Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Interpreting feature importance under spatial heterogeneity: Evidence from urban development stages and regional migration

**dongwoo kim**

dwkim0302@gmail.com

Baekseok University https://orcid.org/0000-0003-2219-083X

**Research Article**

**Keywords:** internal migration, urban development stages, feature importance, SHAP, explainable machine learning, spatial heterogeneity, South Korea

**Posted Date:** July 31st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10520341/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Interpreting feature importance under spatial heterogeneity: Evidence from urban development stages and regional migration


## Abstract

Feature importance from explainable machine learning increasingly guides regional analysis and policy advice. Existing cautions establish what importance is not: not a causal effect, not a policy lever. But they were developed in aspatial settings, and they leave a prior question open. Is an importance ranking stable across the spatial regimes that regional processes exhibit? We test that question on internal migration across 229 South Korean municipalities (2017–2024). We partition the attributions of a single fitted gradient-boosting model by urban development stage and evaluate the resulting rank differences against a permutation null obtained by reassigning stage labels across municipalities. Twelve of twenty predictors are stage dependent. Stage dependency, however, is largely unrelated to global importance. The four leading predictors move by at most four rank positions, whereas the largest re-orderings occur in the middle of the ranking. Magnitude also fails to disclose direction. Network prominence is the most informative predictor in sparsely populated municipalities, yet its contribution there is negative: the model relies on the absence of network position. Childcare provision, the leading predictor overall, reverses the sign of its contribution across the settlement hierarchy. Cross-model comparison, cluster bootstrapping, and spatial transfer tests indicate that these patterns are not artefacts of model choice. Feature importance in spatially heterogeneous systems is therefore best read as evidence about the spatial organization of predictive information. Its direction and its spatial uniformity must be established separately, because the ranking encodes neither.



Keywords: internal migration; urban development stages; feature importance; SHAP; explainable machine learning; spatial heterogeneity; South Korea


## 1. Introduction

Interregional migration redistributes population across space and reshapes the demographic, economic, and social trajectories of regions. Unlike natural change, it responds quickly to shifts in economic opportunity, housing, accessibility, and public services. Across OECD countries it is the mechanism through which concentration at the core and depopulation, ageing, and service withdrawal elsewhere have advanced together (OECD,

2023). Migration shapes not only the fortunes of individual municipalities but the spatial organization of national development.

The stakes are acute in South Korea. By the end of 2023, 50.7 per cent of the national population lived in the capital region of Seoul, Incheon, and Gyeonggi, an area covering little more than a tenth of the national territory (Statistics Korea, 2024). Many non-capital municipalities face decline, rapid ageing, and doubts about long-term viability. Explaining why some municipalities consistently attract migrants while others lose population has become a central objective of regional science, human geography, and urban policy research.

A substantial literature has pursued that objective through place-based determinants. Demographic structure, employment, income, housing, accessibility, and public services all shape migration decisions (Greenwood, 1997), with no single factor dominating. Their relative weight also shifts over time: between the 2000 and 2010 Chinese censuses, the balance between origin-side push and destination-side pull reversed (Cao et al., 2018). That determinants are not fixed in time invites the question this study takes up. Are they fixed in space?

Much of this literature, assumes that determinants operate consistently across regions, implicitly treating estimated relationships as broadly transferable (e.g., Pu et al., 2019; Yu et al., 2019). That assumption sits uneasily with a long geographical tradition holding that relationships between regional variables are rarely stable over space (Fotheringham & Li, 2023): not only outcomes but the mechanisms producing them may differ geographically, so the contribution of any factor may depend on the context in which it operates rather than representing a universal determinant.

The methodological frontier has meanwhile shifted toward nonlinear machine learning, which accommodates interactions without restrictive functional assumptions and now outperforms classical specifications in flow generation (Simini et al., 2021). This transition changes how influence is represented: coefficients give way to feature importance, computed most prominently with SHAP (Lundberg & Lee, 2017), which has become a central basis for identifying migration determinants in predictive modelling.

These two developments expose the gap this study addresses. If relationships are context-dependent, feature importance may likewise reflect the regional context in which a model is learned rather than universally influential determinants. Yet rankings are routinely read as evidence that a predictor is inherently important, and existing cautions have been developed in largely aspatial settings: predictive importance cannot be equated with causal relevance (Janzing et al., 2020), and Shapley-based measures carry limitations of their own (Kumar et

al., 2020). Whether differences in feature importance represent stable characteristics of migration processes or context-dependent expressions of regional heterogeneity therefore remains unresolved, and on that answer depends whether importance can be read as evidence situated within the geographical context from which it is derived.

This study addresses these gaps through three sequential research questions, examined on a balanced panel of 229 South Korean municipalities observed from 2017 to 2024 (1,832 municipality-year observations) with the net migration rate as the outcome, a setting whose extreme spatial polarization makes it a demanding test case rather than a merely local application. How are migration outcomes spatially organized (RQ1)? Do the predictive contributions of their determinants vary across urban development stages (RQ2)? And what does such variation imply for the interpretation of feature importance (RQ3)?

The contribution is threefold, and none of it consists in proposing a new determinant of migration. First, the study supplies a test where the literature has offered description: the null hypothesis that a predictor's importance is invariant across spatial regimes is evaluated against a permutation distribution obtained by reassigning regime labels across municipalities, so that an observed re-ordering can be distinguished from what random assignment would produce. Second, it establishes that global importance and spatial contingency are orthogonal properties, the predictors at the top of the ranking prove the most uniform across urban contexts, while the largest re-orderings occur among predictors of intermediate importance, so that a ranking does not disclose which of its own entries are context dependent. Third, it shows that magnitude does not disclose direction, and shows this for the two predictors that matter most to the literatures at issue: network prominence, which is most informative precisely where municipalities lack it, and childcare provision, whose contribution reverses sign across the settlement hierarchy. Together these findings sharpen a caution stated in aspatial terms by explainable-AI methodology (Janzing et al., 2020; Kumar et al., 2020) into a reading rule for spatially heterogeneous systems, and they extend the geographical tradition of spatially varying relationships (Fotheringham & Li, 2023), which describes how coefficients vary across space but is silent on whether importance rankings do.

## 2. Literature review

This review is an argument rather than a survey. It moves from internal migration as the object of explanation (2.1), through the network turn that supplies this study's distinctive predictors (2.2) and the heterogeneity literature that establishes context dependence as the

rule in regional processes (2.3), to the interpretation of predictive importance in explainable machine learning, where that rule has yet to be absorbed (2.4).

## 2.1 Internal migration and regional population change

Internal migration has long been recognized as a key mechanism of regional population redistribution (Greenwood, 1997). In low-fertility societies, where natural increase has weakened, its demographic significance has become even greater, and migration intensities vary systematically across levels of national development (Bell et al., 2015). Explanations of who moves and where have emphasized economic opportunity, amenities, and life-course considerations (Massey et al., 1993). Agglomeration economics explains why migration reinforces concentration once it begins (Fujita et al., 1999); and stage frameworks of population redistribution (urbanization, suburbanization, counterurbanization) relate migration patterns to successive phases of urban development (Champion, 2001). South Korea offers an acute case: decades of migration toward the capital region have produced one of the most concentrated settlement systems among developed countries (OECD, 2023), and comparable amenity- and opportunity-driven concentration has been documented in China (Yu et al., 2019).

Concentration's counterpart is decline, now a first-order policy concern across OECD countries (OECD, 2023), and recent scholarship cautions against treating it as a single phenomenon: shrinking regions differ in the processes generating population loss and in the responses they require (Martinez-Fernandez et al., 2012; Mallach et al., 2017). Methodologically, most determinant studies share one template, gravity, regression, spatial-econometric, or dynamic-panel specifications (e.g., Pu et al., 2019) whose coefficients are assumed constant across the study area. Whether that assumption holds in a system as polarized as Korea's is the first question this study confronts.

## 2.2 Migration networks and network centrality

A second strand shifts attention from regional attributes to relations: migration is treated not as independent flows but as an interconnected system (Mabogunje, 1970), in which migration decisions and trajectories are embedded in social networks rather than made in isolation (Bilecen & Lubbers, 2021). Network analysis operationalizes this view at the regional scale (regions as nodes, flows as weighted edges) revealing hierarchy, heterogeneous connectivity, and community structure that flow-by-flow models cannot capture (Danchev & Porter, 2018, 2021), with applications to internal migration in settings such as Croatia (Pitoski et al., 2021), Turkey (Gürsoy & Badur, 2022), and China, where the architecture of the intercity network proved stable over a decade even as the distribution of flows within it shifted markedly (Wu & Liu, 2022).

Centrality indices summarize a region's position: prominence measures such as weighted PageRank capture recursive importance as a destination within the system, whereas closeness-type measures capture accessibility to all other regions (Freeman, 1978; Page et al., 1999). Central positions are typically read as structural advantages, and this reading is typically assumed to apply across regions (e.g., Danchev & Porter, 2021; Pitoski et al., 2021). What existing work has not asked is whether the importance of network position is itself conditional: whether centrality informs migration outcomes equally in peripheral, intermediate, and mature urban contexts. Geography has long answered that such conditionality is plausible, relationships between regional variables are rarely stable over space (Fotheringham et al., 2002), the third strand of this review.

## 2.3 Spatial heterogeneity and context-dependent relationships

That relationships between regional variables vary over space is one of geography's most robust findings: spatial context conditions behaviour, and modelling that context is a research programme in its own right (Fotheringham & Li, 2023). The canonical expression is the family of spatially varying coefficient models, geographically weighted regression and its multiscale extension (Fotheringham et al., 2002, 2017), and applications to migration confirm that the determinants of intercity flows differ systematically across space (Li et al., 2023).

This tradition, frames context dependency almost exclusively in regression coefficients. Migration modelling is meanwhile shifting toward nonlinear machine learning, where relationships are summarized not by coefficients but by feature importance, and recent applications have begun to probe spatially non-stationary effects within explainable machine learning (Ming et al., 2024). Whether the geographic lesson carries over to those measures, and what it implies for their interpretation, is the fourth strand of this review.

## 2.4 Interpreting feature importance in explainable machine learning

Machine-learning models capture the nonlinearities and interactions that pervade migration processes and now outperform classical specifications in flow generation (Simini et al., 2021). Because such models are not directly interpretable, SHAP, which decomposes each prediction into additive feature contributions (Lundberg & Lee, 2017), has become the standard bridge between predictive accuracy and substantive claims (Molnar, 2022; Li, 2025), and importance rankings derived in this way increasingly inform regional analysis and policy advice (e.g., Shams Amiri et al., 2021).

The typical workflow, ends where interpretation should begin: a model is fitted, a global ranking is reported, and top-ranked variables are narrated as key drivers or policy priorities. Explainable-AI methodology has warned against precisely this move: predictive attribution

cannot be equated with causal relevance (Janzing et al., 2020; Ma & Tourani, 2020), and Shapley-based measures carry mathematical and interpretive limitations of their own (Kumar et al., 2020). Yet these cautions specify what importance is not, and they have been developed in a largely aspatial setting. What importance positively means for a regional process cannot be settled in the abstract, because the relationships being modelled are, as the previous section established, rarely stable over space.

Spatial explainable AI has begun to respond along two lines. The first measures spatial effects more faithfully: SHAP-based local interpretation can recover effects comparable to those of spatially varying coefficient models (Li, 2022), GeoShapley treats location as a player in the prediction game (Li, 2024), and a parallel line maps local SHAP values across space, arguing explicitly that the static view offered by an importance ranking should give way to the geography of contributions (Wang et al., 2024). The second line asks how far such explanations can be trusted: bootstrap confidence intervals have been attached to GeoShapley values (Li, 2024), and conformal prediction has been used to attach calibrated uncertainty to geospatial explanations at the level of individual predictions (Lou et al., 2025).

Both responses take the ranking's limitations as a reason to look past it, either replacing it with a map of local contributions or attaching uncertainty to those contributions one place at a time. Yet the ranking is what analysts and policymakers actually read, and neither response asks the question that the acknowledgement of context dependence makes unavoidable: whether the ordering of predictors is itself invariant across spatial regimes, and how it should be read if it is not. The literature thus documents heterogeneity and measures importance, but supplies no rule connecting the two.

## 2.5 Research gap and research questions

These strands are successive layers of one problem. Migration research shows that outcomes are spatially organized. Network research supplies positional predictors. The heterogeneity literature establishes that the relationships linking predictors to outcomes shift with spatial context; and explainable machine learning supplies the importance measures through which such relationships are now read. The unresolved question is not whether feature importance can be computed, but how it should be interpreted once spatial heterogeneity is acknowledged.

Each strand leaves a gap that the next inherits. Network studies assume that positional advantage operates uniformly. Heterogeneity research stops at spatially varying coefficients without engaging importance measures; and explainable machine learning produces importance rankings without a theory of their meaning. Existing research has yet to connect these layers explicitly by asking what feature importance means when the process being

predicted is spatially heterogeneous, despite the growing reliance of regional analysis and policy advice on importance rankings.

The present study addresses this intersection through three sequential research questions. RQ1 asks how migration outcomes are spatially organized. RQ2 asks whether the predictive contributions of migration determinants, network position foremost among them, vary across urban development stages. RQ3 asks what such variation implies for the interpretation of feature importance.

# 3. Methods

## 3.1 Research design

The objective of this study is not merely to identify important predictors of municipal migration but to examine how spatial heterogeneity affects the interpretation of feature importance in explainable machine learning. The empirical design therefore follows a single analytical sequence in which each step motivates the next, accompanied by robustness analyses that assess the stability of the resulting interpretation (Figure 1).

