

Supplementary Information

Machine Learning Forecasts Reveal a Concentration Paradox in UK International Student Mobility

Ruth Neville, Francisco Rowe, Emilio Zagheni

Contents

S1. Data sources, preprocessing, and coverage

S2. Forecasting models

S2.1 ARIMA benchmark

S2.2 Negative binomial gravity model

S2.3 XGBoost (primary model)

S3. Forecasting strategy, evaluation, and uncertainty

S4. Extended performance results

S5. Validation of UCAS acceptances against public HESA data

S6. Reproducibility and code availability

15 **S1. Data sources, preprocessing, and coverage**

16 This study draws on administrative data from the Universities and Colleges Admissions Service
17 (UCAS), which provides the principal route through which students apply for undergraduate
18 degrees at UK higher education institutions. UCAS data capture approximately 95% of entrants
19 from the European Union (EU) and around 60% from non-EU countries, making them a robust
20 representation of undergraduate international student mobility to the UK.

21 We focus on successful international undergraduate applications (acceptances) by country
22 of domicile for the period 2010–2024. The dependent variable is the annual count of successful
23 acceptances from each origin country.

24 To capture drivers of these flows, we augment the UCAS data with origin- and destination-
25 specific covariates widely used in the migration and international student mobility literature.
26 These include economic and demographic indicators (GDP per capita and population size),
27 structural relationships (common official language, historical colonial ties), and bilateral dis-
28 tance (great-circle distance to the UK). Covariates are obtained from the World Bank World
29 Development Indicators, the United Nations World Population Prospects (WPP, 2022 Revision),
30 and the CEPII Gravity Database.

31 **S1.1 Projected covariates and the information set at prediction time**

32 For both the forecast period (2025–2030) and the evaluation period (2019–2023), we rely on
33 externally produced projections rather than realised values for economic and demographic co-
34 variates. This ensures that forward-looking predictions are based on information that would
35 have been available at the time of forecasting, thereby avoiding information leakage.

36 Specifically, projections for GDP per capita and population are taken from the most recent
37 World Bank and WPP releases available prior to the forecast year. For example, predictions for
38 2019 use projection series released before 2019; predictions for 2024 use series released in 2019.
39 For the final forecast horizon (2025–2030) we use projections released in 2024.

40 This design mirrors realistic decision-making conditions, where policy makers must rely on
41 macroeconomic and demographic forecasts rather than their eventual realisation. It also necessi-
42 tates the omission of some known drivers of international student mobility—such as institution-
43 specific rankings, subject-level composition, and detailed measures of diaspora size—for which
44 no globally consistent prospective data are available.

45 **S1.2 Preprocessing and feature engineering**

46 Several preprocessing steps and feature engineering operations are applied prior to model esti-
47 mation:

- 48 • **Standardisation of covariates.** Continuous covariates (GDP per capita and population
49 size for origin and destination, and distance) are standardised to have mean zero and
50 unit variance. This facilitates model convergence and comparability of feature scales,
51 particularly for the XGBoost model.
- 52 • **Lagged dependent variable.** We construct a lagged application feature, $y_{i,t-1}$, repre-
53 senting the number of successful applications from origin country i in the previous year.

54 This captures temporal persistence in application flows.

- 55 • **Recursive extension to 2030.** For the forecasting period (2025–2030), the model uses
56 projected covariates and, for the lagged response, its own predicted values. Forecasts for
57 year t thus depend on forecasts from year $t - 1$, reflecting the recursive nature of the
58 forecasting process.
- 59 • **Country inclusion criteria.** To reduce the influence of highly stochastic series, we
60 exclude origin countries with fewer than 10 total successful applications over the full 2010–
61 2024 period. The final dataset comprises 86 origin countries.

