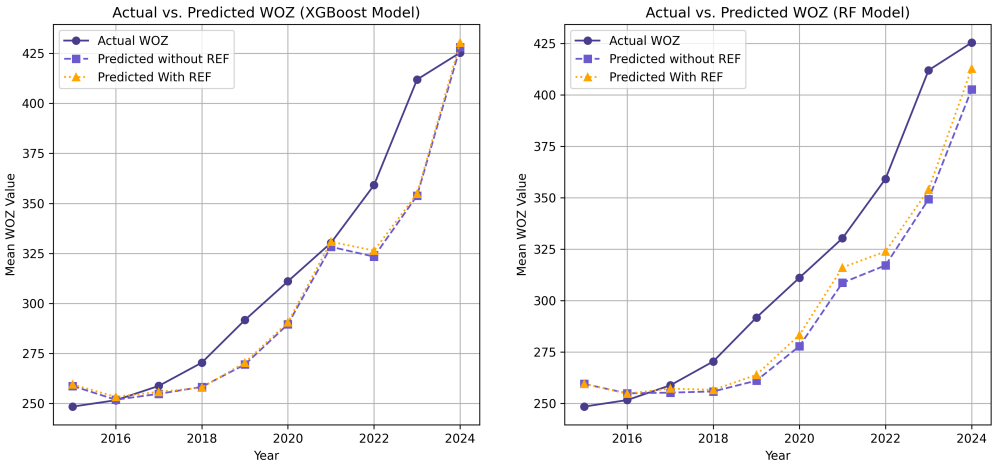

Figure S6. Mean actual and predicted neighborhood WOZ values (€1,000) by year in the original analysis run, for XGBoost (left) and random forest (right), with and without recursive feature elimination

E Dashboard

The dashboard (name and address withheld for anonymous review), built with Streamlit and Plotly, contains two maps. The national map (Figure S7) shows the predicted percentage change in the WOZ value of each neighborhood for a year chosen from a menu, and displays its most important predictors on hovering. The Utrecht map (Figure S8) shows the predicted WOZ values of individual properties on top of their neighborhoods, with a slider to filter properties by value.

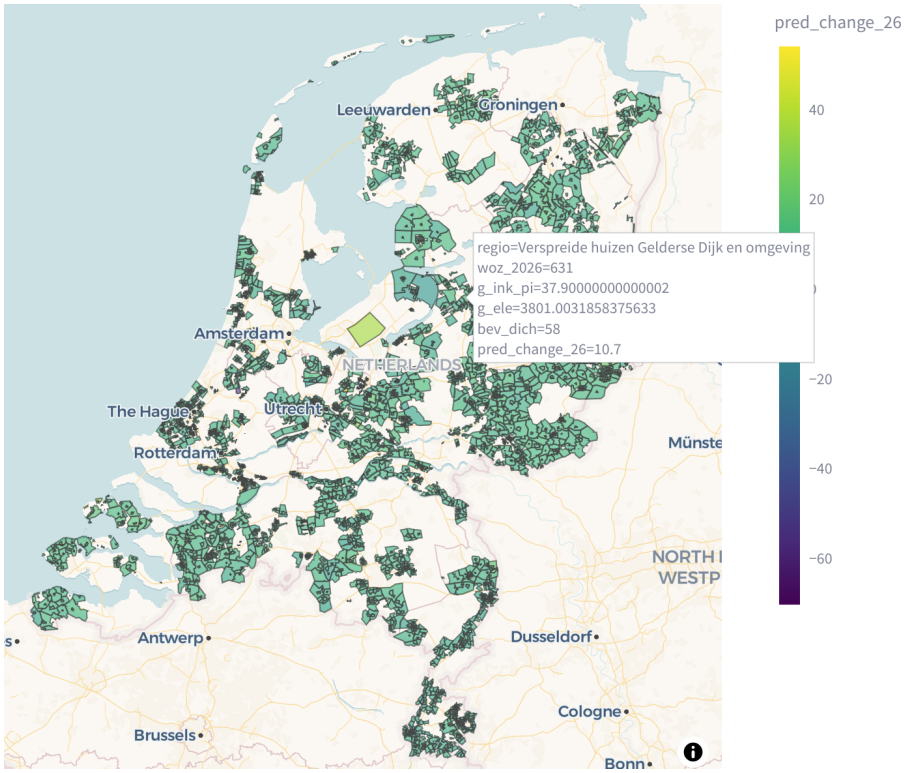

Figure S7. Dashboard map of predicted changes in neighborhood WOZ values in the Netherlands

F Additional results

Figure S9 shows the importance of the neighborhood predictors retained by recursive feature elimination.

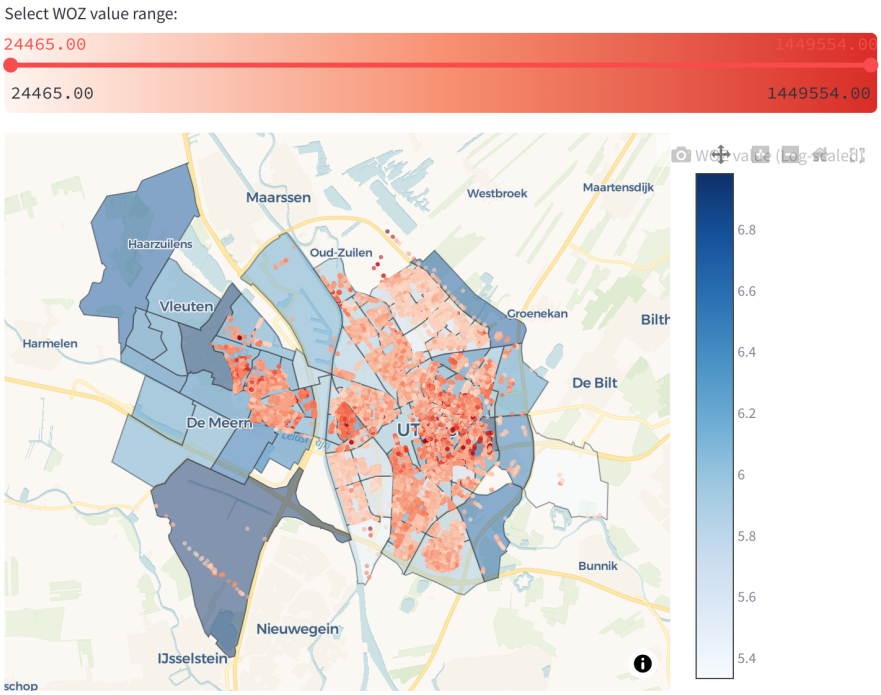

Figure S8. Dashboard map of predicted WOZ values of individual properties and their neighborhoods in Utrecht

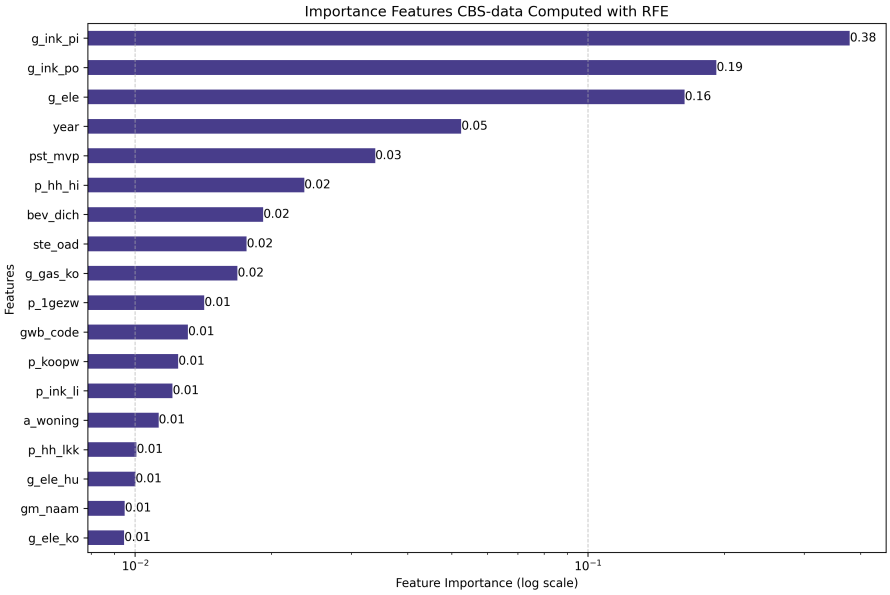

Figure S9. Importance of the neighborhood predictors retained by recursive feature elimination with a random forest regressor (logarithmic scale). Variable names are defined in Table S1; for example, `g_ink_pi` is average income per resident, `g_ele` average electricity consumption and `bev_dich` population density

References

Centraal Bureau voor de Statistiek. (2025b). *Kerncijfers wijken en buurten 2004–2024* [Key figures for districts and neighborhoods 2004–2024] [dataset]. Retrieved from https://www.cbs.nl/nl-nl/reeksen/publicatie/kerncijfers-wijken-en-buurten. Accessed February 21, 2025.

WOZ-waardeloket. (2025). *WOZ-waardeloket* [WOZ value portal] [dataset]. Retrieved from https://www.wozwaardeloket.nl/. Accessed February 21, 2025.