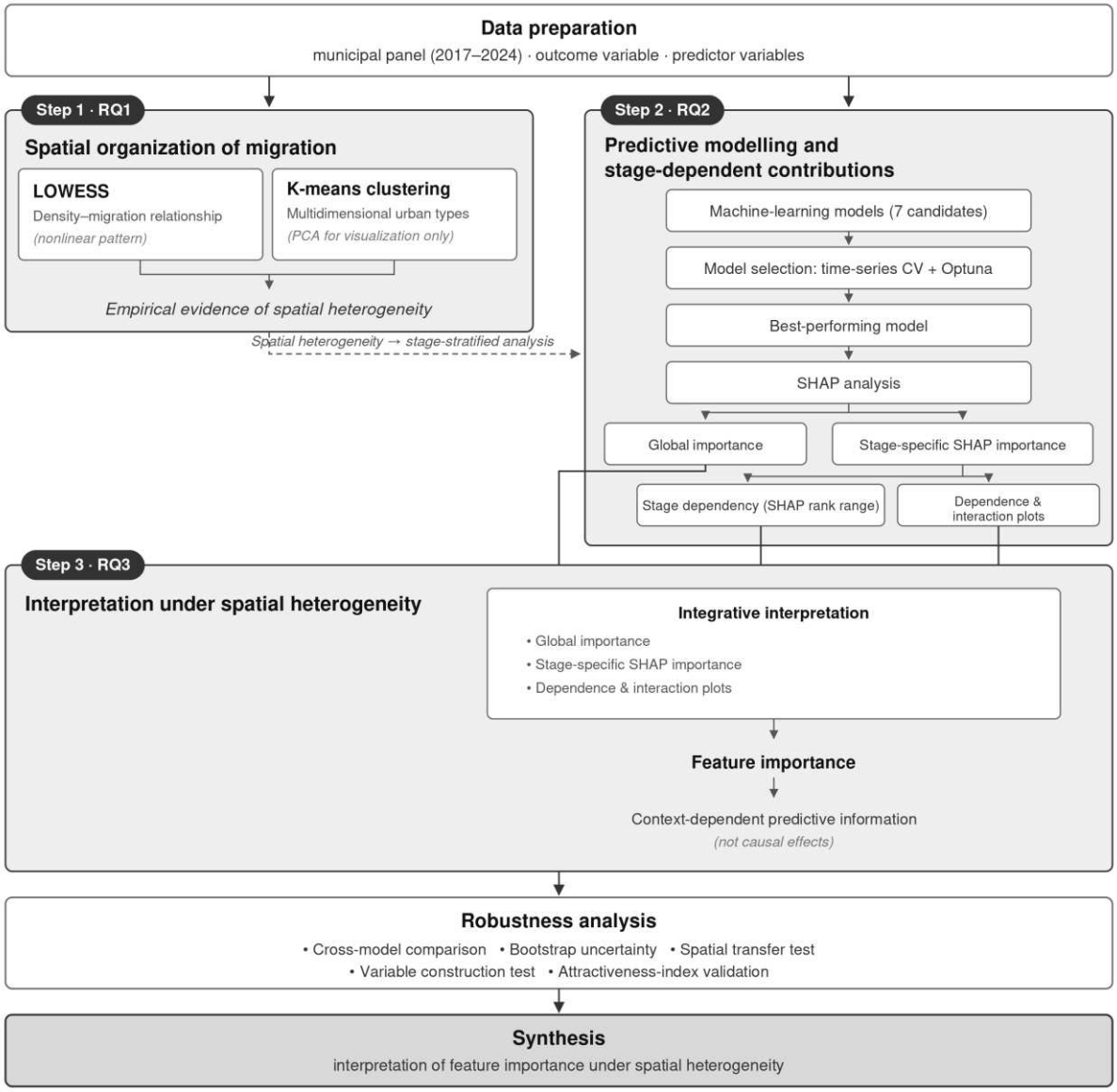



**Figure 1**

**Overall analytical framework and workflow. Prepared data feed Step 1 and Step 2 in parallel. Step 1 combines LOWESS smoothing and k-means clustering to establish empirical evidence of the spatial organization of migration outcomes (RQ1). Step 2 develops the predictive model and decomposes its SHAP contributions across density-based development stages (RQ2), and Step 3 examines how the resulting feature importance should be interpreted (RQ3). Robustness analyses assess the stability of the inferred predictive structure before the final synthesis.**

The design proceeds in three stages, one per research question. The first combines LOWESS smoothing with k-means clustering to establish whether migration outcomes are geographically differentiated. If they are, the determinants contributing to them are unlikely to operate uniformly. The second partitions the attributions of a single fitted predictive model by development stage, using network position as the representative test case because migration-network research has treated centrality as a generally beneficial attribute

(Section 2.2). The third asks what the resulting pattern implies for reading an importance ranking.

## 3.2 Data and variables

The empirical analysis uses a balanced municipal panel covering 229 municipalities in South Korea over the period 2017–2024, yielding 1,832 municipality-year observations. Municipalities constitute the primary unit of analysis because migration policies and demographic planning are largely implemented at the local-government level. The panel is structurally balanced in its municipality-year dimensions. Item-level missingness in individual predictors is reported in Appendix Tables A1 and A2. Missing predictor values were imputed with year-specific variable medians for model training, and observations with missing values were excluded from the stage-stratified analyses.

Missing values are concentrated in a few predictors, the employment rate excepted (17.4% of municipality-years; Appendix Table A2). Model selection and estimation use the full panel with year-specific median imputation, so that all municipalities contribute to the fitted model and the hold-out evaluation. Interpretation is restricted to complete records: all SHAP quantities are evaluated on the 1,480 municipality-year observations with valid values on the outcome and all twenty predictors, so that no feature importance is attributed to an imputed value, and the same sample underlies the global and the stage-stratified analyses, making their ranks directly comparable. Sensitivity to both decisions is reported in Section 4.5.

The dependent variable is the annual municipal net migration rate (‰), calculated as net migrants relative to the resident population. Net migration is adopted because it summarizes the balance of inflows and outflows and thus provides an overall indicator of a municipality's relative migration attractiveness, the focus of this study, rather than migration turnover. Net rates are widely used in studies of regional population redistribution (e.g., Fan, 2005).

We compile twenty explanatory variables from official statistics and grouped into five domains, migration-network structure, demographic characteristics, economic conditions, infrastructure and housing, and public-service provision, consistent with the determinants reviewed in Sections 2.1–2.2. Table 1 reports the variables, their definitions, and their sources. Appendix Table A1 reports descriptive statistics.

**Table 1**
**Dependent and explanatory variables included in the predictive model.**

| Domain | Variable | Description |
|---|---|---|
| Dependent variable | Net migration rate (‰) | Annual net migrants (in-migrants − out-migrants) per 1,000 municipal residents |
| Network structure | PageRank (t−1) | PageRank centrality in the national inter-municipal migration-flow network, entered with a one-year lag to prevent target leakage |
| | Closeness centrality | Closeness centrality of the municipality within the migration network |
| Demographic | Population (ln) | Natural logarithm of the resident population |
| | Population density | Residents per km² |
| | Ageing ratio | Share of residents aged 65 and over (%) |
| | Youth ratio | Share of residents aged 20–39 (%) |
| | Fertility rate | Total fertility rate |
| | Extinction-risk index | Ratio of women aged 20–39 to residents aged 65 and over |
| Economic | Employment rate | Municipal employment rate (%) |
| | Fiscal independence | Share of locally raised revenue in the municipal budget (%) |
| | Business establishments | Number of registered business establishments |
| Infrastructure and housing | Housing age | Average age of the housing stock (years) |
| | Sewerage supply | Sewerage coverage rate (%) |
| | Distance from Seoul | Distance between the municipality and Seoul (km) |
| Public services | Childcare facilities | Childcare facilities per 1,000 residents |
| | Doctors | Physicians per 1,000 residents |
| | Hospital beds | Hospital beds per 1,000 residents |
| | Senior-care facilities | Senior-care facilities per 1,000 residents |
| | Private academies | Private educational academies per 1,000 residents |
| | Cultural facilities | Number of cultural facilities |

*Note: PageRank is lagged by one year (t−1). Population density is additionally used to define urban development stages (Section 3.3.2). Business establishments and cultural facilities are raw counts. Other service variables are rates per 1,000 residents. All variables are compiled from official Korean administrative statistics (KOSIS).*

The network domain contains two measures of municipal position: PageRank and closeness centrality. Because PageRank is computed from the same annual migration flows that define the outcome, contemporaneous values would introduce target leakage. It therefore enters with a one-year lag (t−1).

We compute both measures with NetworkX on annual directed, weighted migration networks in which the 229 municipalities are nodes, the edge weight is the number of migrants moving from i to j, and within-municipality moves are excluded. PageRank measures the stationary probability that a municipality is reached through the directed network, weighting inflows from influential origins more heavily (Page et al., 1999):$w_{ij}$

$$PR_i = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{w_{ji}}{\sum_k w_{jk}} PR_j \quad (1)$$

where is the number of municipalities, denotes the set of municipalities sending migrants to municipality i, is the migration flow from j to i, and is the conventional damping factor. Closeness centrality measures how easily a municipality is reached from all other municipalities within the migration network:$NM(i) w_{ji} d$=0.85

$$CC_i = \frac{N-1}{\sum_{j \neq i} \delta(j,i)} \quad (2)$$

where is the shortest-path distance to municipality i computed with edge length, so that municipality pairs connected by larger migration flows are treated as closer. The resulting values are therefore not bounded above by one.$\delta(j,i) \frac{1}{w_{ij}}$

The remaining four domains follow Table 1: demographic composition and settlement characteristics. Labour-market conditions and municipal fiscal capacity. Built-environment attributes, including housing age, sewerage supply, and distance from Seoul; and local public-service provision.

Population density plays two distinct roles in the analysis. First, it is included as one of the candidate predictors in the machine-learning model. Second, independently of its predictive role, it serves as the operational criterion for stratifying municipalities into broad urban development stages when evaluating spatial heterogeneity.

Population density has frequently been used as an indicator of settlement intensity and urban maturity (Champion, 2001; Dijkstra et al., 2021). Urban development is multidimensional and cannot be reduced to one indicator, but density provides a transparent

and reproducible basis for stratification. The operational procedure is described in Section 3.3.2.

## 3.3 Analytical framework

### 3.3.1 Step 1: Identifying the spatial organization of migration (RQ1)

The first analytical stage examines whether municipal migration exhibits a systematic spatial structure independently of any predictive model. Establishing such structure is important because context-dependent predictor contributions are meaningful only if migration outcomes themselves are spatially differentiated.

To characterize the relationship between urbanization and migration outcomes we fit a locally weighted scatterplot smoothing regression (LOWESS; Cleveland, 1979) between municipal population density and the net migration rate, using municipality-level means over 2017–2024 with density on a logarithmic scale (span = 0.5; sensitivity to the smoothing parameter is reported in Appendix Figure A1). LOWESS imposes no functional form. We use it as a descriptive diagnostic of whether the density–migration relationship is monotonic.

Whereas LOWESS evaluates whether migration follows a nonlinear density gradient, k-means clustering asks whether municipalities form distinct multidimensional types that a single gradient would obscure. We use nine standardized indicators spanning demographic, economic, infrastructural, and service conditions, excluding the net migration rate so that the typology can serve as an external criterion. The number of clusters is selected from silhouette, Calinski–Harabasz, and elbow diagnostics over k = 2–8 (Appendix Table A3, Figure A2), with the seed fixed for reproducibility. Principal component analysis is used only to visualize the solution and does not influence it.

This stage is descriptive rather than predictive: together the two analyses establish whether migration outcomes are spatially organized, which is the premise on which the stage-stratified analysis of predictor contributions rests.

### 3.3.2 Step 2: Predictive modelling and evaluation of stage-dependent predictor contributions (RQ2)

The second analytical stage investigates whether influential migration determinants exhibit context-dependent predictive contributions across urban development stages.

*3.3.2.1 Predictive model development and selection*

Because migration determinants are expected to exhibit substantial nonlinearity, complex interactions, and multicollinearity, several machine-learning algorithms were evaluated under a common validation framework to identify the best-performing predictive model. Seven candidate algorithms were considered: conventional linear regression, a single

decision tree, random forests, gradient boosting, XGBoost, LightGBM, and CatBoost. To ensure a fair comparison, all candidate models were evaluated using expanding-window time-series cross-validation (Appendix Table A4), while observations from 2023–2024 were reserved as an independent hold-out set for final evaluation. This design preserves the temporal ordering of observations and prevents look-ahead bias: the first fold trains on 2017–2019 and validates on 2020, with the training window expanding by one year until the final fold, which validates on 2022.

We optimize hyperparameters for all candidate algorithms with the Optuna framework (Akiba et al., 2019) with Bayesian optimization (TPE sampler with a fixed seed; 80 trials per algorithm) over predefined search spaces. Candidate models were ranked primarily by the mean validation $R^2$ across the expanding-window folds. When the difference in validation performance between competing models was negligible, the independent hold-out set was used as the tie-breaking criterion. The selected model was subsequently used for all explainability analyses. The search spaces and tuned hyperparameter values for each algorithm are reported in Appendix Table A5. Random seeds were fixed for all stochastic procedures, including clustering, hyperparameter optimization, model fitting, and bootstrap resampling, to ensure reproducibility.

We estimate bootstrap confidence intervals (2,000 replications) for hold-out performance, including pairwise differences in hold-out $R^2$ relative to the selected model, using both observation-level and province-block resampling, the latter preserving within-province spatial dependence. These intervals quantify uncertainty in hold-out predictive performance only and do not incorporate variability from model estimation or hyperparameter optimization (Appendix Table A6).

*3.3.2.2 SHAP-based interpretation*

After model selection we interpret predictor contributions with SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017), which attributes each prediction to additive feature contributions based on Shapley values. Throughout, feature importance refers to the mean absolute SHAP value aggregated across observations, whereas SHAP contributions denote local attributions for individual predictions. All SHAP quantities reported below come from a single fitted model, the selected algorithm trained on the 2017–2022 estimation window, and are evaluated on the complete-case observations defined in Section 3.2. Stage-specific importance is therefore a partition of one attribution basis, not a comparison across separately estimated models. Both quantities describe the predictive behaviour of the fitted model. We interpret them as measures of predictive information, not as estimates of causal effects.

SHAP dependence and interaction plots were additionally used to visualize how local contributions vary across the population-density gradient. These are presented as descriptive evidence rather than statistical tests (Section 4.3).