62 A summary of variables, definitions, temporal coverage, and data sources is presented in
63 Supplementary Table 1.

Table 1: **Supplementary Table 1 — Covariates, definitions, and data sources.**

Variable	Definition	Source	Temporal coverage
Applications	Annual count of successful undergraduate acceptances from origin country to the UK	UCAS	2010–2024
GDP per capita (origin)	GDP per capita (constant prices), origin country	World Bank	2010–2030 (observed + projected)
GDP per capita (destination)	GDP per capita, UK	World Bank	2010–2030 (observed + projected)
Population (origin)	Total population, origin country	UN WPP	2010–2030 (observed + projected)
Population (destination)	Total population, UK	UN WPP	2010–2030 (observed + projected)
Distance	Great-circle distance between capital city of origin and London	CEPII	Time-invariant
Common language	Indicator for shared official language (English) between origin and UK	CEPII	Time-invariant
Colonial tie	Indicator for historical colonial link with the UK	CEPII	Time-invariant
Lagged applications	Successful applications from origin country in the previous year	Constructed	2011–2030
Year	Calendar year	Constructed	2010–2030

64 S2. Forecasting models

65 We compare three forecasting approaches: a time-series benchmark (ARIMA), a theory-aligned
66 regression model (negative binomial gravity), and a machine learning model (XGBoost). XG-
67 Boost is the primary model used to generate forward-looking forecasts.

68 S2.1 ARIMA benchmark

69 We estimate autoregressive integrated moving average (ARIMA) models separately for each
70 origin country. ARIMA models extrapolate historical trends in application counts without
71 incorporating external covariates and are widely used in institutional planning and migration
72 forecasting.

73 Let y_t denote the annual number of successful applications from a given origin country. The
74 ARIMA(p, d, q) model can be written as:

$$\hat{y}_t = \mu + \sum_{i=1}^p \phi_i y_{t-i} - \sum_{j=1}^q \theta_j e_{t-j} + e_t, \quad (1)$$

75 where μ is a constant term, ϕ_i are autoregressive (AR) coefficients, θ_j are moving average (MA)
76 coefficients, and e_t are residual errors. Differencing of order d is applied as required to achieve
77 stationarity.

78 Country-specific orders (p, d, q) are selected using information-criterion-based optimisation
79 (AICc). We use the `auto.arima()` function from the `forecast` package in R to identify the
80 best-fitting model for each origin. Across origin countries, selected specifications predominantly
81 feature first-order differencing and low-order AR and/or MA components, consistent with rela-
82 tively short annual series. As ARIMA serves as a pragmatic time-series benchmark rather than a
83 structural model of international student mobility, we do not tabulate all country-specific orders
84 for brevity.

85 S2.2 Negative binomial gravity model

86 To provide a theory-aligned comparator that incorporates structural, economic, and demo-
87 graphic characteristics, we specify a gravity-inspired negative binomial model for count out-
88 comes.

89 Let y_{it} denote the number of successful applications from origin country i in year t . We
90 model the conditional mean as:

$$\mathbb{E}(y_{it} | \cdot) = \lambda_{it} = \exp(\alpha + \alpha_i + \beta_1 d_i + \beta_2 s_i + \beta_3 p_{it} + \beta_4 p_{jt} + \rho y_{i,t-1}), \quad (2)$$

91 where:

- 92 • α is a global intercept;
- 93 • α_i is a country-specific random intercept capturing unobserved heterogeneity;
- 94 • d_i is the log of great-circle distance between origin i and the UK;
- 95 • s_i is a vector of structural indicators (common language, colonial tie);

- 96 • p_{it} is a vector of time-varying origin-country characteristics (e.g., GDP per capita, popu-
97 lation);
- 98 • p_{jt} is a vector of time-varying destination-country characteristics (e.g., UK GDP per capita,
99 population);
- 100 • $y_{i,t-1}$ is the lagged dependent variable capturing temporal dependence, with coefficient ρ .

101 We assume a negative binomial distribution for y_{it} with mean λ_{it} and overdispersion param-
102 eter κ , allowing variance to exceed the mean, as is typical in migration and application count
103 data. The model is estimated using the `glmTMB` package in R, with a log link and random
104 intercepts at the country level.

105 Parameter estimates are directionally consistent with prior work on migration and interna-
106 tional student mobility: GDP per capita and population sizes are positively associated with
107 application volumes, distance is negatively associated, and the lagged dependent variable cap-
108 tures strong temporal persistence. As the gravity model is included as a benchmark rather than
109 for substantive interpretation of individual coefficients, we summarise these patterns narratively
110 and do not reproduce the full coefficient table here.