*3.3.2.3 Development-stage analysis*

We operationalize urban development stages as municipality-level population-density tertiles, classifying municipalities into three equal-sized groups labelled low-density, intermediate-density, and high-density tiers. The labels describe relative density and mark positions along the urban development hierarchy, not contemporary growth or decline. Because the stages are defined independently of the predictive model, differences in SHAP importance across stages reflect differences in municipal context rather than artefacts of stage construction. We obtain within-stage importance by partitioning the SHAP matrix of the single fitted model by tier, averaging absolute contributions within each partition, and ranking predictors within each stage. Stage dependency is summarized by the rank range: the difference between a predictor's maximum and minimum rank across the three tiers. Two inferential procedures accompany the comparison. First, we quantify uncertainty in the within-stage ranks with a cluster bootstrap. Municipalities, rather than municipality-years, are resampled with replacement (2,000 replications), so that the panel structure is preserved, and tier-specific ranks are recomputed at each replication to obtain percentile 95% confidence intervals. Resampling-based intervals for Shapley quantities are established practice (Li, 2024; Lou et al., 2025). What they cannot answer is whether an observed difference between regimes exceeds what random assignment of regimes would produce. The second procedure supplies that test. We evaluate the null hypothesis that a predictor's importance is independent of development stage by permutation, reassigning stage labels at random across municipalities (2,000 permutations) and recomputing the rank range, which yields a predictor-specific null distribution against which the observed rank range is assessed. Because the test is applied to each of the twenty predictors, we convert p-values to Benjamini–Hochberg false-discovery rates and report significance at $q < 0.05$. Both procedures hold the fitted model fixed, so the inference concerns the tier composition of the model's attributions. Variability arising from model estimation is addressed separately in Section 3.4.

### 3.3.3 Step 3: Interpreting feature importance under spatial heterogeneity (RQ3)

Once the presence of stage-dependent contributions among influential predictors has been assessed, the final analytical step examines how global feature importance should be interpreted under spatial heterogeneity. Rather than treating globally important predictors as universally relevant, this step evaluates whether their predictive importance is consistent across urban development stages or concentrated within particular urban contexts.

The interpretive framework integrates three sources of evidence from Step 2: global feature importance, stage-specific importance rankings, and SHAP dependence and interaction plots. Together, these analyses distinguish predictors exhibiting broadly consistent predictive contributions from those whose importance varies substantially across development stages.

Throughout this study we interpret neither feature importance nor SHAP contributions as estimates of causal effects or policy effectiveness. We treat them as measures of predictive information that characterize the predictive structure the model has learned under spatial heterogeneity. The final step evaluates whether globally important predictors can be read as universally applicable, and what should replace that reading when they cannot.

## 3.4 Robustness analyses

Six robustness analyses evaluate whether the principal findings depend on model specification, attribution method, spatial partitioning, variable construction, or the treatment of missing data. We compare importance rankings across five attribution approaches. We quantify uncertainty in the global ranking by bootstrap resampling with model refitting (50 replications), which complements the fixed-model cluster bootstrap of Section 3.3.2.3 by propagating estimation variability. We assess spatial generalizability by leave-one-province-out block cross-validation and by transferring a capital-region model to other provinces. We replace the per-capita service indicators with absolute facility counts. We estimate a specification in which a separate model is trained within each development stage, testing whether the reordering also emerges when the model itself may differ across stages. Finally, we vary the imputation and complete-case decisions of Section 3.2. Results are reported in Section 4.5.

All analyses were run in Python 3.10 using CatBoost 1.2 for the selected model, XGBoost 3.2, LightGBM 4.6 and scikit-learn 1.7 for the competing algorithms, NumPy 2.2, pandas 2.3 and SciPy 1.15 for numerical work, statsmodels for the LOWESS smoother, NetworkX for the centrality measures, and Optuna for hyperparameter search. Random seeds were fixed for clustering, hyperparameter optimization, model fitting, bootstrap resampling and permutation. Exact pinned versions, the municipal panel, and a single script that regenerates every table, figure and in-text value reported below are available in the replication package (see Data availability).

Taken together, these analyses evaluate the stability of the inferred predictive structure rather than the predictive performance of any single model.

# 4. Results

## 4.1 Model performance

Before interpreting feature importance, the predictive performance of the candidate models was evaluated to identify an appropriate model for the subsequent SHAP analysis. Because SHAP explains the behaviour of a fitted predictive model, interpretation is meaningful only after adequate predictive performance has been established.

We evaluate seven candidate models under the expanding-window cross-validation framework described in Section 3.3.2, followed by independent evaluation on the 2023–2024 hold-out set (Table 2).

**Table 2**
**Predictive performance of candidate models under expanding-window cross-validation and 2023–2024 hold-out evaluation.**

| Model | Validation $R^2$ | Test $R^2$ | 95% CI (block bootstrap) | Test RMSE | Test MAE | $\Delta R^2$ vs CatBoost (95% CI) |
|---|---|---|---|---|---|---|
| Linear regression (ridge) | −0.069 | −0.116 | (−0.564, 0.098) | 15.86 | 10.71 | 0.443 (0.318, 0.572) |
| Decision tree | 0.094 | −0.104 | (−0.539, 0.029) | 15.77 | 10.17 | 0.438 (0.277, 0.629) |
| Random forest | 0.341 | 0.303 | (0.137, 0.384) | 12.53 | 7.89 | 0.018 (−0.044, 0.077) |
| Gradient boosting | 0.336 | 0.274 | (0.051, 0.354) | 12.79 | 8.33 | 0.054 (−0.008, 0.130) |
| XGBoost | 0.350 | 0.298 | (0.125, 0.412) | 12.58 | 7.78 | 0.025 (−0.053, 0.104) |
| LightGBM | 0.323 | 0.252 | (0.005, 0.356) | 12.98 | 8.57 | 0.073 (0.015, 0.130) |
| CatBoost | 0.353 | 0.324 | (0.128, 0.392) | 12.34 | 8.17 | — |

*Note: Validation $R^2$ is the mean coefficient of determination across the three expanding-window cross-validation folds (validation years 2020–2022; Appendix Table A4). The hold-out set comprises all municipality-year observations from 2023–2024 (n = 458) and was excluded from model fitting and hyperparameter optimization. Confidence intervals are percentile intervals from 2,000 province-block bootstrap replications of the fixed hold-out predictions, which redraw entire provinces with replacement and thus preserve within-province spatial dependence. $\Delta R^2$ is the paired difference in hold-out $R^2$ relative to CatBoost, with the same resampling scheme. Model selection and performance evaluation use the full panel with year-specific median imputation of missing predictor values (Section 3.2). SHAP interpretation is subsequently restricted to complete-case observations. The linear baseline is a ridge regression with a fixed penalty and was not tuned.*

Table 2 shows a substantial performance advantage of nonlinear ensemble models over the linear and single-tree baselines. Ridge regression has no out-of-sample predictive skill (test $R^2 = -0.116$), and a single decision tree fares no better (−0.104). The difference between each and the selected model is large and its bootstrap confidence interval excludes zero. All five ensembles, by contrast, attain positive hold-out $R^2$ between 0.252 and 0.324.

Among the ensembles, differences are small. CatBoost attains the highest mean validation $R^2$ (0.353), marginally ahead of XGBoost (0.350) and random forests (0.341). Because these margins are negligible, the tie-breaking rule specified in Section 3.3.2 was applied, and CatBoost also attains the highest hold-out $R^2$ (0.324, against 0.298 for XGBoost). Block-bootstrap confidence intervals for the paired difference in hold-out $R^2$ nonetheless include zero for random forests, gradient boosting, and XGBoost: on predictive performance alone the ensembles are statistically indistinguishable from one another.

The purpose of this comparison was not to identify the "true" migration model but to select an appropriate predictive model for SHAP-based interpretation, and the statistical equivalence of the ensembles makes the consequences of that choice the relevant question. They are limited: refitting the analysis with XGBoost reproduces the global importance ranking closely (Spearman $\rho = 0.93$ across the twenty predictors) and yields the same stage-dependent pattern for network prominence (PageRank ranks 1, 4, and 7 across the low-, intermediate-, and high-density tiers, against 1, 6, and 9 under CatBoost; Appendix Table A12). All subsequent analyses are therefore based on the fitted CatBoost model.

## 4.2 RQ1, Spatial organization of municipal migration

RQ1 asks whether municipal migration exhibits systematic spatial structure rather than a simple urban–rural gradient, a question about the organization of outcomes, addressed independently of predictor importance.

### 4.2.1 Density and migration

To examine the overall settlement structure, a LOWESS curve was fitted between population density and municipal net migration (Figure 2).

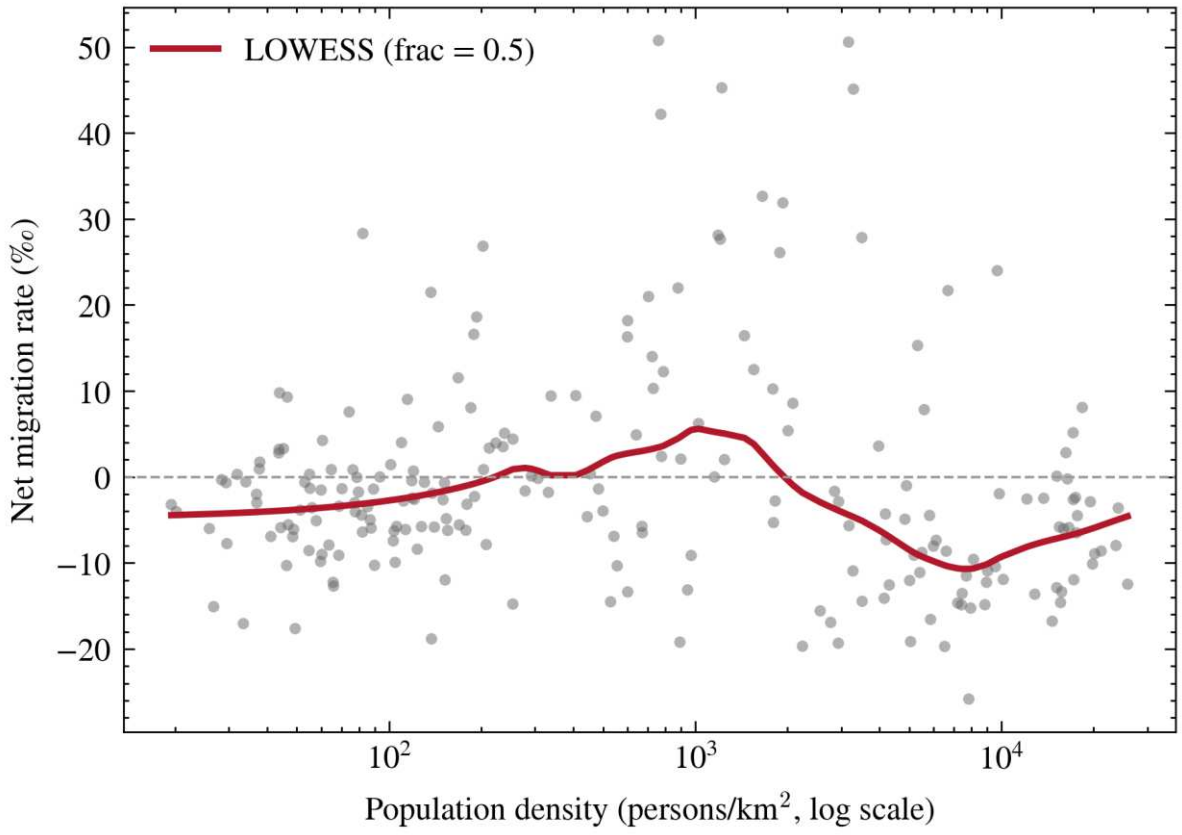



**Figure 2**
**LOWESS relationship between population density and net migration.**

*Note: Each point represents one municipality, based on municipality-level averages of population density and the net migration rate over 2017–2024 (n = 224 municipalities with complete records on the typology variables, the same sample as the k-means analysis; Appendix Table A2). The solid curve is a locally weighted regression (LOWESS, span = 0.5). Population density is displayed on a logarithmic scale, and the dashed horizontal line marks zero net migration. Sensitivity to alternative smoothing parameters is shown in Appendix Figure A1.*

Rather than increasing monotonically with urbanization, net migration follows a pronounced inverted-U pattern: gains concentrate at intermediate densities, while both sparsely populated municipalities and the densest municipalities lose population (Figure 2). Only the intermediate tier records positive average net migration (+5.29‰, against −2.58‰ and −5.33‰).

An obvious objection is that this pattern merely reproduces the geography of the capital region. Intermediate-density municipalities are the suburban ring of Seoul, so what looks like a density effect may be an accessibility effect. We test that objection in Table 3 by cross-classifying municipalities on density and on distance from the capital.

**Table 3**
**Mean municipal net migration rate (‰) by population-density tertile and distance from the capital, 2017–2024.**

| Density tertile | Near capital | Intermediate distance | Far from capital | All (mean; SD; n) |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Low density | −7.10 (5) | −1.58 (41) | −4.03 (29) | −2.58; 7.18; 76 |
| Intermediate density | +13.99 (24) | +0.55 (25) | +2.44 (26) | +5.29; 14.49; 75 |
| High density | −2.68 (46) | −8.79 (9) | −9.14 (21) | −5.33; 13.20; 75 |
| $\eta^2$ (explained variance) | Distance tertiles: 0.034 | | Density tertiles: 0.130 | |

*Note: Entries are means of municipality-level average net migration rates over 2017–2024, with the number of municipalities in parentheses. 226 municipalities with complete records on both stratifying variables are included. Density tertiles define the urban development stages used throughout the study (boundaries ≈ 135 and 1,961 persons/km$^2$; Section 3.3.2). Distance tertiles are computed from road distance to Seoul. $\eta^2$ is the share of between-municipality variance in net migration explained by each tertile classification.*

The inverted-U survives. Within every distance band (near the capital, at intermediate distance, and far from it) the intermediate-density tier is the only one with positive mean net migration, and the high-density tier is always the most negative. Proximity to the capital amplifies the peak substantially (+13.99‰ for intermediate-density municipalities near the capital, against +2.44‰ for those far from it) but it does not create it. The two classifications are also unequal in explanatory power: density tertiles account for 13.0% of the between-municipality variance in net migration, distance tertiles for 3.4%. Density, not accessibility, is the primary axis along which Korean municipal migration is organized; accessibility modulates it.