111 **S2.3 XGBoost (primary model)**

112 Our primary forecasting model is Extreme Gradient Boosting (XGBoost), implemented with a
113 Poisson objective to reflect the count nature of the outcome and the log-link structure used in
114 the gravity model.

115 Let y_i denote the number of successful applications for observation i (a country–year pair),
116 and let x_i be the vector of covariates including lagged applications, economic indicators, demo-
117 graphic variables, structural characteristics, and the calendar year. XGBoost approximates the
118 log of the expected count via an additive ensemble of regression trees:

$$\log \hat{y}_i = f(x_i) = \sum_{k=1}^K f_k(x_i), \quad (3)$$

119 where each f_k is a decision tree and K is the number of boosting rounds. Under the Poisson
120 objective, the loss function is:

$$L(\theta) = \sum_{i=1}^n [\exp\{f(x_i)\} - y_i f(x_i)] + \lambda \sum_j \theta_j^2, \quad (4)$$

121 where θ denotes the set of tree parameters and λ is an L2 regularisation term controlling model
122 complexity. Predicted application counts are obtained as:

$$\hat{y}_i = \exp\{f(x_i)\}. \quad (5)$$

123 **Hyperparameter tuning**

124 Hyperparameters are selected using 10-fold cross-validation with early stopping on the training
125 period. We train candidate models on data from 2010–2018 and evaluate performance on 2019–

126 2023, varying the number of boosting rounds and allowing the cross-validation procedure to
 127 select the optimal number via early stopping. In practice, we maintain a maximum depth of 6
 128 and learning rate of 0.3, with the number of boosting rounds typically around 150–200, trading
 129 off flexibility against overfitting.

130 The main hyperparameter configuration is summarised in Supplementary Table 2. This
 131 specification is then used for the validation exercise (training 2010–2023, predicting 2024) and
 132 for the final forecasting model (training 2010–2024, forecasting 2025–2030).

Table 2: **Supplementary Table 2 — XGBoost hyperparameter configuration.**

Hyperparameter	Value	Notes
Objective	<code>count:poisson</code>	Poisson regression for modelling count outcomes
Max depth	6	Controls the maximum depth of each tree
Learning rate (<code>eta</code>)	0.3	Shrinks the contribution of each tree
Number of boosting rounds	~150–200	Selected via 10-fold cross-validation with early stopping
Subsample	1.0	No row subsampling used
Column subsampling	1.0	No feature subsampling used
L2 regularisation (<code>lambda</code>)	Via early stopping	Helps reduce overfitting

133 **S3. Forecasting strategy, evaluation, and uncertainty**

134 **S3.1 Train-test splits and recursive forecasting**

135 For model evaluation, all three forecasting approaches are trained on data from 2010–2018 and
136 evaluated on an out-of-sample period from 2019–2023. This period encompasses the implemen-
137 tation of Brexit and the COVID-19 pandemic, providing a stringent test of model performance
138 under structural volatility rather than smooth trend continuation.

139 After evaluation, the XGBoost model is retrained on the full observed period (2010–2024)
140 using the tuned hyperparameters and is used to generate forecasts for 2025–2030. Forecasts are
141 produced recursively: the forecast for 2025 uses observed lagged applications from 2024, while
142 forecasts for 2026–2030 use model-predicted values for the lagged dependent variable.

143 **S3.2 Evaluation metrics**

144 We evaluate predictive accuracy using four standard measures of forecast error and one measure
145 of association between predicted and realised values.

Mean absolute error (MAE).

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (6)$$

Root mean squared error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (7)$$

Normalised mean absolute error (NMAE).

$$\text{NMAE} = \frac{\text{MAE}}{\bar{y}}, \quad (8)$$

146 where \bar{y} is the mean number of applications for a given country. NMAE enables relative com-
147 parison of predictive performance across countries with different application volumes.

Mean absolute percentage error (MAPE).

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100. \quad (9)$$

148 MAPE provides an intuitive percentage measure of prediction error but can be inflated for small
149 denominators. For origin countries recording fewer than approximately 50 applications annually,
150 MAPE values should therefore be interpreted with caution.

151 **Pearson correlation.** We compute the Pearson correlation coefficient between predicted and
152 observed values to summarise the strength of linear association across the evaluation period.

153 Aggregate performance statistics pooled across origin countries and years 2019–2023 are
 154 summarised in Supplementary Table 3, with country-level out-of-sample predictions shown in
 155 Supplementary Figure 1.

Table 3: **Supplementary Table 3** — **Aggregate performance of benchmark and primary models, 2019–2023.**

Model	MAE	RMSE	MAPE (%)	Correlation
ARIMA	327	1065	88	0.87
Negative binomial gravity	282	818	77	0.93
XGBoost	255	713	81	0.94

156 S3.3 Bootstrap prediction intervals

157 To quantify uncertainty around the XGBoost forecasts, we construct 95% prediction intervals
 158 using a model-retraining bootstrap. For each forecast year and origin country:

- 159 1. We draw a bootstrap sample of the historical training data by resampling country–year
 160 observations with replacement.
- 161 2. We re-estimate the XGBoost model on this bootstrap sample using the same hyperparam-
 162 eters.
- 163 3. We generate a prediction for the focal forecast year (and, in the recursive setting, update
 164 the lagged dependent variable with the bootstrap prediction).

165 This process is repeated 250 times, yielding a distribution of forecast values for each country–
 166 year. The 2.5th and 97.5th percentiles of this distribution form the bounds of the 95% prediction
 167 interval. Because forecasts for later years depend on earlier predicted values, interval widths
 168 naturally widen with the forecast horizon, reflecting increasing uncertainty and the accumulation
 169 of structural volatility over time.

S4. Extended performance results

S4.1 Country-level out-of-sample predictions from the primary model (2019–2023)

Supplementary Figure 1 shows out-of-sample predictions from the XGBoost model for the 2019–2023 evaluation period. Coloured lines represent model-predicted counts, and black points represent observed values. The figure provides a country-level view of validation performance and highlights three characteristics of the forecasting environment:

1. **Scale effects:** alignment is closer in absolute terms for high-volume sending countries, where signal dominates stochastic variation, while proportional errors are comparatively larger for low-volume countries.
2. **Volatility and discontinuity:** temporary divergences align with periods of structural disruption, such as the COVID-19 pandemic and policy changes affecting visa or fee status.
3. **Heterogeneity in trend persistence:** some countries exhibit smooth trend continuation, whereas others display non-linear trajectories that are challenging for all forecasting approaches.

This figure complements the aggregate performance metrics presented in Section 3 of the main paper by illustrating how system characteristics—rather than model specification alone—shape predictive accuracy.

S4.2 Separate origin-level forecasts for Mainland China and Hong Kong

Mainland China and Hong Kong are treated as distinct origin systems throughout model estimation and forecasting. Given substantive differences in policy exposure and mobility dynamics between Mainland China and Hong Kong, we present separate origin-level forecasts to demonstrate distinct trajectories and uncertainty ranges in Supplementary Figure 2. Mainland China and Hong Kong are reported as separate domiciles in the UCAS data, and do experience different relationships between their student mobility to the UK. For example, students from Hong Kong do have a British National Overseas (BNO) visa arrangement, perhaps making it easier to study in the UK. Hong Kong’s colonial history with the UK is also likely to lower barriers to entry in terms of language, administration, and culture - as explored in Neville et al. (2024). However, the substantive text refers to aggregated patterns of Mainland China and Hong Kong.

Aggregated values shown in the main figures (e.g., Mainland China and Hong Kong combined) are constructed by aggregating origin-specific forecasts; uncertainty intervals are obtained by aggregating bootstrap predictions across origins (summing bootstrap draws within iteration and then taking quantiles), rather than through pooled or joint model estimation.