Figure 3 maps the spatial expression of these differences, anticipating the geographical analysis in Section 4.2.3.

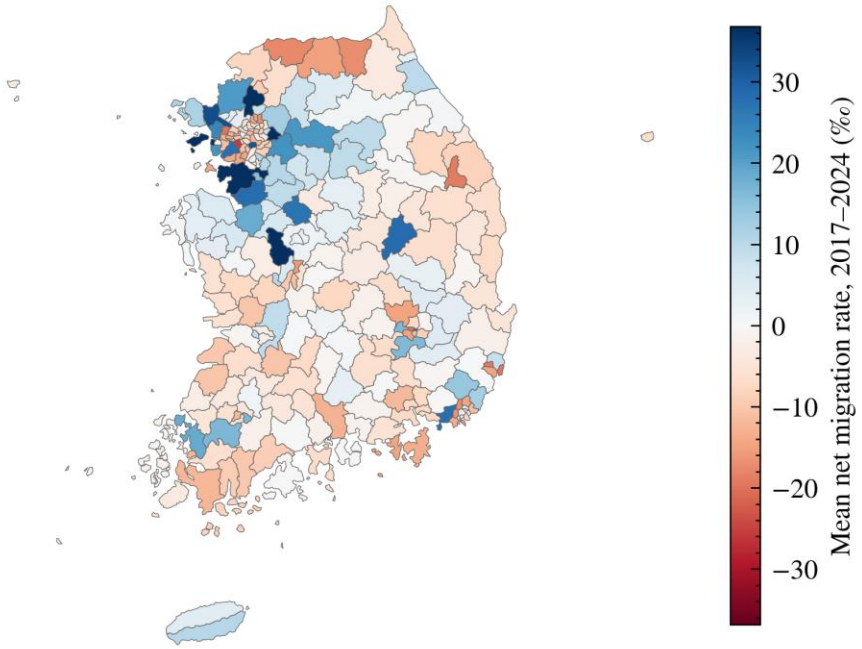


**Figure 3**
**Spatial distribution of the mean municipal net migration rate, 2017–2024 (‰).**

*Note: Net in-migration (blue) concentrates in suburban municipalities within commuting distance of Seoul and in planned administrative centres, whereas most rural municipalities and mature metropolitan cores record net losses (red). Values are municipality-level means over 2017–2024. Boundaries follow 2013 administrative divisions matched to 2024 units.*

### 4.2.2 Multidimensional municipal typology

K-means clustering of the nine standardized demographic, economic, infrastructure, and service indicators (Section 3.3.1) identified four distinct multidimensional municipal types beyond the population-density gradient.

Cluster-number diagnostics for k = 2–8 did not identify a single dominant solution: the silhouette coefficient and the Calinski–Harabasz index favoured very small numbers of clusters, whereas the elbow curve flattened between k = 3 and k = 5 (Appendix Table A3 and Figure A2). Within this range, we adopt the four-cluster solution because it distinguishes substantively different urban types while remaining parsimonious. The separation of the resulting clusters in the space of the clustering indicators is shown in Appendix Figure A8. Principal component analysis was used solely for visualization and did not influence the clustering.

**Table 4**
**Characteristics of the four municipal clusters (municipality-level means, 2017–2024).**

| Characteristic | Cluster 2: Capital-region growth core | Cluster 0: Intermediate | Cluster 3: Non-capital metropolitan | Cluster 1: Rural decline |
|---|---|---|---|---|
| Population density (per km²) | 9,030 | 389 | 4,322 | 219 |
| Housing age (years) | 17.3 | 28.5 | 22.7 | 44.9 |
| Ageing ratio (%) | 14.7 | 23.4 | 17.0 | 34.0 |
| Distance from Seoul (km) | 34.9 | 156.8 | 294.7 | 229.3 |
| Fiscal independence (%) | 33.4 | 16.6 | 22.7 | 9.4 |
| Net migration rate (‰) | +2.98 | −2.10 | −2.97 | −2.85 |
| Municipalities (n) | 65 | 56 | 40 | 63 |

*Note: Entries are cluster means of municipality-level averages over 2017–2024. Five of the nine clustering indicators (Section 3.3.1) are shown. The omitted indicators are strongly correlated with those displayed or vary little across clusters, and full profiles are available in the replication package. The net migration rate was not used in clustering and serves as an external criterion. Cluster labels are descriptive names assigned after estimation.*

Cluster 2 (Table 4) comprises municipalities in and adjacent to the Seoul Capital Area: the highest density, the youngest housing, the least ageing, the strongest fiscal capacity, and positive net migration, the principal growth regime of the settlement system. Cluster 1 is predominantly rural: low density, old housing, advanced ageing, persistent out-migration. Cluster 3 comprises non-capital metropolitan municipalities, substantially more urbanized

than Cluster 1 yet also losing population despite developed infrastructure. Cluster 0 occupies an intermediate position with only modest loss.

Two forms of decline therefore coexist, and they differ less in depth than in kind. Among non-capital municipalities, the low-density (rural) group and the high-density (metropolitan) group lose population at statistically indistinguishable average rates (−2.94‰ and −3.60‰; Mann–Whitney p = 0.25). What separates them is dispersion: the standard deviation of net migration is 7.2‰ among rural municipalities and 11.7‰ among non-capital metropolitan ones. Rural decline is pervasive (39 of 53 municipalities lose population, and few gain appreciably) whereas metropolitan decline is selective, combining sharply losing districts with a minority that grow. The two regimes call for different diagnoses even though their averages coincide.

The geographical expression of the typology is shown in Figure 4. The capital-region growth core (cluster 2) forms a contiguous block within and around the Seoul Capital Area, the non-capital metropolitan cluster (cluster 3) concentrates in the south-eastern urban belt, and the rural-decline cluster (cluster 1) covers most of the remaining periphery.

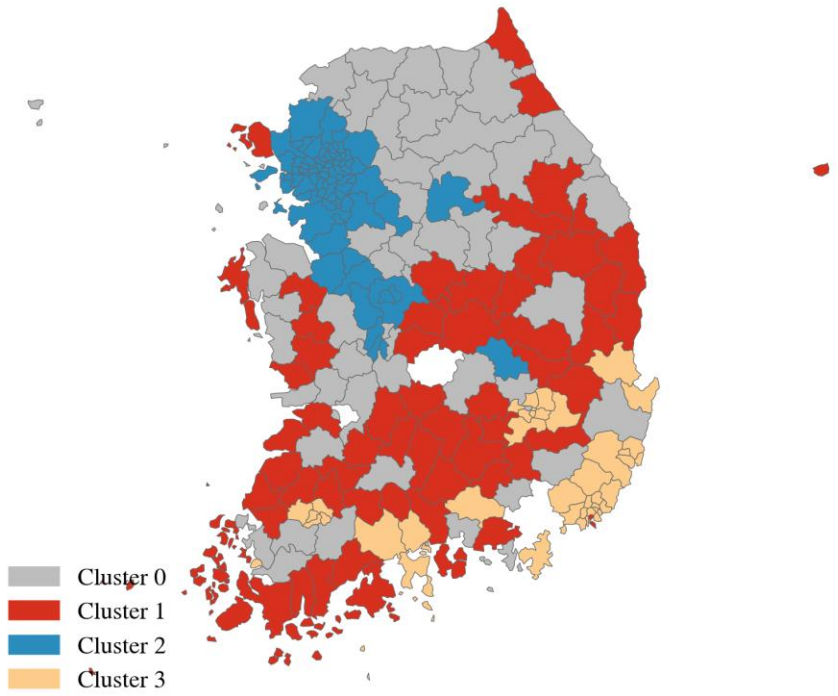



**Figure 4**
**Geographical distribution of the four municipal clusters.**

*Note: Cluster labels follow Table 4. The five municipalities shown in white (Yeongdong, Changwon, Jeonju, Jeju, and Seogwipo) were excluded from clustering because they lack complete records for one or more of the nine clustering indicators (e.g., fiscal independence for the two Jeju municipalities).*

### 4.2.3 Spatial polarization

The geography of gains and losses (Figures 3 and 4) shows that migration is spatially differentiated rather than randomly distributed, and it locates the growth not in the capital but around it.

Within the Seoul Capital Area, the core loses and the ring gains. The twenty-five districts of Seoul record a mean net migration rate of −7.17‰, and twenty-three of them lose population. The thirty-one municipalities of Gyeonggi record +9.20‰, with nineteen gaining. Incheon lies between (+3.69‰). The municipalities with the largest gains are suburban and peri-urban, such as Hanam (+50.6‰) and Hwaseong (+45.3‰), together with the planned administrative city of Sejong (+50.8‰). The largest losses within the capital region occur in dense Seoul districts such as Nowon (−16.8‰) and Dobong (−14.6‰) (Appendix Table A9).

Capital-region concentration is therefore no longer centripetal. It has become redistributive: population is moving outward from the mature core into the surrounding ring, not inward toward Seoul. High density is not in itself associated with migration gains, municipalities of comparable density record opposite outcomes depending on their position within the settlement system.

Collectively, these findings answer RQ1. Municipal migration is organized as a geographically differentiated system with three regularities. First, net migration is an inverted-U function of density that holds within every distance band from the capital, so the pattern is not an artefact of capital-region accessibility. Density explains four times as much of the between-municipality variance as distance does. Second, capital-region concentration is now internally redistributive: Seoul's districts lose while the Gyeonggi ring gains. Third, decline outside the capital region takes two forms of equal average depth but unequal dispersion, pervasive in rural municipalities, selective in non-capital metropolitan ones. This organization motivates the second research question: if migration outcomes are spatially differentiated in this way, the contribution of individual predictors may likewise vary across urban development stages rather than remaining constant throughout the settlement system.

## 4.3 RQ2, Predictor contributions vary across urban development stages

Given that migration outcomes are spatially organized (Section 4.2), RQ2 asks whether the predictive contribution of their determinants likewise differs across urban development stages.

The analysis proceeds in three steps. First, the global importance ranking is established from the fitted CatBoost model. Second, predictor importance is recalculated separately within each development stage defined by population-density tertiles. Finally, SHAP interaction plots are examined to visualize how local SHAP contributions vary across urban contexts.

Importantly, all results reported below concern predictive contribution within the fitted model rather than causal effects.

### 4.3.1 Global predictor importance

The global predictor ranking provides the baseline for the stage-stratified analysis (Figure 5; Table 5).

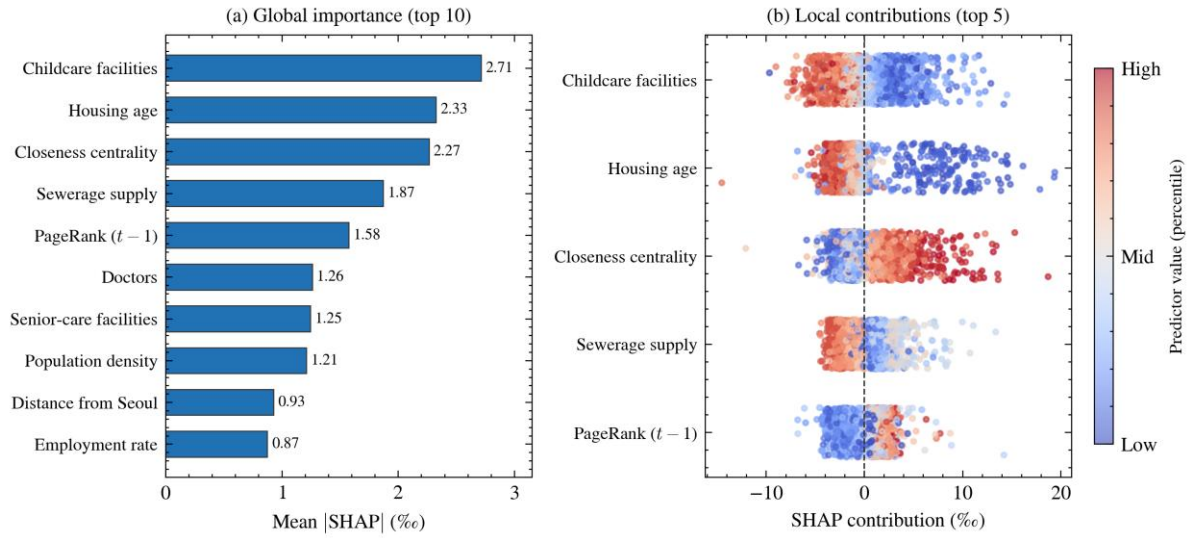



**Figure 5**

**Global SHAP feature importance derived from the fitted CatBoost model.**

*Note: Panel (a) ranks the ten leading predictors by mean absolute SHAP value (‰), computed from the single fitted CatBoost model and evaluated on the 1,480 complete-case municipality-year observations (Section 3.2). Values correspond to Table 5. Panel (b) shows the distribution of local SHAP contributions for the five leading predictors: each point is one municipality-year observation, colour denotes the predictor's percentile value, and points to the right (left) of the dashed zero line raise (lower) the predicted net migration rate. SHAP values describe predictive contributions within the fitted model, not causal effects.*

**Table 5**

**Global SHAP importance and dominant contribution direction of the ten leading predictors.**

| Rank (95% CI) | Predictor | Mean \|SHAP\| (‰) | Dominant direction | Rank range across stages | Permutation p (FDR q) |
|---|---|---|---|---|---|
| 1 (1–3) | Childcare facilities | 2.59 | − | 1 | 0.001 (0.003)** |
| 2 (1–4) | Housing age | 1.99 | − | 2 | 0.378 (0.473) |
| 3 (1–6) | Closeness centrality | 1.87 | + | 3 | 0.023 (0.038)* |
| 4 (1–8) | Sewerage supply | 1.85 | − | 4 | <0.001 (0.002)** |
| 5 (2–9) | PageRank (t−1) | 1.59 | + | 8 | <0.001 (0.002)** |

| | | | | | |
|---|---|---|---|---|---|
| 6 (3–12) | Senior-care facilities | 1.41 | + | 2 | 0.116 (0.155) |
| 7 (4–14) | Doctors | 1.36 | − | 3 | 0.004 (0.010)* |
| 8 (5–13) | Population density | 1.07 | − | 13 | <0.001 (0.002)** |
| 9 (6–17) | Distance from Seoul | 0.98 | − | 3 | 0.107 (0.153) |
| 10 (8–16) | Ageing ratio | 0.91 | + | 6 | 0.004 (0.010)* |

*Note: Mean |SHAP| values are computed from the single fitted model and evaluated on the complete-case sample (n = 1,480 municipality-year observations; Section 3.2). Confidence intervals for the global rank are percentile intervals from 50 bootstrap replications with model refitting. Direction denotes the dominant sign of the association between feature values and SHAP contributions over the full sample. As Sections 4.3.3 and 4.4 show, that sign may itself vary across development stages. The rank range is the difference between a predictor's maximum and minimum within-stage rank. The permutation p-value tests the null hypothesis that this range arises under random assignment of development stages (2,000 permutations), and q is the Benjamini–Hochberg false-discovery rate across the twenty simultaneous tests. The test concerns the invariance of a predictor's rank across spatial regimes, not whether a coefficient differs from zero: SHAP contributions are attributions of a fitted model, not estimated parameters. Asterisks refer to q: * q < 0.05. ** q < 0.01. *** q < 0.001. Sensitivity to the missing-data treatment is reported in Appendix Table A13.*

The global SHAP ranking (Table 5) identifies childcare provision (mean |SHAP| = 2.59‰), housing age (1.99‰), and closeness centrality (1.87‰) as the three leading predictors, followed by sewerage supply (1.85‰) and PageRank (1.59‰).