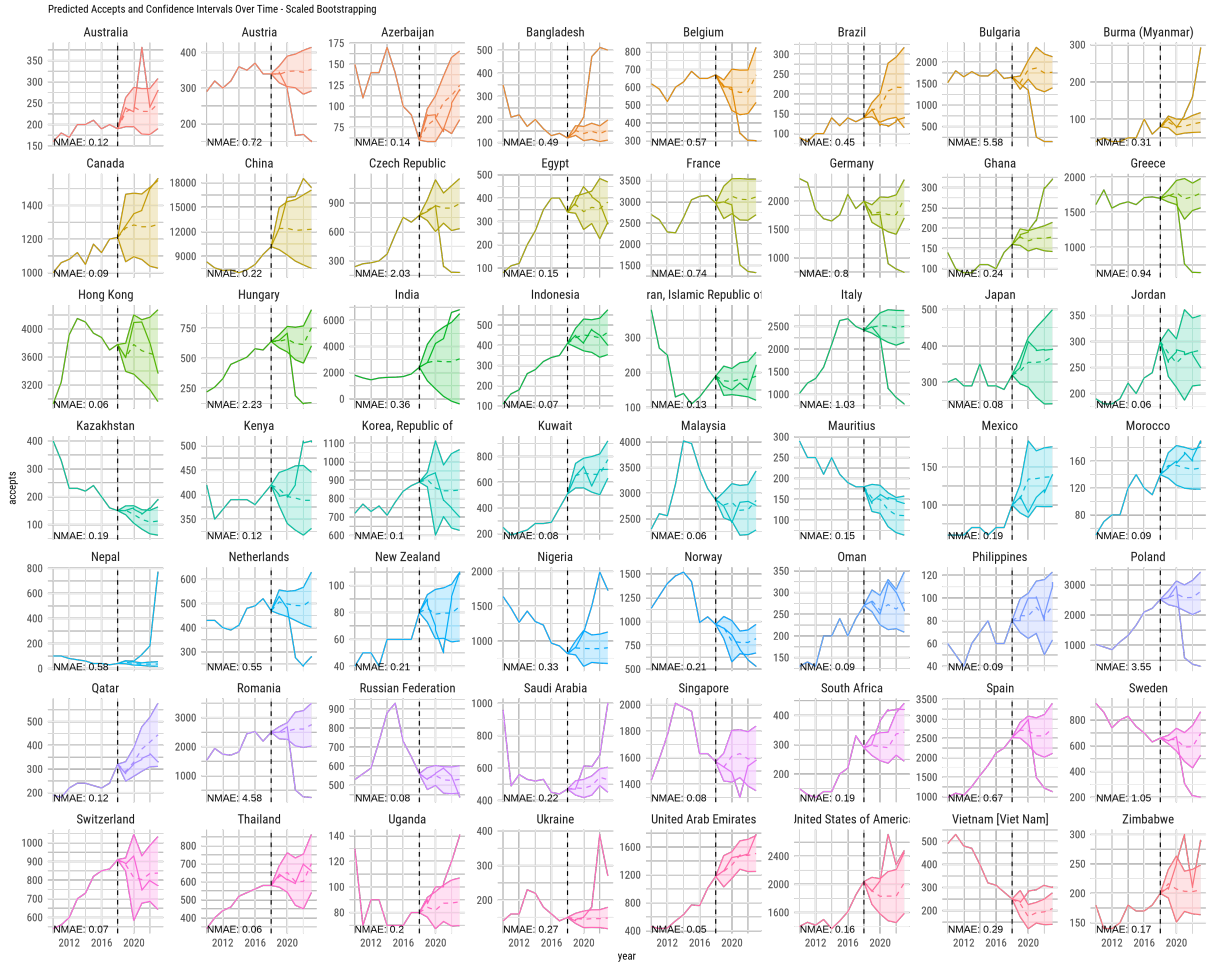


Figure 1: **Supplementary Figure 1 — Country-level out-of-sample predictions from the XGBoost model, 2019–2023.** Panels display predicted (dashed lines) and observed (solid line) application counts for 60 major origin countries and NMAE for each country. The figure illustrates heterogeneous predictive accuracy shaped by country size, volatility, and exposure to structural shocks.

202 S5. Validation of UCAS acceptances against public HESA data

203 To assess whether UCAS undergraduate acceptances provide a reasonable proxy for broader
 204 international undergraduate mobility patterns, we compared the UCAS series used in the main
 205 analysis with publicly available HESA undergraduate statistics. These checks are intended
 206 as validation exercises rather than as evidence of equivalence between the two data sources.
 207 UCAS records acceptances within the admissions system, whereas HESA records undergradu-
 208 ate entrants or enrolments, and the two sources differ in timing, coverage, and administrative
 209 definition.