Older housing stock, childcare provision, and sewerage supply contribute negatively to predicted migration, whereas closeness centrality and PageRank contribute positively. Childcare provision therefore combines the largest overall importance with a predominantly negative contribution direction, a combination examined in Section 4.4.

These rankings summarize the predictive structure of the fitted model but do not indicate whether contributions are spatially uniform: high global importance may arise from consistent contributions everywhere or from strong contributions confined to particular urban contexts. Global importance therefore serves only as the point of departure for evaluating context dependency (predictor correlations are documented in Appendix Figure A4).

### 4.3.2 Development-stage comparison

To evaluate stage dependency we partition the SHAP matrix of the fitted model by population-density tertile and re-rank predictors within each stage (Section 3.3.2.3). The stage-stratified analysis covers 1,480 municipality-year observations with complete data (low-density: 588; intermediate: 528; high-density: 364). Table 6 reports the within-stage ranks of five focal predictors with bootstrap confidence intervals and permutation q-values.

**Table 6**
**Within-stage SHAP-importance ranks of the five focal predictors, with bootstrap confidence intervals and permutation tests of stage dependence.**

| Predictor | Low density | Intermediate | High density | Rank range | Permutation p (FDR q) |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Childcare facilities | 2 (1–2) | 1 (1–1) | 1 (1–2) | 1 | 0.001 (0.003)** |
| Housing age | 4 (3–4) | 2 (2–4) | 3 (2–5) | 2 | 0.378 (0.473) |
| Closeness centrality | 3 (3–4) | 3 (2–4) | 6 (4–7) | 3 | 0.023 (0.038)* |
| Sewerage supply | 6 (5–8) | 4 (2–4) | 2 (1–3) | 4 | <0.001 (0.002)** |
| PageRank (t−1) | 1 (1–2) | 6 (5–8) | 9 (8–9) | 8 | <0.001 (0.002)** |

*Note: Ranks are computed within each stage among all twenty predictors by partitioning the attributions of the single fitted model (n = 1,480). Values in parentheses are 95% percentile confidence intervals from the municipality-level cluster bootstrap (2,000 replications), which resamples municipalities rather than municipality-years so that the panel structure is preserved. The rank range is the difference between the maximum and minimum rank across tiers. P-values are from the stage-label permutation test (2,000 permutations), whose null hypothesis is that a predictor's importance rank is invariant across development stages, and q is the Benjamini–Hochberg false-discovery rate across the twenty simultaneous tests (* q < 0.05; ** q < 0.01; *** q < 0.001). Significance and magnitude must be read together: a predictor occupying the same one or two positions in every tier has a compressed null distribution, so a rank range of one may be statistically detectable yet substantively negligible. Full results for all twenty predictors are reported in Appendix Table A10.*

Stage dependency is widespread: twelve of the twenty predictors have rank ranges exceeding their permutation null distribution, and all twelve remain significant after Benjamini–Hochberg correction for the twenty simultaneous tests ($q < 0.05$; Appendix Table A10). Its distribution across the ranking, however, is not what global importance would suggest. The four leading predictors move by at most four rank positions across tiers (childcare provision 2/1/1; housing age 4/2/3; closeness centrality 3/3/6; sewerage supply 6/4/2) and each remains within the top six in every tier. PageRank, ranked fifth globally, is the first predictor to display substantial contextual variation: it is the single most informative predictor in the low-density tier (rank 1; 95% CI 1–2) and falls to rank 9 (8–9) in the high-density tier, a rank range of 8 that lies far outside its permutation null ($q = 0.002$), with non-overlapping confidence intervals. Statistical significance and substantive magnitude must be read together: childcare provision attains $q = 0.003$ with a rank range of one, because a predictor that occupies first or second position in every tier has an extremely compressed null distribution. The rank range, not the p-value alone, measures the extent of contextual variation.

Global importance and stage dependency are therefore largely orthogonal. The largest re-orderings occur among predictors of intermediate global importance: population density moves from rank 17 in the low-density tier to rank 4 in the high-density tier (range 13; $q = 0.002$), the ageing ratio from 7 to 13 (range 6; $q = 0.010$), and fiscal independence from 10 to 15 (range 7; $q = 0.002$), whereas the predictors at the top of the ranking are comparatively stable. A predictor's position in the global ranking therefore carries little information about whether its contribution is uniform across urban contexts or concentrated within particular ones.

Figure 6 extends the comparison to all twenty predictors and marks those whose rank range exceeds the permutation null after correction for multiple testing. The result also shows why a fixed threshold is inadequate: because the null distribution of the rank range differs across

predictors, an absolute cut-off both misses genuinely stage-dependent predictors with small ranges (doctors: range 3, q = 0.010) and admits predictors whose apparent variation lies within the null (business establishments: range 1, q = 1.00; fertility rate: range 2, q = 0.58).

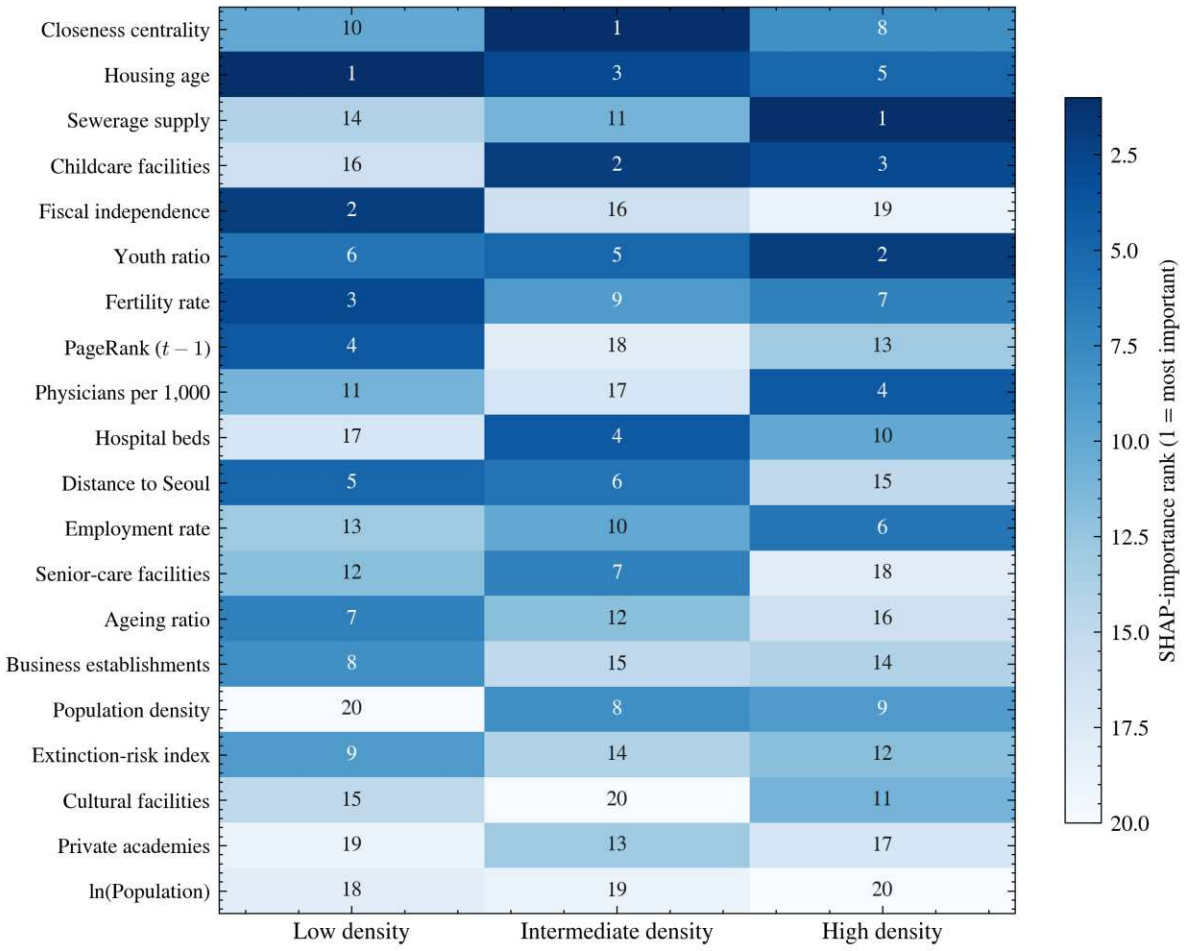



**Figure 6**

**Within-stage SHAP-importance ranks of the twenty predictors, with permutation tests of stage dependence.**

*Note: Cells give the within-stage SHAP-importance rank among all twenty predictors (1 = most important), obtained by partitioning the attributions of the single fitted model by population-density tertile (n = 1,480). Predictors are ordered by global importance (Table 5). Asterisks mark predictors whose observed rank range exceeds the predictor-specific permutation null distribution after Benjamini–Hochberg correction across the twenty simultaneous tests (* q < 0.05; ** q < 0.01; 2,000 permutations of the stage labels across municipalities; Section 3.3.2.3). Full results, including bootstrap confidence intervals, are reported in Appendix Table A10.*

### 4.3.3 Network position as the representative case

Network position is examined as the representative case because migration-network research has treated centrality as a universally beneficial structural characteristic (Section 2.2). If any predictor were expected to retain stable importance across municipalities, it would be network centrality. The two network measures, behave differently from one another, and only one of them is stage dependent.

Figure 7 shows that PageRank is the single most informative predictor within low-density municipalities (rank 1 of 20) and that its importance declines monotonically across the density gradient to rank 9, whereas closeness centrality remains within the top six in every tier (ranks 3, 3, and 6). Importance is a magnitude. Its direction must be read separately. Within the low-density tier the mean SHAP contribution of PageRank is −1.70‰. In the intermediate and high-density tiers it is +0.72‰ and +1.18‰.

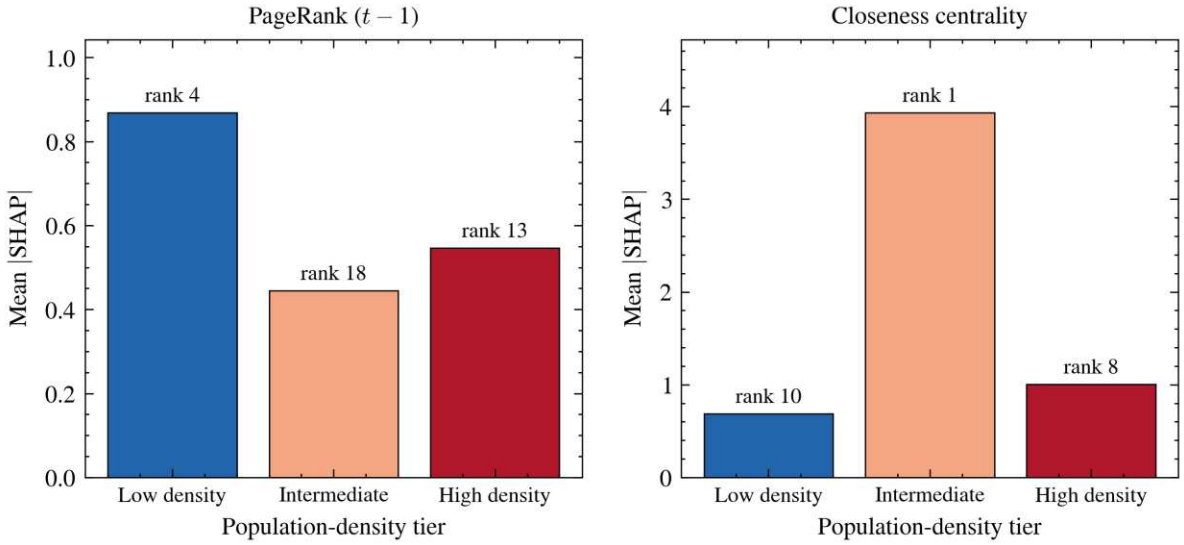



**Figure 7**

**Development-stage variation in the SHAP importance of PageRank and closeness centrality, with cluster-bootstrap confidence intervals.**

*Note: Bars show the mean absolute SHAP contribution of each predictor within each population-density tier, partitioned from the single fitted model (n = 1,480). Whiskers give 95% confidence intervals from the municipality-level cluster bootstrap (2,000 replications), which resamples municipalities rather than municipality-years so that the panel structure is preserved. Labels report the within-tier importance rank among all twenty predictors and its bootstrap confidence interval. Permutation p-values test the null hypothesis that the observed rank range arises under random assignment of development stages. Bars measure magnitude, not direction: the mean signed contribution of PageRank is −1.70‰ in the low-density tier and +1.18‰ in the high-density tier (Section 4.3.3).*

The interpretation therefore differs from the one that importance alone would suggest. PageRank dominates prediction in low-density municipalities not because network prominence lifts them, but because their weak position within the migration network is what the model relies on most heavily to predict their population loss: the contribution is large in magnitude and negative in sign. Where PageRank contributes positively, in the intermediate and high-density tiers, its importance is comparatively modest. Prominence within the migration network is thus most informative precisely where municipalities lack it, which is not the same as being an asset that peripheral municipalities could acquire. Relative accessibility, by contrast, carries predictive information across the settlement hierarchy, and its contribution shifts from negative in sparsely populated municipalities (−1.60‰) to positive in dense ones (+0.93‰) without a comparable change in importance.

The predictive relevance of network prominence is therefore not uniform but conditional upon urban development stage, whereas that of network accessibility is not; and in neither case does the magnitude of a predictor's contribution disclose its sign.

It should be emphasized that this interpretation concerns variation in model attribution, not variation in underlying causal effects. The results indicate that the fitted predictive model relies on PageRank differently across urban contexts. They do not establish that the causal influence of network position itself changes across development stages.

### 4.3.4 The direction of contributions

The rank comparisons above measure how strongly each predictor contributes within a development stage. They cannot measure in which direction it contributes, because a rank is computed from absolute contributions. Direction is what a policy reading of an importance ranking implicitly supplies (the top-ranked variable is assumed to be one whose increase would improve the outcome) so it must be examined separately. Figure 8 plots the signed local contributions of the four focal predictors against population density.

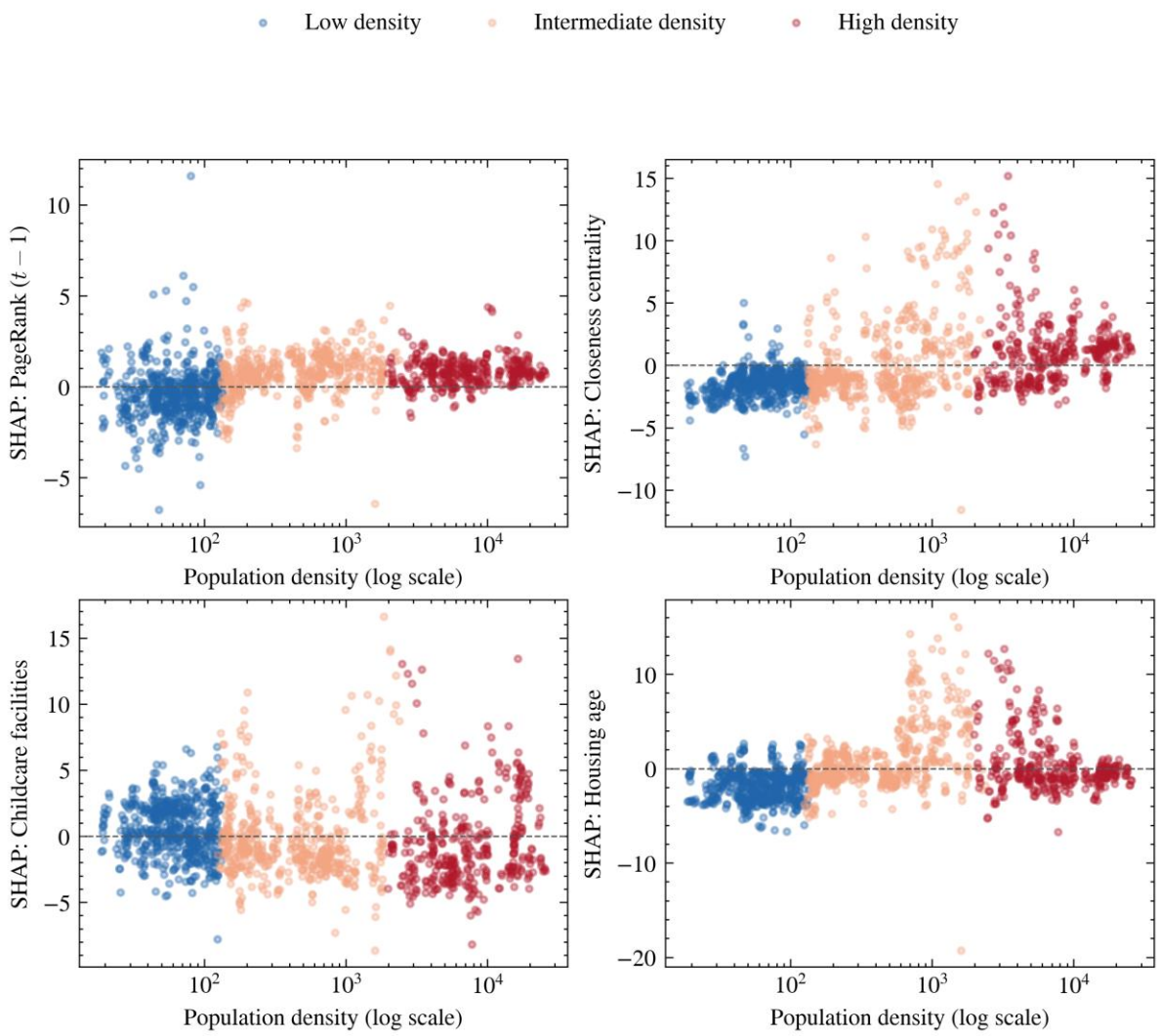



**Figure 8**

**Signed SHAP contributions of the four focal predictors along the density gradient.**

*Note: Each panel plots the signed local SHAP contribution of one predictor (vertical axis, ‰) against municipal population density (horizontal axis, log scale) for the 1,480 complete-case municipality-year observations. Colours denote population-density tiers (blue = low, orange = intermediate, red = high; consistent with Figure 7) and the dashed line marks zero contribution. Unlike the rank comparisons in Table 6, which are computed from absolute contributions and therefore measure magnitude only, this figure displays direction: PageRank and childcare provision each reverse the sign of their contribution across the settlement hierarchy (Section 4.3.4).*

The signed contributions change systematically along the density gradient, and for two predictors they change sign. PageRank contributes strongly negatively among sparsely populated municipalities (mean −1.70‰) and positively as municipalities become more urbanized (+0.72‰ and +1.18‰), tracking the underlying gradient in network prominence: the predictor that dominates prediction in the low-density tier does so by registering what those municipalities lack. Closeness centrality follows the same qualitative pattern with a smaller amplitude (−1.60‰ to +0.93‰).

Childcare provision moves in the opposite direction: its contribution is slightly positive in the low-density tier (+0.33‰) and becomes increasingly negative as density rises (−0.91‰ and −1.26‰). The globally leading predictor therefore reverses the sign of its contribution across the settlement hierarchy while its importance rank stays essentially constant (rank range 1; Table 6). Housing age, by contrast, contributes negatively throughout, most strongly in the low-density tier (−1.53‰), where the housing stock is oldest.

Variation in importance and variation in direction are distinct properties, and they do not coincide. PageRank varies in both: its rank falls from 1 to 9 across the density gradient, and the sign of its contribution turns from negative to positive. Childcare provision varies only in direction: its rank is effectively constant (2/1/1), yet the sign of its contribution reverses, a change that no comparison of ranks could detect, because ranks are computed from absolute contributions. Housing age varies in neither, contributing negatively and with similar strength throughout. The rank test therefore establishes one form of context dependence and is blind to the other.

This is why the two analyses are complementary rather than redundant. The permutation test establishes that the ordering of predictors is not invariant across spatial regimes. The signed contributions establish that neither is the meaning of a predictor's position in that ordering. An importance ranking, read on its own, conceals both.

Taken together, the results answer RQ2. The predictive contribution of migration determinants is systematically conditioned by urban development stage for twelve of twenty predictors, all of which survive correction for multiple testing. The conditioning is not concentrated among the most important predictors. Network prominence provides the clearest illustration: PageRank moves from rank 1 to rank 9 across the density gradient ($q = 0.002$), while the four globally leading predictors vary by at most four positions. The assumption that globally important predictors contribute uniformly across municipalities is therefore not supported. Neither is the converse assumption that importance and context dependency travel together.

## 4.4 RQ3, Under spatial heterogeneity, feature importance indicates neither effect direction nor policy relevance

RQ3 concerns what follows for interpretation once predictor contributions are shown to be stage dependent. If they vary across urban contexts, a global ranking should not be read as identifying universally applicable priorities.

Childcare provision provides the clearest illustration. Although childcare is ranked as the most important predictor in the fitted model (Figure 5), its SHAP contributions are

predominantly negative across municipalities, and its contribution becomes increasingly negative within dense municipalities close to the capital region (Figure 9).

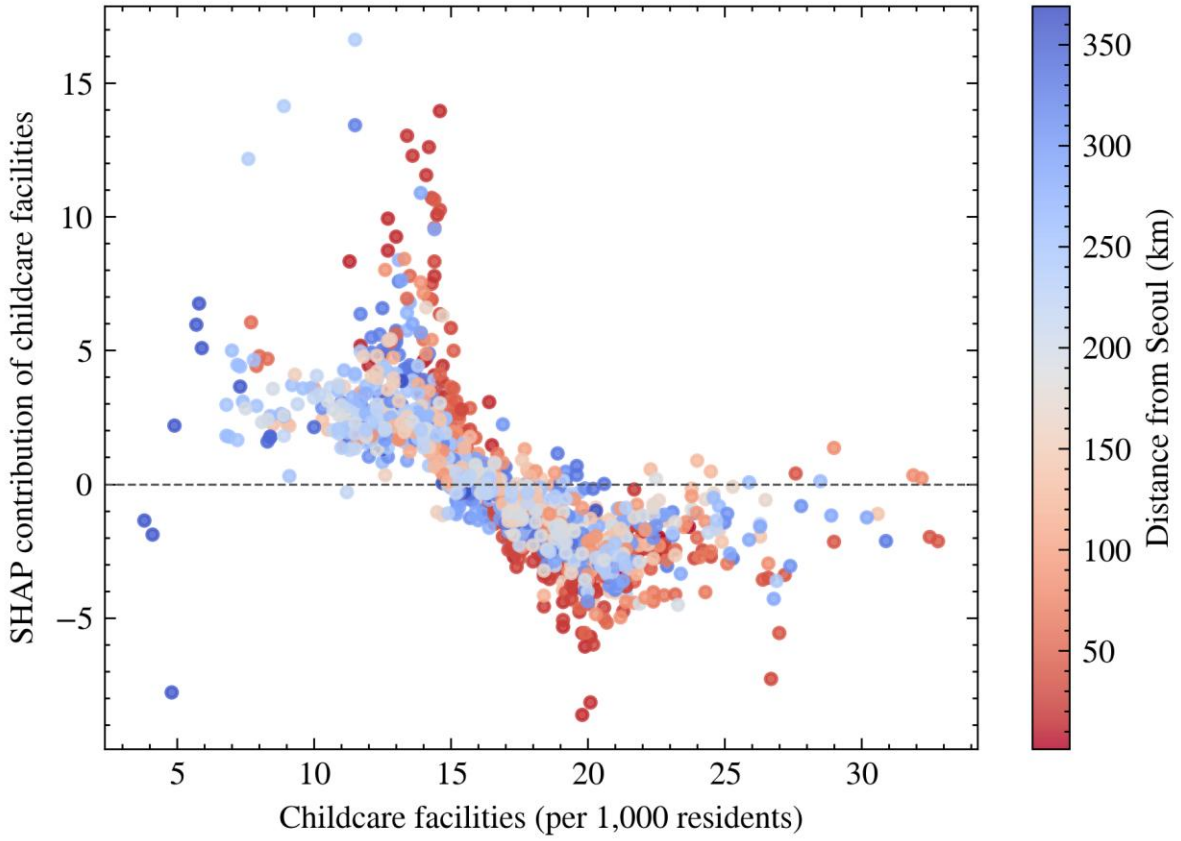



**Figure 9**
**SHAP dependence plot for childcare provision.**

*Note: Each point is one of the 1,480 complete-case municipality-year observations. The horizontal axis shows childcare facilities per 1,000 residents, the vertical axis the local SHAP contribution of childcare provision (‰), and colour denotes distance from Seoul (red = near the capital, blue = far). The dashed line marks zero contribution. The mean contribution is +0.33‰ in the low-density tier and −1.26‰ in the high-density tier (Section 4.4).*

Unlike PageRank, childcare provision is not concentrated within any development stage: it ranks second, first, and first across the low-, intermediate-, and high-density tiers, a rank range of one (Table 6). Its importance is therefore uniform, but its contribution is not: the mean SHAP contribution shifts from +0.33‰ in the low-density tier to −1.26‰ in the high-density tier. Uniformly high importance thus coexists with a contribution whose sign depends on urban context.

This distinction is important. A predictor may possess high predictive importance because it captures substantial variation in migration outcomes, even when its association with migration is predominantly negative.

To examine whether this result merely reflects the use of per-capita indicators, the major service variables were re-estimated using absolute facility counts (Appendix Table A7).

The robustness analysis shows that childcare retains both its negative SHAP direction and its leading importance rank (first to second) after conversion to absolute counts, as do hospital beds. The transformation does reshuffle importance among the remaining service indicators (doctors fall from rank 7 to 16 and senior-care facilities from 6 to 15, while private academies rise from 19 to 7) but no service indicator reverses from a negative to a positive contribution direction.

The negative association observed for childcare therefore cannot be attributed to a denominator effect arising from per-capita measurement.

Several mechanisms could plausibly generate this pattern. Municipalities experiencing demographic decline may expand childcare provision on a per-capita basis as population decreases. Alternatively, extensive childcare infrastructure may be concentrated within high-demand metropolitan municipalities that nevertheless experience family out-migration because of housing costs. The observed relationship may also reflect broader ecological associations operating at the municipal scale.

The present analysis cannot distinguish among these mechanisms because SHAP explains prediction rather than causal processes.