210 S5.1 Aggregate entrant validation over time

211 For the main validation exercise, we used the public HESA Figure 9 extract filtered to entrant
 212 marker = Entrant, level of study = All undergraduate, mode of study = All, and country of HE
 213 provider = All. Because the public HESA entrant table is available only at aggregate domicile-
 214 group level, we compared UCAS and HESA for two groups: European Union and non-European

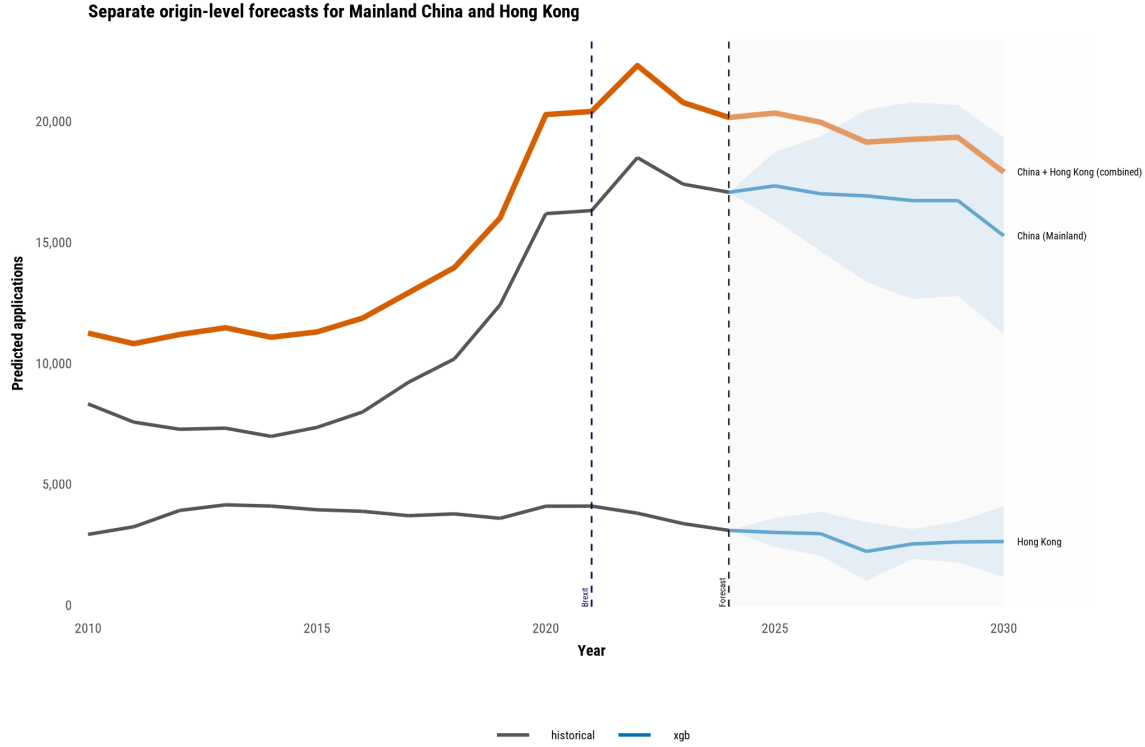


Figure 2: **Supplementary Figure 2 — Separate origin-level forecasts for Mainland China and Hong Kong.** Observed acceptances (2010–2024) are shown in grey; XGBoost forecasts for 2025–2030 are shown in blue with shaded 95% bootstrap prediction intervals. Mainland China and Hong Kong are modelled independently at the origin–year level. The orange line shows the combined China and Hong Kong series, constructed by aggregating origin-level forecasts, for comparison.

215 Union.

216 For each group, we aggregated the UCAS series to the corresponding category and calculated
 217 the Pearson correlation between annual UCAS acceptances (2020–2024) and HESA undergrad-
 218 uate entrant counts for the overlapping academic years (2020/21–2024/25). This correlation
 219 measures similarity in temporal movement over time rather than equality in absolute levels.

220 The resulting correlations are reported in Supplementary Table 4. Correspondence was very
 221 strong for both groups, indicating close temporal co-movement between UCAS acceptances and
 222 HESA undergraduate entrants at aggregate level for European Union students ($r = 0.998$) and
 223 Non-European Union students ($r = 0.975$).

Table 4: **Supplementary Table 4 — Aggregate and composition-based validation of UCAS against public HESA undergraduate data.**

Comparison	Years / units	Pearson r
EU: UCAS acceptances vs HESA undergraduate entrants	2020–2024	0.998
Non-EU: UCAS acceptances vs HESA undergraduate entrants	2020–2024	0.975

Notes: Aggregate entrant comparisons use public HESA Figure 9 undergraduate entrant totals. These checks are interpreted as validation exercises rather than like-for-like equivalents of UCAS acceptances.

224 **S5.2 Comparison with HESA undergraduate enrolment counts**

225 To provide additional evidence beyond aggregate categories and a single-year composition check,
 226 we used the yearly public HESA country-of-domicile files: Table 28, "Non-UK domiciled HE
 227 students by HE provider and country of domicile, 2014/15 onward". These files were filtered to
 228 undergraduate study only (First degree and Other undergraduate), aggregated across providers,
 229 and aligned to the corresponding UCAS annual acceptances series for the principal Non-EU
 230 origins highlighted in the paper.

231 We calculated Pearson correlations between the UCAS and HESA series over the overlapping
 232 period 2014–2024 for each origin separately. These results, reported in Supplementary Table 5,
 233 indicate strong temporal correspondence for most of the major Non-EU origin systems: China
 234 ($r = 0.973$), India ($r = 0.984$), Malaysia ($r = 0.860$), Nigeria ($r = 0.925$), Singapore ($r = 0.812$),
 235 the United Arab Emirates ($r = 0.989$), and the United States ($r = 0.886$). Correspondence was
 236 weaker for Hong Kong ($r = 0.670$), consistent with the fact that UCAS acceptances and HESA
 237 undergraduate enrolments are related but non-equivalent measures and may differ more for
 238 some origins than for others. These correlations indicate that UCAS and HESA series exhibit
 239 consistent co-movement over time across the principal origin countries, despite differences in
 240 absolute levels arising from coverage and definition.

241 We also examined the average gap between the two sources over time. Across all origins,
 242 HESA undergraduate enrolment counts are consistently higher than UCAS acceptances, reflect-
 243 ing differences in coverage and the fact that HESA captures all enrolled students (including
 244 non-UCAS routes and continuing cohorts), whereas UCAS captures admissions flows within a
 245 single application cycle. These gap statistics therefore help contextualise the level differences
 246 between the datasets, while the correlations summarise the extent to which they move together
 247 over time.

Table 5: **Supplementary Table 5 — Origin-specific temporal correspondence between UCAS acceptances and public HESA undergraduate enrolments, 2014–2024.**

Origin	Pearson r
China	0.973
Hong Kong	0.670
India	0.984
Malaysia	0.860
Nigeria	0.925
Singapore	0.812
United Arab Emirates	0.989
United States	0.886

Notes: Correlations are calculated between annual UCAS acceptances and yearly public HESA country-of-domicile undergraduate enrolment counts aggregated across providers for the overlapping period 2014–2024. These data are drawn from HESA Table 28, "Non-UK domiciled HE students by HE provider and country of domicile, 2014/15 onward". HESA values are not entrant-specific in this table and should therefore be interpreted as a validation check rather than a direct equivalent of UCAS acceptances.

248 **S5.3 Comparison with HESA all-student enrolment counts**

249 As an additional robustness check, we compared annual UCAS undergraduate acceptances with
 250 all-student enrolment counts from the same public HESA country-of-domicile table. This com-
 251 parison uses records aggregated across all levels and modes of study, including undergraduate
 252 and postgraduate students, full-time and part-time modes, and both new and continuing enrol-
 253 ments. It is therefore broader than, and not directly equivalent to, the UCAS undergraduate
 254 acceptances series, which captures undergraduate admissions flows.