Precisely for this reason, childcare illustrates the central methodological argument of this study, and PageRank illustrates it a second time. High feature importance indicates that a variable contains substantial predictive information, but it establishes neither the direction of its association nor that increasing it would improve migration outcomes: childcare provision is the leading predictor and contributes negatively wherever municipalities are dense, and PageRank is the leading predictor in low-density municipalities precisely because they lack network prominence. Nor does importance indicate whether that information is spatially concentrated: childcare is the most important predictor and the most uniformly informative one, whereas PageRank is moderately important and strongly stage dependent. Feature importance identifies where predictive information is concentrated within the fitted model, and, as these contrasts show, the ranking itself reveals neither where nor in which direction.

Taken together, these findings answer RQ3 by demonstrating that predictor importance and policy relevance belong to different inferential domains: identifying the variables on which a predictive model relies is not the same as identifying the levers that policy could pull. Once predictor contributions become conditional upon urban development stage, global

importance rankings should be read as evidence on the spatial organization of prediction (where, and through which variables, the model finds its predictive information) rather than as universal intervention priorities.

Table 7 consolidates this interpretive framework for the four focal predictors examined in Sections 4.3–4.4.

**Table 7**
**Interpretation matrix of the focal predictors under spatial heterogeneity.**

| Predictor | Global importance (rank) | Dominant SHAP direction | Tier pattern (rank: low/int./high) | Interpretive status |
|---|---|---|---|---|
| Childcare facilities | High (1) | − | 2 / 1 / 1 (rank range 1) | Uniformly the leading source of predictive information; negative direction; not a policy lever |
| Housing age | High (2) | − | 4 / 2 / 3 (rank range 2; p = 0.38) | Structural condition; informative in every tier |
| Closeness centrality | High (3) | + | 3 / 3 / 6 (rank range 3; p = 0.02) | Accessibility advantage; broadly informative, modestly weaker in dense cores |
| PageRank (t−1) | Moderate (5) | + | 1 / 6 / 9 (rank range 8; p < 0.001) | Predictive information concentrated in low-density municipalities |

*Note: Tier ranks are within-stage SHAP-importance ranks among all twenty predictors (low-, intermediate-, and high-density tiers; Figure 6), with permutation p-values from the stage-label permutation test (Section 3.3.2.3). Direction denotes the dominant sign of SHAP contributions across the full sample. As Sections 4.3.3 and 4.4 show, the sign of a predictor's contribution may itself vary across tiers. Interpretive status summarizes the reading supported by the evidence in Sections 4.3–4.5. None of the entries implies a causal effect.*

## 4.5 Robustness analyses

We examine whether the principal findings depend upon modelling choices, attribution methods, or variable construction (Section 3.4).

First, importance rankings were compared across five attribution approaches (CatBoost, XGBoost, and LightGBM SHAP; random-forest gain; permutation importance). Rank agreement with CatBoost was moderate to strong (Spearman $\rho = 0.76–0.93$), and childcare provision, housing age, and closeness centrality appear among the three leading predictors under every method (Appendix Figure A5). The stage-dependent pattern is likewise not algorithm specific: under XGBoost, PageRank is again the leading predictor in the low-density tier and falls to rank 7 in the high-density tier, while closeness centrality remains uniformly informative (Appendix Table A12).

Second, bootstrap resampling with model refitting (50 replications) reproduced the leading ranks: childcare provision, housing age, and closeness centrality occupied the first three median ranks and PageRank a median rank of 5 (95% CI: 2–9; Appendix Figure A6). Intervals

widen markedly below the top of the ranking, so fine distinctions among mid-ranked predictors should not be over-interpreted.

Third, under leave-one-province-out cross-validation, mean out-of-block performance fell to $R^2 = 0.05$ (range: −0.68 for Seoul to 0.38) against 0.36 under temporal validation (Appendix Figure A7), and a model trained on the Seoul Capital Area produced negative out-of-region $R^2$ in ten of thirteen provinces. Limited transferability cannot by itself establish spatial heterogeneity, but it is consistent with the differentiation reported in Section 4.2.

Fourth, predictor construction was evaluated by replacing the per-capita service indicators with absolute facility counts and repeating the analysis (Appendix Table A7). The conclusions concerning childcare and hospital beds are unchanged (both retain a negative contribution direction, and childcare remains among the two leading predictors) indicating that the central interpretation does not depend on indicator normalization.

Fifth, an alternative specification in which a separate model is trained within each development stage reproduces the focal finding, PageRank still declines across the density gradient (ranks 4, 6, and 11), but produces systematically larger rank ranges (mean 7.7 against 4.0) and agrees only moderately with the partitioned ranking (Spearman $\rho = 0.36$–$0.73$ across tiers; Appendix Table A11). Because each stage-specific model is estimated on 456 training observations with hyperparameters tuned on the full panel, part of this dispersion reflects estimation noise rather than urban context. Holding the attribution basis fixed isolates the two, and the partitioned specification is therefore retained.

Sixth, evaluating SHAP on the full imputed panel rather than on complete cases leaves the global ranking essentially unchanged (Spearman $\rho = 0.99$), confirming that the reported attributions are not driven by imputed values. Re-estimating the model on complete cases only (discarding a quarter of the training records, most owing to missing employment-rate data) produces a more noticeable change ($\rho = 0.82$; sewerage supply falls from rank 4 to 8), and under that specification XGBoost rather than CatBoost attains the highest validation $R^2$ (Appendix Table A13). PageRank remains strongly stage dependent under both training samples and both algorithms. The result is a further instance of the paper's argument: a preprocessing decision that leaves predictive performance essentially unchanged can move predictors several positions in the importance ranking.

Overall, the robustness analyses consistently support the three principal findings of this study. First, municipal migration exhibits a geographically differentiated spatial structure. Second, predictor contributions vary systematically across urban development stages. Third,

under such spatial heterogeneity, feature importance should be interpreted as a measure of predictive information rather than as evidence of universally applicable policy priorities.

# 5. Discussion

The three contributions are developed in turn: the spatial organization of migration outcomes (5.1), the stage dependence of predictor contributions (5.2), and what both imply for reading feature importance (5.3).

## 5.1 Spatial organization of migration: revisiting Korean migration geography

The first contribution of this study is empirical. Municipal migration forms not a simple urban–rural gradient but a geographically differentiated system organized around capital-region concentration, suburban redistribution, and dual patterns of regional decline. The inverted-U density–migration relationship and the geography of gains and losses (Sections 4.2.1–4.2.3) show that urbanization alone cannot explain municipal population change: losses characterize both ends of the density spectrum.

This finding refines the common characterization of Korean internal migration as unidirectional concentration toward the capital region in two respects.

First, contemporary concentration is redistributive rather than centripetal. Earlier stages of Korean urbanization were characterized largely by migration toward Seoul itself (Mills & Song, 1979; Joo, 2018). The present results instead show the capital region redistributing internally: Seoul's twenty-five districts lose population (−7.17‰; twenty-three of twenty-five), while the Gyeonggi ring gains (+9.20‰). Capital concentration should therefore be understood not as the growth of Seoul but as the continuing outward expansion of the capital-region urban system.

Second, population decline outside the capital region is not a single process. Rural and non-capital metropolitan municipalities lose population at indistinguishable average rates, but rural decline is pervasive and metropolitan decline is selective (Section 4.2.2). Averaging the two conceals that difference, and policies calibrated to the average will fit neither.

This distinction provides empirical support for recent arguments in the shrinking-city literature that different forms of regional decline require different conceptual explanations and policy responses (Martinez-Fernandez et al., 2012; Khavarian-Garmsir, 2023).

Taken together, the results locate the organizing dimension of contemporary Korean migration in the density gradient rather than in capital-region accessibility, which modulates

the gradient without producing it (Section 4.2.1). This qualifies a common reading of the Korean case, in which distance from Seoul is treated as the master variable. Accessibility determines how large the suburban peak is. It does not determine where the peak lies.

## 5.2 Stage-dependent predictor contributions: extending migration-network theory

Having established that migration outcomes themselves are spatially differentiated, the second contribution concerns the structure of migration determinants.

Existing migration-network research has applied network analysis to characterize the structural position of regions within migration systems (Danchev & Porter, 2021; Pitoski et al., 2021; Gürsoy & Badur, 2022), and network position has generally been regarded as an intrinsically advantageous characteristic. Municipalities occupying more central positions within migration networks are typically assumed to attract migrants because of their structural accessibility and connectivity.

Our results qualify this assumption, and in doing so invert it. PageRank is the single most informative predictor within the low-density tier (rank 1 of 20) and recedes to rank 9 in the high-density tier (permutation $q = 0.002$; Section 4.3.2). Its contribution there is negative (mean SHAP = −1.70‰): the model relies on network position most heavily in sparsely populated municipalities not because prominence lifts them, but because the absence of prominence is what most strongly predicts their population loss. Where PageRank contributes positively, in denser tiers, it is comparatively unimportant. Closeness centrality, by contrast, remains within the top six predictors throughout. Because the tiers are cross-sectional, this pattern describes differences across urban contexts rather than a temporal trajectory of individual municipalities.

Network position is therefore neither a single attribute nor a uniformly beneficial one. Prominence and accessibility behave differently from one another, only prominence is stage dependent, and prominence is most informative precisely where municipalities lack it, which is not the same as being an asset that peripheral municipalities could acquire. The aggregate treatment of centrality as a positional advantage in the existing literature does not admit these distinctions.

This does not contradict econometric analyses of migration (Section 2.1): the present study evaluates a different quantity, the relative contribution of predictors within a nonlinear prediction model. A variable may possess a significant causal effect yet contribute modestly to prediction when correlated variables carry the same information, and a highly important predictor need not have a large causal effect.

The two analytical approaches are therefore complementary rather than competing. Panel models estimate average causal relationships under explicit identifying assumptions, whereas interpretable machine-learning models describe how information is organized for prediction across heterogeneous urban contexts.

Stage dependency is not confined to network position: comparable patterns across the wider predictor set (Section 4.3.2) indicate that context dependency characterizes the predictive system as a whole. Urban development stage should therefore be regarded as a boundary condition on predictor contributions in municipal migration models, a conclusion that converges with geographically weighted evidence that migration determinants vary systematically across space (Li et al., 2023), and that extends it from spatially varying coefficients to the interpretation of feature importance in nonlinear models.

The principal implication of this section is that predictor contributions, like migration outcomes themselves, are stage dependent. This stage dependency raises a broader methodological question: how should feature importance be interpreted when predictor contributions vary systematically across space?

## 5.3 Interpreting feature importance under spatial heterogeneity

The principal theoretical contribution of this study concerns the interpretation of feature importance in explainable machine learning. While SHAP has become widely adopted for identifying influential predictors in regional science, feature importance is frequently interpreted as if it represented a universally beneficial determinant or a direct policy priority. Our findings suggest that such an interpretation is not generally warranted when the underlying process exhibits spatial heterogeneity. Recent work in explainable AI has likewise emphasized that predictive feature attribution should not be conflated with causal explanation (Janzing et al., 2020; Ma & Tourani, 2020). The present study extends this distinction to spatially heterogeneous regional systems.

The logic is cumulative. Spatial heterogeneity in migration outcomes (RQ1) alters the joint distribution of predictors and outcomes, and it is through that joint distribution that the predictive information carried by individual variables becomes context dependent (RQ2). Feature importance is therefore not spatially invariant, and, crucially, the ranking discloses neither which of its entries are context dependent nor in which direction they act.

Childcare provision illustrates the second point. It is the leading predictor globally and the most uniformly informative one, yet the sign of its contribution reverses across the settlement hierarchy, from positive in sparsely populated municipalities to strongly negative in dense ones (Section 4.4). Read literally as policy importance, its top ranking would imply

that expanding childcare provision should stem migration decline. The predictive analysis supports no such inference, and indeed points the other way wherever municipalities are dense.

The childcare result shows instead that feature importance identifies where predictive information is concentrated within the fitted model, not the mechanisms generating it. Several processes could produce the pattern (Section 4.4), and SHAP cannot distinguish among them, which is the point: predictive importance remains useful for prediction without establishing mechanism or policy relevance.

Network position carries the same lesson from the opposite direction. Often regarded as a universally advantageous structural characteristic, it proves both stage dependent and, where it matters most, negative in sign (Section 4.3.3). Between them, the two predictors exhaust the ways in which an importance ranking can mislead: the most important predictor is uniformly informative but directionally unstable, and the most context-dependent predictor is most informative where its contribution is adverse.

The broader methodological implication is straightforward. High SHAP importance indicates that a variable contains substantial predictive information within the fitted model, but it establishes neither the sign of its contribution, nor whether that contribution is spatially uniform, nor the magnitude of any causal effect, nor the effectiveness of policy intervention. These are distinct inferential questions requiring complementary analytical frameworks (Janzing et al., 2020; Ma & Tourani, 2020). Shapley-based importance measures, carry well-documented mathematical and interpretive limitations that reinforce the need for such caution (Kumar et al., 2020).

This stance lies between two traditions. Geographically weighted approaches explain how regression relationships vary across space (Li et al., 2023). SHAP applications report importance rankings without asking whether their interpretation is itself spatially conditional. The contribution here is to connect them.

We therefore propose an interpretive principle for explainable machine learning in regional science. Whenever the underlying phenomenon exhibits spatial heterogeneity, feature importance should be read as evidence about the spatial organization of predictive information, where, and through which variables, a model finds what it needs to predict, rather than as a measure of universal policy priority. Direction and spatial uniformity must be established separately, because the ranking encodes neither. We developed the principle on municipal migration in South Korea, but it is transferable to other geographically differentiated processes, including housing markets, regional innovation, accessibility,

environmental exposure, and urban resilience, where predictor contributions are likewise likely to vary across places.

## 5.4 Policy implications

The findings suggest that regional development policy should be differentiated according to municipalities' positions within the national migration system rather than applied as a uniform national strategy. Growing capital-region suburbs, declining non-capital metropolitan cities, and ageing rural municipalities occupy fundamentally different positions within the settlement system and face different structural challenges.

For rapidly growing suburban municipalities, policy priorities concern housing supply and the management of continued population concentration while mitigating congestion and escalating housing costs. In contrast, non-capital metropolitan municipalities appear to require policies that strengthen regional economic functions and expand long-term employment opportunities. Their continued population decline despite relatively developed infrastructure suggests that infrastructure investment alone is unlikely to reverse migration losses. Rural municipalities face a different set of challenges associated with demographic ageing, service provision and long-term population sustainability, making policies that maintain essential public services and quality of life more appropriate than attempts to replicate metropolitan growth trajectories.