255 The results show strong temporal alignment for most major origins, although the correspon-
 256 dence is weaker than in the undergraduate-only validation checks for some comparisons. Cor-
 257 relations between UCAS undergraduate acceptances and HESA all-student enrolment counts
 258 were high for China ($r = 0.957$), India ($r = 0.991$), Malaysia ($r = 0.940$), Nigeria ($r = 0.951$),
 259 Singapore ($r = 0.807$), the United Arab Emirates ($r = 0.970$), and the United States of America
 260 ($r = 0.953$). Hong Kong was the main exception, with a weaker correlation ($r = 0.366$). The
 261 selected-origin composition correlation for 2024 was also lower for the all-student HESA com-
 262 parison ($r = 0.787$) than for the undergraduate-only HESA comparison ($r = 0.978$), consistent
 263 with the broader and less directly comparable nature of the all-student HESA measure.

Table 6: **Supplementary Table 6 — UCAS–HESA validation using HESA all-student enrolment counts.**

Origin	Pearson r
China	0.957
Hong Kong	0.366
India	0.991
Malaysia	0.940
Nigeria	0.951
Singapore	0.807
United Arab Emirates	0.970
United States of America	0.953
Selected-origin composition, 2024	
UCAS vs HESA all students	0.787

Notes: Annual origin-level correlations compare UCAS undergraduate acceptances with all-student enrolment counts by country of permanent address from HESA Table 28, "Non-UK domiciled HE students by HE provider and country of domicile, 2014/15 onward", for 2014–2024. HESA all-student counts are aggregated across undergraduate and postgraduate levels, full-time and part-time modes, and both new and continuing enrolments. The 2024 selected-origin composition correlation compares the cross-sectional distribution of students across selected origins. The undergraduate-only composition result is discussed in the text for comparison. These comparisons are interpreted as sensitivity checks rather than direct equivalence tests, since UCAS acceptances capture undergraduate admissions flows whereas HESA all-student counts capture broader enrolment stocks.

264 **S5.4 Interpretation and limitations**

265 Taken together, these checks support the use of UCAS acceptances as a reasonable adminis-
 266 trative indicator of broad patterns in undergraduate international student mobility. Aggregate
 267 validation shows that UCAS tracks public HESA undergraduate entrant dynamics closely for
 268 EU and non-EU groups, while the country-composition and origin-specific analyses indicate

269 similar correspondence for the major non-EU origins emphasised in the paper. The additional
270 all-student HESA comparison further shows that broad temporal alignment remains strong for
271 most major origins even when compared with wider enrolment-stock measures, although the
272 lower composition correlation confirms that all-student HESA counts are less directly compara-
273 ble to UCAS undergraduate acceptances than undergraduate-specific HESA measures.

274 However, these checks do not remove the conceptual distinction between admissions out-
275 comes and realised enrolments. UCAS acceptances should therefore be interpreted as confirmed
276 undergraduate admissions placements within the UCAS system rather than exact measures of
277 final enrolment.

278 Several limitations should be noted. First, UCAS records acceptances, whereas HESA
279 records entrants or enrolments. Second, HESA public data are reported by academic year,
280 whereas the UCAS analytical panel is organised by calendar year. Third, public HESA entrant
281 data are available only at broad domicile-group level rather than as a full country-year panel.
282 Fourth, the DT051 country tables are undergraduate-only but not entrant-specific.

283 Hong Kong is a clear exception in the origin-specific comparisons. One plausible explana-
284 tion is that UCAS records applicant domicile at the admissions stage, whereas HESA's public
285 country-of-domicile table reports students by non-UK permanent address. Following the in-
286 troduction of the British National Overseas route, some Hong Kong-origin students may have
287 reported, or later come to report, a UK permanent address by the time of enrolment or continued
288 study, reducing their visibility in HESA's Hong Kong non-UK permanent-address counts.

289 **S6. Reproducibility and code availability**

290 All modelling was conducted in R (version 4.3) using the packages `xgboost`, `forecast`, `glmmTMB`,
291 `dplyr`, and `ggplot2`. Code for data preprocessing, model estimation, evaluation, and figure
292 generation will be made openly available via the GitHub upon publication. Documentation will
293 be provided to allow replication of the main analyses and extension to additional origin countries
294 or alternative forecasting horizons.

295 **References**

296 Neville, R., Rowe, F., and Singleton, A. (2024). Nonlinear relationships between human devel-
297 opment and international student mobility: The prominent role of employment prospects and
298 cultural and linguistic ties. *Population, Space and Place*, 30(5).