The findings also caution against translating explainable-machine-learning outputs directly into policy priorities. Highly ranked predictors do not necessarily represent positive migration drivers, and policy decisions should therefore not rely on feature-importance rankings without considering predictor direction, spatial context and plausible causal mechanisms.

Explainable machine learning is instead likely to be most valuable as a diagnostic tool that identifies where predictive information is concentrated. Monitoring network position within the low-density tier, for example, can flag peripheral municipalities that are gaining or losing ground within the migration system, without implying that strengthening network connectivity would itself attract migrants (Section 5.3). Such diagnostics help policymakers distinguish regional development contexts before interventions are selected on the basis of complementary causal evidence.

## 5.5 Limitations and future research

Several limitations should be acknowledged when interpreting the present findings.

First, the analysis is predictive rather than causal. SHAP decomposes a fitted model's predictions into feature contributions. It does not identify the mechanisms generating

migration. Stage-dependent contributions should therefore not be read as evidence that the causal effects of predictors differ across municipalities, only that the predictive information available to the model does. Establishing causal heterogeneity requires designs built for identification: spatial panel models, natural experiments, and heterogeneous-treatment-effect methods (e.g., Wager & Athey, 2018), against which the structures documented here can be tested.

Second, development stage was operationalized using population-density tertiles. Density is a widely used indicator of urbanization but does not exhaust it. The study therefore does not claim that density represents urban development in full, only that it provides a transparent and reproducible basis for asking whether predictive relationships differ across broad urban contexts. Whether comparable patterns emerge under multidimensional development indices remains open.

Third, the scope of the evidence is bounded. The findings rest on municipal-level data for 2017–2024, a period that includes the COVID-19 pandemic, and describe aggregate differences among municipalities rather than individual migration decisions. The external validity of the specific patterns is likewise limited by the pronounced single-pole structure of the Korean settlement system. It is the interpretive principle proposed in Section 5.3, rather than the particular importance rankings, that is expected to travel across national contexts.

Despite these limitations, the principal contribution of this study remains robust. Rather than proposing a new causal explanation for Korean migration, the study advances an interpretive framework for explainable machine learning under spatial heterogeneity. By demonstrating that migration outcomes are spatially differentiated, that predictor contributions vary systematically across urban development stages, and that feature importance should therefore be interpreted as the spatial organization of predictive information rather than universal policy relevance, the study provides a conceptual foundation for integrating interpretable machine learning with spatial and causal analyses in future regional research.

# 6. Conclusion

This study examined how feature importance from explainable machine learning should be interpreted when the process being modelled is spatially heterogeneous. The setting is municipal migration across 229 South Korean municipalities between 2017 and 2024. Rather than asking only which predictors matter most, we asked what mattering most means when migration itself is spatially organized.

Three findings answer this question in sequence. First, municipal migration follows a geographically differentiated structure rather than a single urban–rural gradient: net migration is an inverted-U function of population density, gains concentrate in suburban municipalities within commuting distance of the capital region, and two structurally distinct forms of decline, rural and non-capital metropolitan, coexist within the national settlement system.

Second, the predictive contributions of migration determinants are systematically stage dependent, but not in the way a global ranking would suggest. Twelve of twenty predictors show rank ranges exceeding a permutation null, yet the leading predictors are the most stable across urban contexts and the largest re-orderings occur among predictors of intermediate importance. Network prominence illustrates the point most sharply. PageRank is the single most informative predictor within low-density municipalities and recedes to ninth in dense ones. Its contribution in the tier where it matters most is negative, reflecting the absence of network position rather than its advantage.

Third, and consequently, feature importance cannot be read as a universal ranking of policy priorities. High importance indicates where predictive information is concentrated within the fitted model. It establishes neither the direction of a predictor's contribution (the leading predictor here reverses sign across the settlement hierarchy), nor whether that contribution is spatially uniform, nor the effectiveness of intervening on it. Predictive importance and intervention relevance belong to different inferential domains. We therefore propose an interpretive principle for explainable machine learning in regional science: under spatial heterogeneity, feature importance should be read as evidence about the spatial organization of predictive information rather than as a measure of universal policy priority.

For practice, the results imply a division of labour. Explainable machine learning is most valuable as a diagnostic tool that reveals how predictive information is organized across regional development contexts. It can flag, for instance, where weak network position carries early signals about the trajectories of peripheral municipalities. Decisions about intervention require complementary causal evidence. Regional policy should be

differentiated across municipal types rather than derived from a single national strategy or a global importance ranking.

The contribution lies not in proposing a new determinant of migration but in showing how existing determinants should be read under spatial heterogeneity. Spatial heterogeneity is the norm rather than the exception in regional phenomena. The interpretive principle developed here, grounded in Korean municipal migration but not specific to it, therefore offers a template for reading explainable machine learning in housing, innovation, accessibility, and other geographically differentiated processes.

## Acknowledgements

The authors are grateful to Statistics Korea (KOSIS), the Ministry of the Interior and Safety, and the Health Insurance Review and Assessment Service for making the underlying public statistics available.

## Funding Statement

This study was supported by the Academic Research Funding of Baekseok University in 2026.

## Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the study, authorship, and/or publication of this article.

## Data Availability Statement

The data supporting the findings of this study are official aggregate statistics from Statistics Korea (KOSIS), the Ministry of the Interior and Safety, and the Health Insurance Review and Assessment Service. The analysis code and canonical result tables underlying all reported estimates are openly available in the study's replication repository at https://github.com/dongwoo2022008/spatial-heterogeneity-feature-importance. The

harmonized origin–destination flow dataset is available from the corresponding author upon reasonable request.

## References


Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623–2631). https://doi.org/10.1145/3292500.3330701

Bell, M., Charles-Edwards, E., Ueffing, P., Stillwell, J., Kupiszewski, M., & Kupiszewska, D. (2015). Internal migration and development: Comparing migration intensities around the world. Population and Development Review, 41(1), 33–58. https://doi.org/10.1111/j.1728-4457.2015.00025.x

Bilecen, B., & Lubbers, M. J. (2021). The networked character of migration and transnationalism. Global Networks, 21(4), 837–852. https://doi.org/10.1111/glob.12317

Cao, Z., Zheng, X., Liu, Y., Li, Y., & Chen, Y. (2018). Exploring the changing patterns of China's migration and its determinants using census data of 2000 and 2010. Habitat International, 82, 72–82. https://doi.org/10.1016/j.habitatint.2018.09.006

Champion, T. (2001). Urbanisation, suburbanisation, counterurbanisation and reurbanisation. In R. Paddison (Ed.), Handbook of Urban Studies (pp. 143–161). Sage. https://sk.sagepub.com/hnbk/edvol/hdbk_urban/chpt/urbanization-suburbanization-counterurbanization

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368), 829–836. https://doi.org/10.1080/01621459.1979.10481038

Danchev, V., & Porter, M. A. (2018). Neither global nor local: Heterogeneous connectivity in spatial network structures of world migration. Social Networks, 53, 4–19. https://doi.org/10.1016/j.socnet.2017.06.003

Danchev, V., & Porter, M. A. (2021). Migration networks: Applications of network analysis to macroscale migration patterns. In M. McAuliffe (Ed.), Research Handbook on International Migration and Digital Technology (pp. 70–90). Edward Elgar Publishing. https://doi.org/10.4337/9781839100611.00014

Dijkstra, L., Florczyk, A. J., Freire, S., Kemper, T., Melchiorri, M., Pesaresi, M., & Schiavina, M. (2021). Applying the Degree of Urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. Journal of Urban Economics, 125, 103312. https://doi.org/10.1016/j.jue.2020.103312

Fan, C. C. (2005). Interprovincial migration, population redistribution, and regional development in China: 1990 and 2000 census comparisons. The Professional Geographer, 57(2), 295–311. https://doi.org/10.1111/j.0033-0124.2005.00479.x

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley. https://www.wiley.com/en-us/Geographically+Weighted+Regression:+The+Analysis+of+Spatially+Varying+Relationships-p-9780471496168

Fotheringham, A. S., & Li, Z. (2023). Measuring the unmeasurable: Models of geographical context. Annals of the American Association of Geographers, 113(10), 2269–2286. https://doi.org/10.1080/24694452.2023.2227690

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers, 107(6), 1247–1265. https://doi.org/10.1080/24694452.2017.1352480

Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. Social Networks, 1(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Fujita, M., Krugman, P., & Venables, A. J. (1999). The Spatial Economy: Cities, Regions, and International Trade. MIT Press. https://direct.mit.edu/books/monograph/2525

Greenwood, M. J. (1997). Internal migration in developed countries. In M. R. Rosenzweig & O. Stark (Eds.), Handbook of Population and Family Economics (Vol. 1B, pp. 647–720). Elsevier. https://doi.org/10.1016/S1574-003X(97)80004-9

Gürsoy, F., & Badur, B. (2022). Investigating internal migration with network analysis and latent space representations: An application to Turkey. Social Network Analysis and Mining, 12, 150. https://doi.org/10.1007/s13278-022-00974-w

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 108, 2907–2916. https://proceedings.mlr.press/v108/janzing20a.html

Joo, Y.-M. (2018). Megacity Seoul: Urbanization and the Development of Modern South Korea. Routledge. https://doi.org/10.4324/9781315278018

Khavarian-Garmsir, A. R. (2023). A systematic review of shrinking cities literature: Lessons from the past and directions for the future. International Planning Studies, 28, 219–238. https://doi.org/10.1080/13563475.2023.2205030

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR 119, 5491–5500. https://proceedings.mlr.press/v119/kumar20e.html

Li, S., Gu, H., & Shen, J. (2023). Detecting spatial heterogeneity in the determinants of intercity migration in China. Population, Space and Place, 29, e2649. https://doi.org/10.1002/psp.2649

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. Computers, Environment and Urban Systems, 96, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Li, Z. (2024). GeoShapley: A game theory approach to measuring spatial effects in machine learning models. Annals of the American Association of Geographers, 114(7), 1365–1385. https://doi.org/10.1080/24694452.2024.2350982

Li, Z. (2025). Explainable AI in spatial analysis. In Handbook of Geospatial Artificial Intelligence. Springer. https://doi.org/10.1007/978-3-031-87421-5_5

Lou, X., Luo, P., Li, Z., Gao, S., & Meng, L. (2025). GeoXCP: Uncertainty quantification of spatial explanations in explainable AI. International Journal of Geographical Information Science, 39(9), 1866–1903. https://doi.org/10.1080/13658816.2025.2574900

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 4765–4774). https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Ma, S., & Tourani, R. (2020). Predictive and causal implications of using Shapley value for model interpretation. Proceedings of the 2020 KDD Workshop on Causal Discovery, PMLR 127, 23–38. https://proceedings.mlr.press/v127/ma20a.html

Mabogunje, A. L. (1970). Systems approach to a theory of rural–urban migration. Geographical Analysis, 2(1), 1–18. https://doi.org/10.1111/j.1538-4632.1970.tb00140.x

Mallach, A., Haase, A., & Hattori, K. (2017). The shrinking city in comparative perspective: Contrasting dynamics and responses to urban shrinkage. Cities, 69, 102–108. https://doi.org/10.1016/j.cities.2016.09.008

Martinez-Fernandez, C., Audirac, I., Fol, S., & Cunningham-Sabot, E. (2012). Shrinking cities: Urban challenges of globalization. International Journal of Urban and Regional Research, 36(2), 213–225. https://doi.org/10.1111/j.1468-2427.2011.01092.x

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1993). Theories of international migration: A review and appraisal. Population and Development Review, 19(3), 431–466. https://doi.org/10.2307/2938462

Mills, E. S., & Song, B.-N. (1979). Urbanization and Urban Problems: Studies in the Modernization of the Republic of Korea, 1945–1975. Harvard University Press.

Ming, Y., Liu, Y., Li, Y., & Song, Y. (2024). Unraveling nonlinear and spatial non-stationary effects of urban form on surface urban heat islands using explainable spatial machine learning. Computers, Environment and Urban Systems, 114, 102200. https://doi.org/10.1016/j.compenvurbsys.2024.102200

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book

OECD (2023). OECD Regional Outlook 2023. OECD Publishing. https://doi.org/10.1787/92cd40a0-en

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web (Technical Report No. 1999-66). Stanford InfoLab.

Pitoski, D., Lampoltshammer, T. J., & Parycek, P. (2021). Network analysis of internal migration in Croatia. Computational Social Networks, 8, 10. https://doi.org/10.1186/s40649-021-00093-0

Pu, Y., Zhao, X., Chi, G., Zhao, J., & Kong, F. (2019). A spatial dynamic panel approach to modelling the space-time dynamics of interprovincial migration flows in China. Demographic Research, 41, 913–948. https://doi.org/10.4054/DemRes.2019.41.31

Shams Amiri, S., Mottahedi, S., Lee, E. E., & Hoque, S. (2021). Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption. Computers, Environment and Urban Systems, 88, 101647. https://doi.org/10.1016/j.compenvurbsys.2021.101647

Simini, F., Barlacchi, G., Luca, M., & Pappalardo, L. (2021). A Deep Gravity model for mobility flows generation. Nature Communications, 12, 6576. https://doi.org/10.1038/s41467-021-26752-4

Statistics Korea (2024). Resident registration population and internal migration statistics. Korean Statistical Information Service (KOSIS). https://kosis.kr

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wang, N., Zhang, H., Dahal, A., Cheng, W., Zhao, M., & Lombardo, L. (2024). On the use of explainable AI for susceptibility modeling: Examining the spatial pattern of SHAP values. Geoscience Frontiers, 15(4), 101800. https://doi.org/10.1016/j.gsf.2024.101800

Wu, S., & Liu, T. (2022). Stability and change in China's geography of intercity migration: A network analysis. Population, Space and Place, 28(4), e2570. https://doi.org/10.1002/psp.2570

Yu, Z., Zhang, H., Tao, Z., & Liang, J. (2019). Amenities, economic opportunities and patterns of migration at the city level in China. Asian and Pacific Migration Journal, 28(1), 3–27. https://doi.org/10.1177/0117196819832054

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx