uc3m | Universidad Carlos III de Madrid

Master's degree in Computational Social Science
2022-2023

Master Thesis

"Accounting for education enhancement effects on population growth dynamics: A wealth gap perspective"

Luis Ignacio Pulido Ruiz

Regina Kaiser

Madrid, 2023

AVOID PLAGIARISM

The University uses the Turnitin Feedback Studio program within the Aula Global for the delivery of student work. This program compares the originality of the work delivered by each student with millions of electronic resources and detects those parts of the text that are copied and pasted. Plagiarizing in a TFM is considered a **Serious Misconduct**, and may result in permanent expulsion from the University.

Abstract

This paper presents an exploration of the relationship between education and population growth rate using machine learning algorithms which, by taking a different perspective from the commonly used time series analysis, aims to shed light on the role of education in understanding demographic dynamics. The research leverages advanced regression techniques with a focus on regularization to test a set of hypotheses concerning gender disparities in education within different types of economies. Our findings reveal that encouraging women to pursue tertiary education holds a more robust marginal effect on population growth rate than simply increasing the overall proportion of the population with educational attainment. One observes how incorporating various education attributes, such as quality, years of schooling, and the proportion of women in universities, yields a remarkable accuracy of up to 88% in predicting countries prone to encountering negative population growth rates. Namely, our analysis indicates a probability of close to 80% for European regions like Spain to exhibit a negative growth rate. Furthermore, we investigate the effectiveness of incorporating a time series evaluation, revealing a modest improvement through the autoregressive component, hence this study emphasizes the role of education as a determinant predictor in population dynamics, showcasing its potential as a viable alternative to traditional time series analysis.

All assessments and modelling were conducted in the R programming language, and the accompanying code is provided as a supplementary resource to this work. Through this research, we contribute to the understanding of the intricate relationship between education and population fluctuations, offering insights that could inform policy interventions and strategic planning for sustainable development.

Keywords

Education; Population growth rate; Machine learning; Regularization; Gender Disparities; Time series analysis; Predictive modelling; Logistic regression; R programming.

Content

Introduction

Empirical evidence suggests that MED (More Economically Developed) countries are undergoing a process of stagnation in terms of population dynamics whereas pre-industrialized economies experience lingering peaks in the number of inhabitants. Economies like Japan, Italy or Spain are currently portraying negative population rates which diverge significantly from the average replacement rate required to sustain population from one generation to another one (2.1 children per woman), and World Health Organization projections point at the shrinkage of up to 10% in Europe in contrast with the African boost of over 50% by 2050.

The uneven nature of population growth rate emerges as the main argument calling for a dual analysis of the latter (e.g., splitting our task into a continental dimension analysis). This spans from the significant decline in demographic expansion present in many European economies to the existing abysmal gap with African countries such that, according to the Wittgenstein Centre for Demography, Nigeria is approaching a fertility rate of 6 children per woman. Therefore, the different objectives to tackle population issues directly stem from the disparities observed between countries in terms of culture, development and economic conditions. Dealing with a problem of overpopulation requires public policy intervention which, as an earlier step, needs a deep understanding of the treatment variables (factors one can have an influence on) which can shape the output as efficiently as possible (population growth). Take, for instance, the one-child policy introduced in China in the early 80s which directly limits the number of children allowed per woman. Such a policy, set aside any moral or individual freedom apology, has proven to lead to an ageing society similar to the one currently experiencing in European economies.

According to Population Council researcher Mark Montgomery, boundless drops in fertility rates were immediately observed when developing countries revised their educational performance in the 70s, portraying a significant correlation between these two-events worth studying. Such correlations were accounted with complex analytical and computational

techniques followed by a longitudinal approach (time series component) which allowed to observe this trend behaviour in many countries and at different points in time. Our departure point will be to reveal the education level effect on population dynamics distinguishing between different classes of economies and continents. Of course, one could claim that such a causal effect could be stained by other confounding biases introduced by cultural spill overs (marriage rate, religious attention etc.) or structural properties (household size, economic openness etc.), and we must devote a section of the paper to depict the causal paths and potential confounders. Assuming our target variable holds a null independence nature, current population movements would behave as a function of past observations, hence suggesting the inclusion of an autoregressive time series model in addition to features related to the current state of it (household size, median age etc.).

Motivation

Classical macroeconomic approaches often resort to already built frameworks in order to model the explanatory behaviour and evolution of unemployment, inflation or any other measure of one could care for. These models are nourished by and evaluated with econometric techniques based on empirical data such that their fundamental objective is to unfold and predict subsequent economic performance. Solowian economics claim for the appearance of country convergence conditional on a set of characteristics whereas the introduction of an endogenous Ramsey model involving the use optimal control theory predicts the possibility of a steady state for a given set of variables. These variables and characteristics common to most models involve the interaction of population dynamics as an argument of any function describing the factors driving long-term economic prosperity, but such component is seldom endowed with an endogenous characteristic, and it is usually announced as a constant.

Despite the inclusion of this parameter on modelling GDP per capita growth rate is often revealed as inversely proportional, the role of population growth rate holds an uncertain nature. Enhanced labour force or shifts in aggregate demand are both linked to the notion of population growth rate, thus impacting in a desirable way economic prosperity. On the other

hand, the denominator rise (per unit of population) when computing the progress of GDP per capita or the potentially excessive demand pushing up price indexes (inflation) might justify the opposite. Yet the issue is far from black and white. Expressions (1) and (2) depict the evolution of per capita GDP following Solow and Ramsey, respectively:

$$\frac{\dot{y}}{y} = \alpha\left[k^{\alpha-1} - \frac{c}{k} - (\delta + n)\right] \quad (\mathbf{1})$$

$$y = \left(\frac{s}{\delta + n + g}\right)^{\frac{\alpha}{1-\alpha}} \quad (\mathbf{2})$$

Solow and Ramsey Illustration of "n" Inclusion

where $\frac{\dot{y}}{y} : Per\ capita\ GDP\ Growth\ Rate$

$n : Population\ Growth\ Rate$

$\delta : Depreciation\ Rate$

Our proposal will hold that education enhancement is strongly linked with the notion of family planning, sex education and female labour participation which eventually gives rise to birth control and gender parity. The main objective revolves around establishing the empirical link between education and population growth rate, delving into the depths of this intricate relationship and measuring the extent to which the latter can serve as a predictive tool in analysing demographic dynamics through time series evaluation. Our opening hypothesis requires us to explore the most relevant components shaping the evolution of population dynamics outside of the confounding effects arising from migration. Both a cultural and economic perception calls for this analysis, locating the existing relationship between such variables with population rates and enabling the researcher to predict for future directions. The relevance of this study lies on the potential contribution to the broader topic of population dynamics and public policies which is one of the major threatening issues to many economies in terms of sustainable development and long-term prosperity.

Data Processing & Methodology

Research design

The first section of the paper will operate under the hypothesis that such an impact on population growth rate will differ across countries portraying different characteristics (civil rights, economic conditions, geography etc.) and this will require a cautious specification of the root model in order to compute marginal effects. Once the non-linear particularization is reached, we will test for the hypothesis of identical marginal effects by replacing in the original equation our restriction and evaluating the sum of squared residuals.

The second segment of the project will deal with the construction of a predictive (dichotomous) model to forecast instances in which a country experiences negative population growth rates in contrast with the opposite case. The results stemming from this section of the paper focus on MEDs in contrast to our previous part on developing economies, and we will compute the marginal effect over a set of variables on contributing to the probability of driving into the region of negative growth rates.

The final section of the thesis introduces a time series approach with population evolution as a target variable to test the hypothesis of the autoregressive component explained earlier. The logic is that, if such a variable is proven to behave as a function in which one of the stronger arguments is a lag of the same variable, then accuracy of predictions using this model would improve the ones from the predictive analysis based on our binary variable. Variable inclusion and its corresponding units are summarised in the following table:

Attribute	Class	Description	Source
Country Name	Character/String	The name of the country starting with upper case	World Bank
Year	Numeric/Date Discrete	Year of the observation	World Bank
Gdp growth rate	Numeric/Float Continuous	Yearly growth rate as a percentage	World Bank

Pop growth rate	Numeric/Float Continuous	Yearly growth rate as a percentage	World Bank
School years	Numeric/Float Continuous	Average number of years of schooling	Our World in Data
Education Expenditure	Numeric/Float Continuous	Expenditure on education as a percentage of GDP	Our World in Data
Continent	Categorical/factor	String indicating the country continent	Penn World Table
Life Expectancy	Numeric/Float Continuous	Number of years expected to live at current year	Our World in Data
Migration rate	Numeric/Float Continuous	Positive/Negative rate indicating the balance	World Bank
Health Expenditure	Numeric/Float Continuous	Public Healthcare expenditure per capita	World Bank
Median Age	Numeric/Float Continuous	Median population age at current year	World Bank
Marriage rate	Numeric/Float Continuous	Number of marriages per 1000 people	Our World in Data
Civil rights	Categorical/factor	1-7 scale (1 scoring the best)	Our World in Data
Religion degree	Numeric/Float Continuous	Percentage of population who thinks religion is important	The International Religious Demographic Project
Latitude	Numeric/Float Continuous	Location variable	github.com/albertyw/avenews
Longitude	Numeric/Float Continuous	Location variable	github.com/albertyw/avenews
Tertiary Education	Numeric/Float Continuous	Percentage of people between 25 and 30	Sustainable Development Report

		holding tertiary education	
Female to Male Ratio	Numeric/Float Continuous	Ratio of females to males in education	Sustainable Development Report

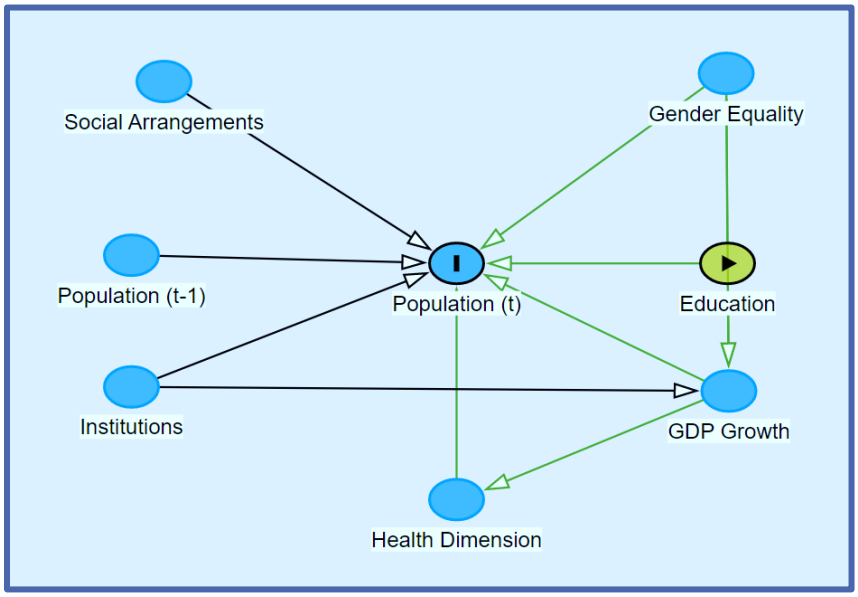
Table 1: Attribute Description, Own elaboration

Causal path and Bias Direction

The direction of causality and pre-process of proxy attributing and operationalization is outlined in plot 1 such that population dynamics is introduced as the outcome and further dimensions are proposed as future covariates. Following the reasoning of our introductory section, the inclusion of a specific variable is justified from the point of view of omitted variable bias and must be rigorously inspected from a qualitative lens. Such a bias will be effective when a given variable, holding a non-null correlation with our target outcome, is sufficiently linked to the incumbent explanatory attributes in the model. For instance, one might suggest that economies with enhanced levels of civil rights or negligible influence from religious authorities are expected to be less stringent on gender equality issues such as abortion laws or access to sex education, which leads to reduced fertility levels as a consequence of unconfined family planning.

There is a wide literature pointing to lower discrimination (on race, gender, religion, or any other characteristic) being highly correlated with having an equal chance to pursue education, job opportunities, and entrepreneurship. This accompanies a more diverse and skilled workforce, which can lead to strengthen productivity, foster innovation, and ultimately drive economic growth. Nevertheless, it is income inequality which is often revealed as one of the major concerns in the context of economic growth and population structure. Social exclusion based on lack of human rights can lead to income disparities and uneven access to resources. By promoting equal treatment and opportunity for all individuals, civil rights enable the

tackling of income inequality, which drives a region to a more stable and prosperous economy.



Plot 1: Directed Acyclic Graph, Own Elaboration

The implicit notion of bias in this example illustrates an underestimation of the actual coefficient parameter capturing the effect of economic growth rate on population dynamics since our model would compare developed regions with low fertility rates (enhanced by institutional factors such as civil rights and legal framework) against LEDCs with huge birth rates (influenced, in many cases, by the opposite). The formal representation of this is captured by expressions (3) and (4) below. If we consider the regression of y against x_1 and under the omission of x_2, then if

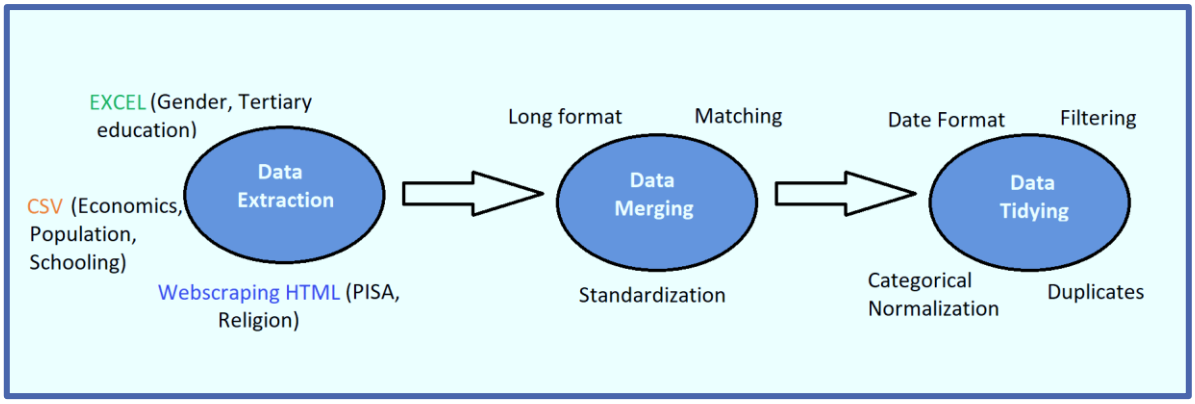
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 \rightarrow E(\widehat{\beta_1}) = \beta_1 + \hat{\beta}_2 \delta \quad (3)$$

Where:

$$\beta_2 : \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1 \quad (4)$$
$$\delta : \hat{x}_2 = \hat{\delta}_0 + \hat{\delta} x_1$$

Data Harvesting

We provide the user with a brief section devoted to detailing the construction of the final database, as it required a considerable investment of time and effort. Data from multiple sources, including the World Bank and the Sustainable Development Report, were integrated and processed into a unified data set which was laboriously homogenized and organized to fill later descriptive and imputation purposes. The web scraping techniques required for gathering information on PISA outcomes and religious attributes are delivered inside the supplement code attached to the project. The respective HTML was parsed, systematically scanned through the source code, and we identified the specific tags that corresponded to the PISA results for science exams supervised every four years. Once we had identified these tags, strings of information including numerical scores and descriptive traits were tokenized and stored for further analysis and interpretation.

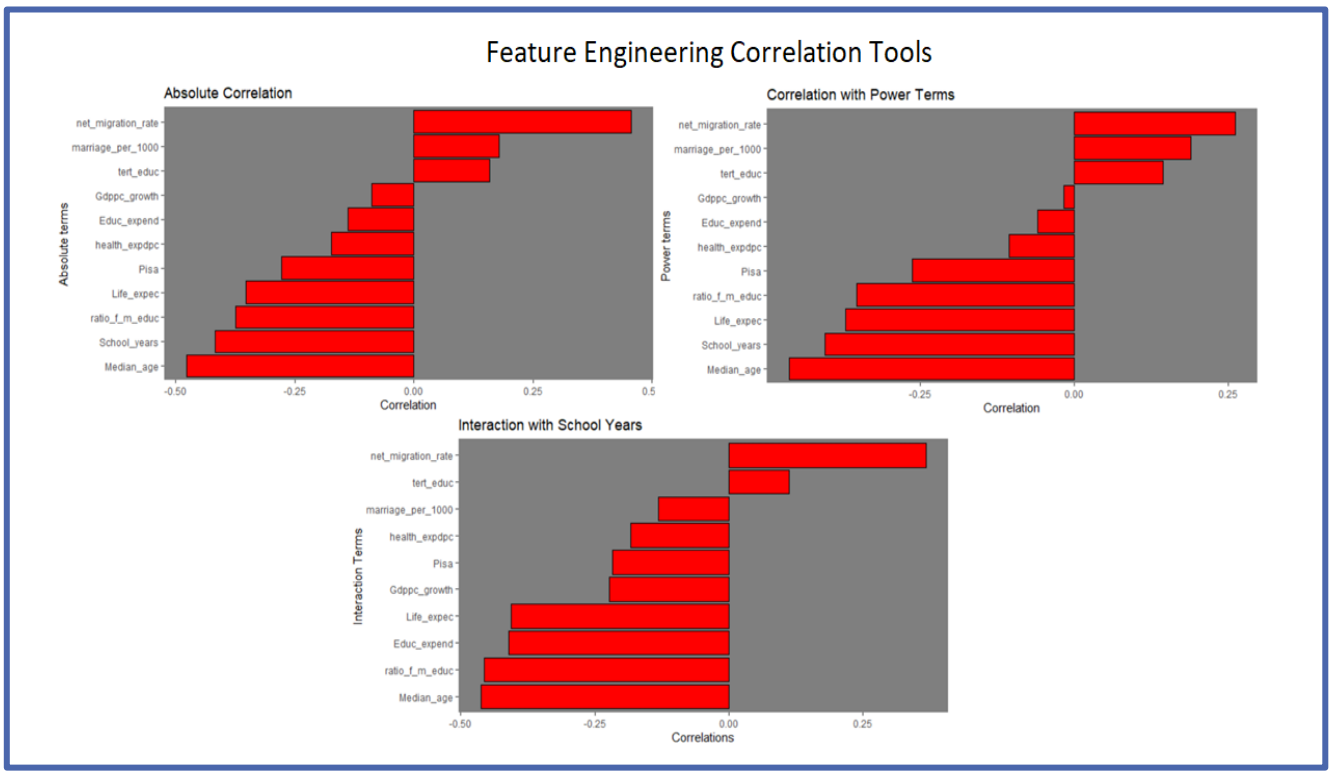


Plot 2: Data Harvesting Phases, Own Elaboration

Feature Engineering and Imputation

We have implemented a rigorous feature engineering process to identify the most relevant variables to be introduced in our predictive models. Such operation consists in plotting the correlation between each regressor and the population growth rate and examining the

association between interaction terms with school years and squared variables (plot 3). By comparing the magnitude of these correlations, we have determined that it is imperative to include a squared term for both life expectancy and marriage in combination with joint effects for education expenditure, life expectancy, and female to male ratio on tertiary education. Our study also calls for logarithmic features for many of our variables, which is a common technique in economics when accounting for growth rates and for the sake of avoiding heteroscedastic issues.

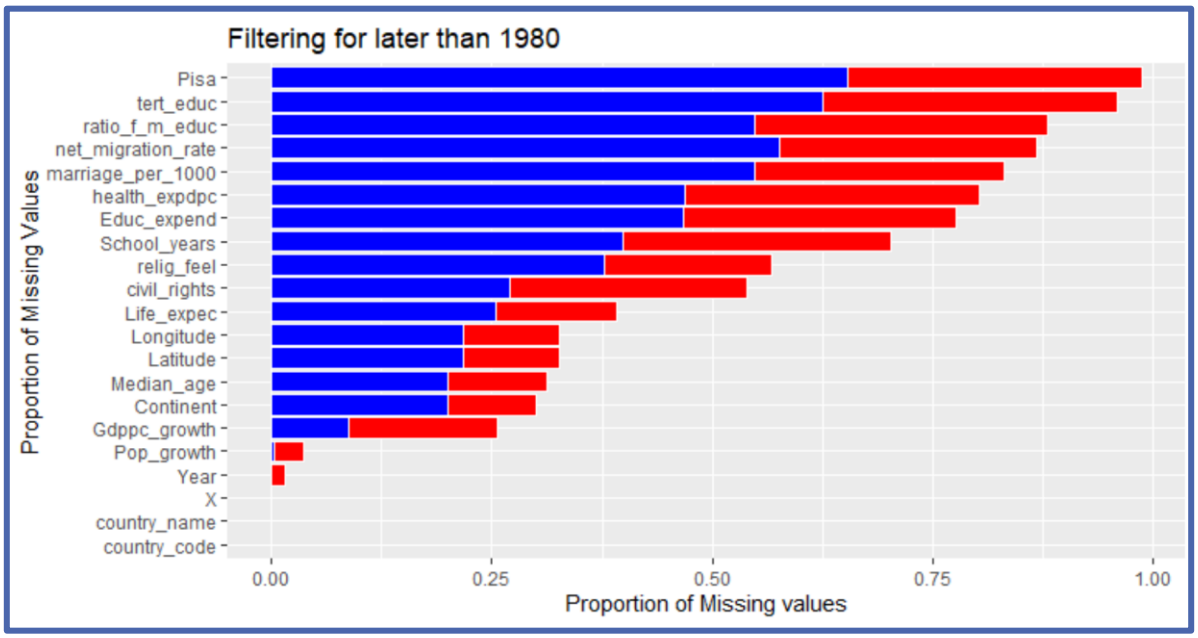


Plot 3: Feature Engineering Correlations, Own Elaboration

One of our main concerns at the data manipulation and encoding stage was the substantial number of missing values stemming from database merging, time measurement inconsistencies or absence of full public disclosure in some cases. Plot 4 identifies the plummeting rate of unavailable data subject to the filtering of observations pertaining to periods later than (and including) 1980. The red bars account for the proportion of missing values when considering the whole dataset whereas the blue ones measure this amount when we filter for data starting in 1980. This illustration is an advocate for the fundamental decision

of restricting the sample period to the one proposed, but complementary tools involving multiple imputation techniques will be applied to enable the piping of such data into the appropriate models.

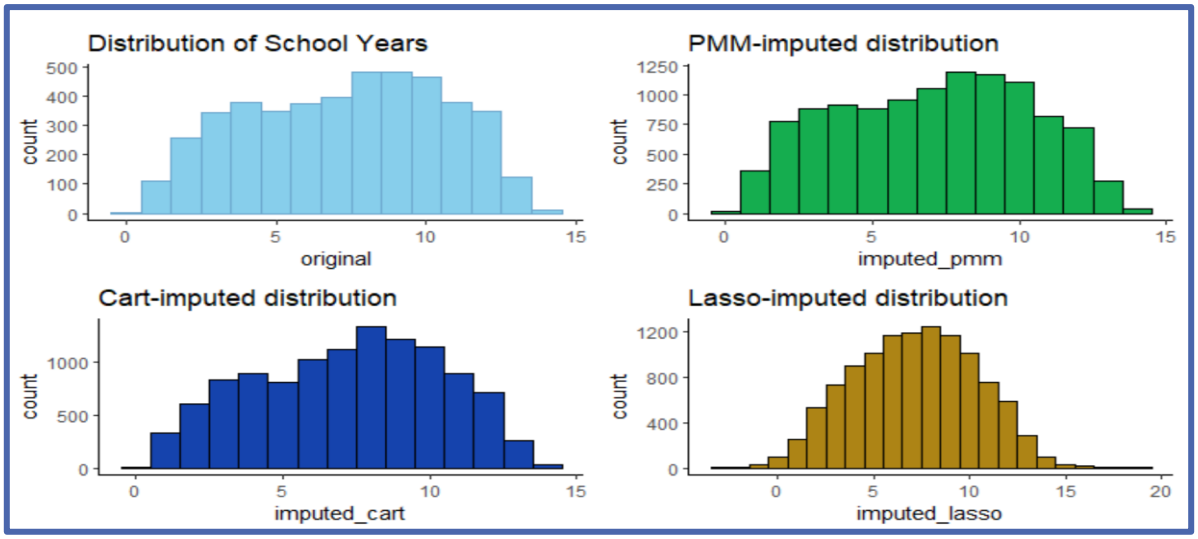
Precisely, we will resort to Multivariate Imputation by Chained Equations (v3.13.0; van Buuren & Groothuis-Oudshoorn, 2011) in R such that we will thoroughly inspect, for a range of approaches and for each of our covariates, the optimum one at fitting the data. Predictive Mean Matching (PMM) will regress the variable holding missing observations against the rest of attributes and compute the most plausible measurement to replace the empty field whereas classification and regression trees (CART) have decision nodes such that, based on certain splitting conditions or "questions", data points meeting such requirements are classified into one direction or another.



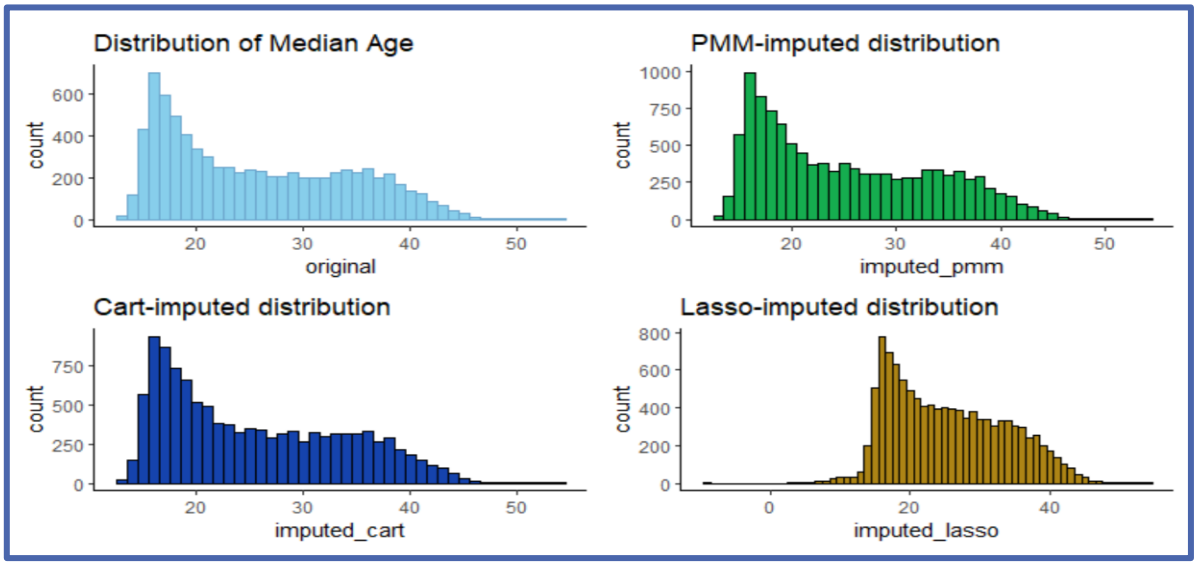
Plot 4: Proportion of Missing observations Pre and Post 1980s, Own Elaboration

We provide an example of the rationale disclosed above covered by plots 5 and 6 which depict the contrasting dispersion and behaviour obtained when applying multiple imputation to number of school years and median age. The choice of the best fitting imputation technique is a sound approach based on the visualization of the distribution before and after imputation

which enables the analyst to discriminate within algorithms and establish which technique preserves the original characteristics of the data. Their initial missing observations are automatically replaced by the predictive values harvested by our algorithms, and we determined to select the predictive matching and decision trees procedures for school years and median age, respectively, due to statistical similarities with the original frequency distribution.



Plot 5: Original vs Imputed School years, Own Elaboration



Plot 6: Original vs Imputed Median Age, Own Elaboration

Research Findings

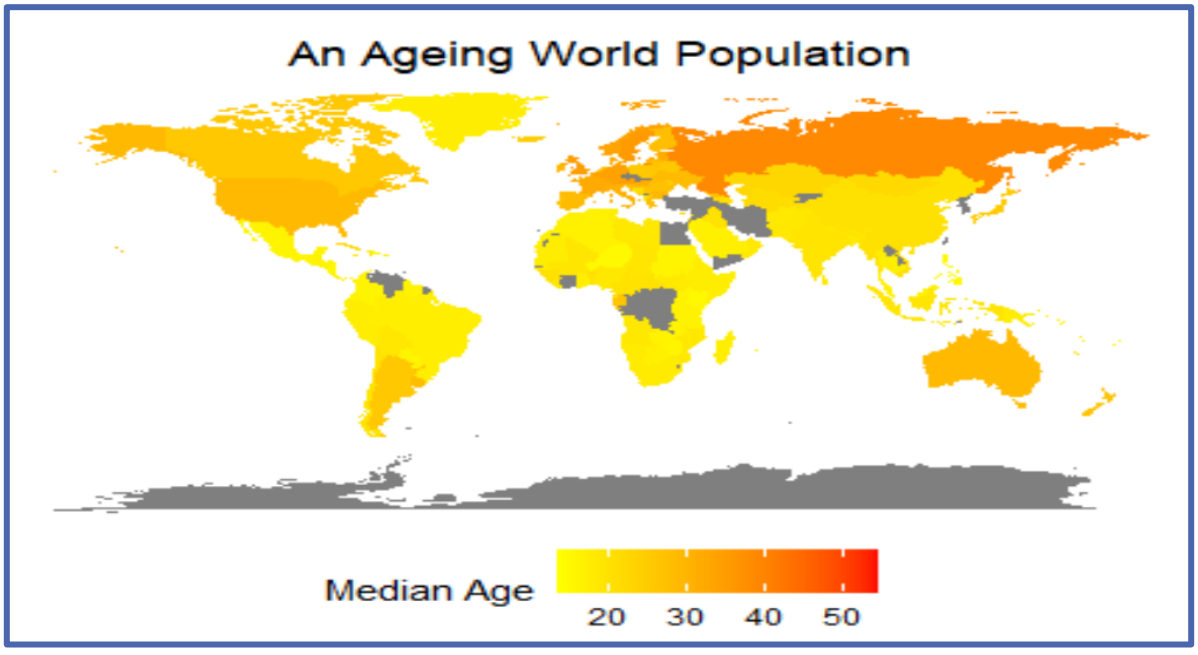
The results stemming from the linear model confirmed our opening hypothesis of the negative empirical relation between education development and population growth rate, exposing the number of years of schooling and proportion of females in education as the two key driving factors behind this phenomenon. The interaction effect between schooling years and continental dimension shows that there is a statistically significant difference in this marginal effect between Africa and Europe, being the latter the group exhibiting the greatest repercussions in terms of population dynamics stagnation during the last decades. On the other hand, we have observed a positive interrelationship between demographic change and health expenditure per capita which likewise holds for life expectancy. While healthcare spending appears to exert a more significant impact on population growth trends, the partial effect of school years reveals a more pronounced effect.

We rejected the null of the symmetrical marginal effect between having more women in tertiary education and the relevance of promoting the aggregate proportion of people tertiary education. Our R squared analysis revealed that the inclusion of women into education can be even more powerful in explaining population trends than the overall education analysis that we completed throughout the paper.

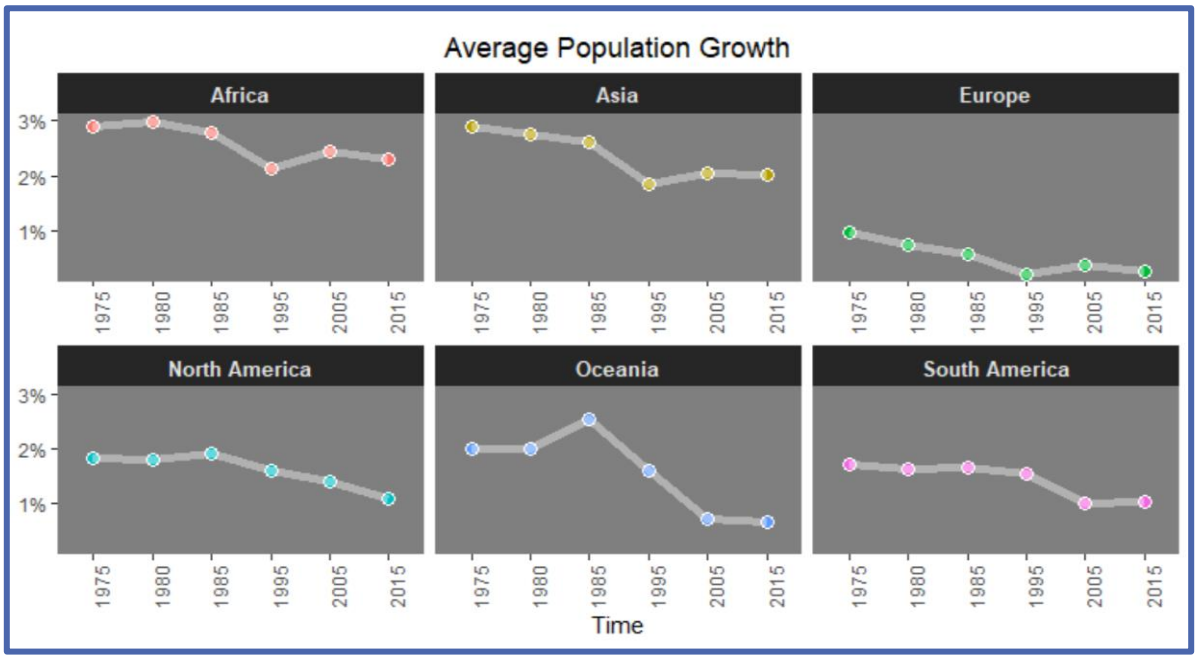
The dichotomous model is able to predict the behaviour of population with an AUC close to 0.9 and a sensitivity of 90% just by resorting to the education aggregates, and the concave marginal effect of education revealed Europe and Asia as the regions with more risk of entering into the dynamics of negative demographic rates. After controlling for comparable education attributes, we observed that a European economy has a 25% higher likelihood of experiencing population contraction compared to an Asian economy.

Descriptive Analysis

We provide an exploratory analysis of the underlying distributions and potential synergies found within the set of variables described in our work as an initial strategy to understand the behaviour of our data. This step will be key to establish the link between what one perceives from the data and the conclusions reached by our modelling strategies, and to assess the quality and consistency of such outcomes. The channel through which we will assemble our point descriptive interpretation with the further generalization requires the inclusion of hypothesis testing and inferential statistics applied to modelling techniques. This section has been constructed supported by a range of *ggplot* operations in R which the user can access as a supplement to this project together with all the code lines applied during the paper. We open this discussion by supplying the reader with a coloured world map snapshot in 2018 portraying regional differences in median age (plot 7) and the evolution of continental population growth rate (plot 8). One notices how Africa is systematically holding the most prominent levels of population increments while at the same time accounting for the less aged society. Many European economies, in combination with USA, Russia and Australia, envelop the top ranking of aged societies and, in particular, Europe shows a negative trend in population dynamics in spite of departing from already depressed conditions.



Plot 7: World Map by Median Age in 2018, Own Elaboration

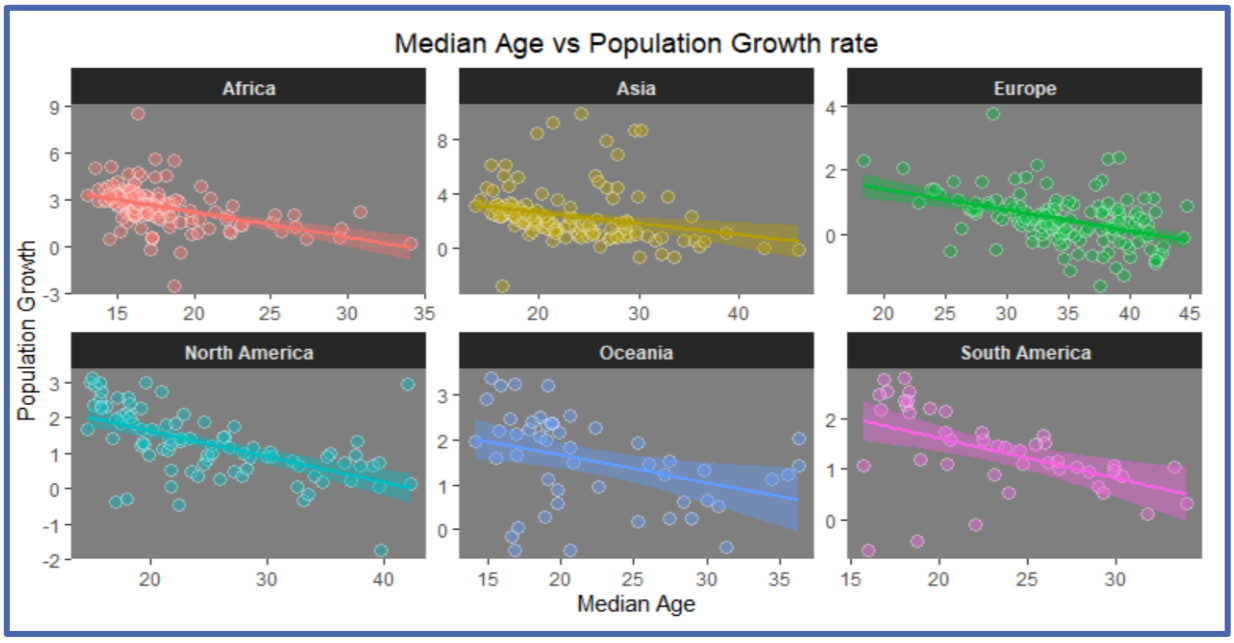


Plot 8: Population Growth by Continent, Own Elaboration

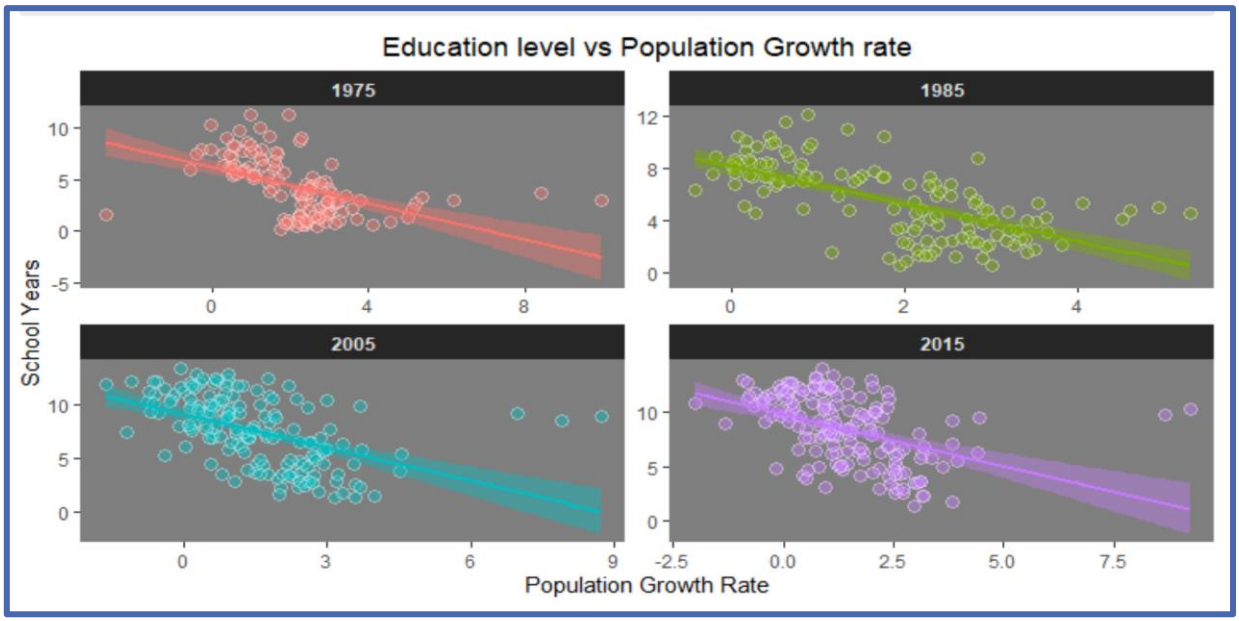
The following step would be to combine this information in a single graph (plot 9) to characterize the relationship between such growth rates and median age. The median age has been considered instead of the mean because the median provides a more robust measure in terms of outliers, and it gives information about the middle point observation. Again, we acknowledge the negative trend which is more consistent for the European and North American cases and more heteroscedastic in the case of Oceania. Older individuals are prone to having lower birth rates than younger people, which contributes to the overall fall in population growth and, additionally, as people age, their mortality rates rise, which magnifies the impact even further.

With respect to our subject of interest, we shall explore the effect of education enhancement on demographic expansion. Plot 10 reveals the inverse relationship between growth rate and number of school years and how this law tends to linger across decades. This is consistent with our opening hypothesis such that schooling factors and access to early education motivate family planning, which can result in individuals making more informed decisions about family planning and having fewer children. Education provides agents with knowledge

about reproductive health, contraception, and family planning methods, and hence chances are that they take reasoned choices about their reproductive preferences.

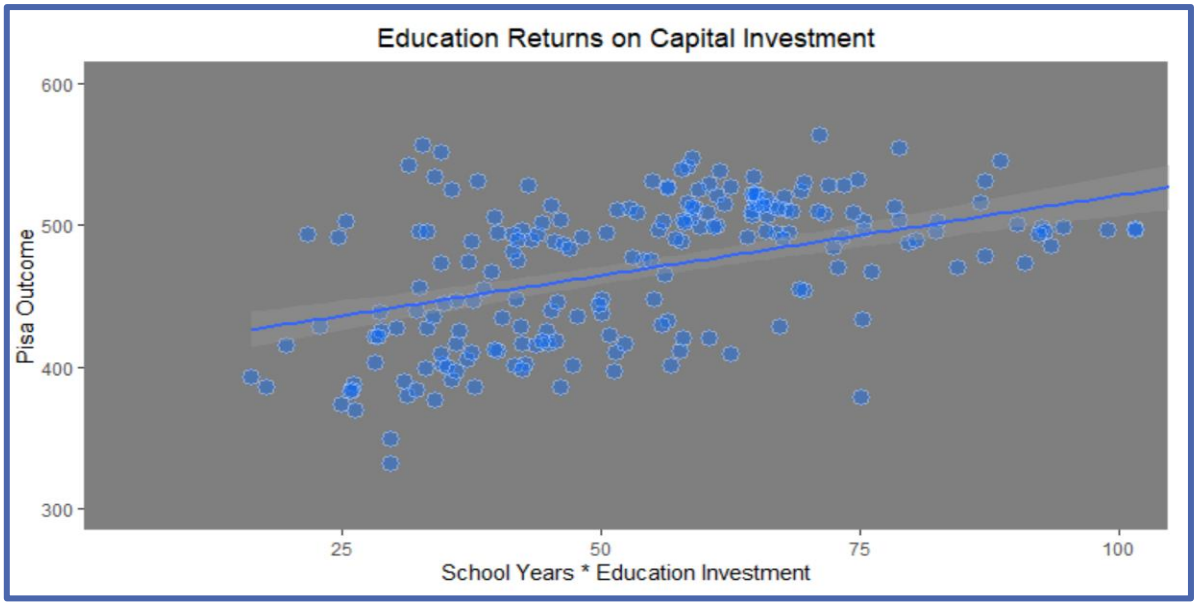


Plot 9: Population Growth vs Median Age by Continent, Own Elaboration

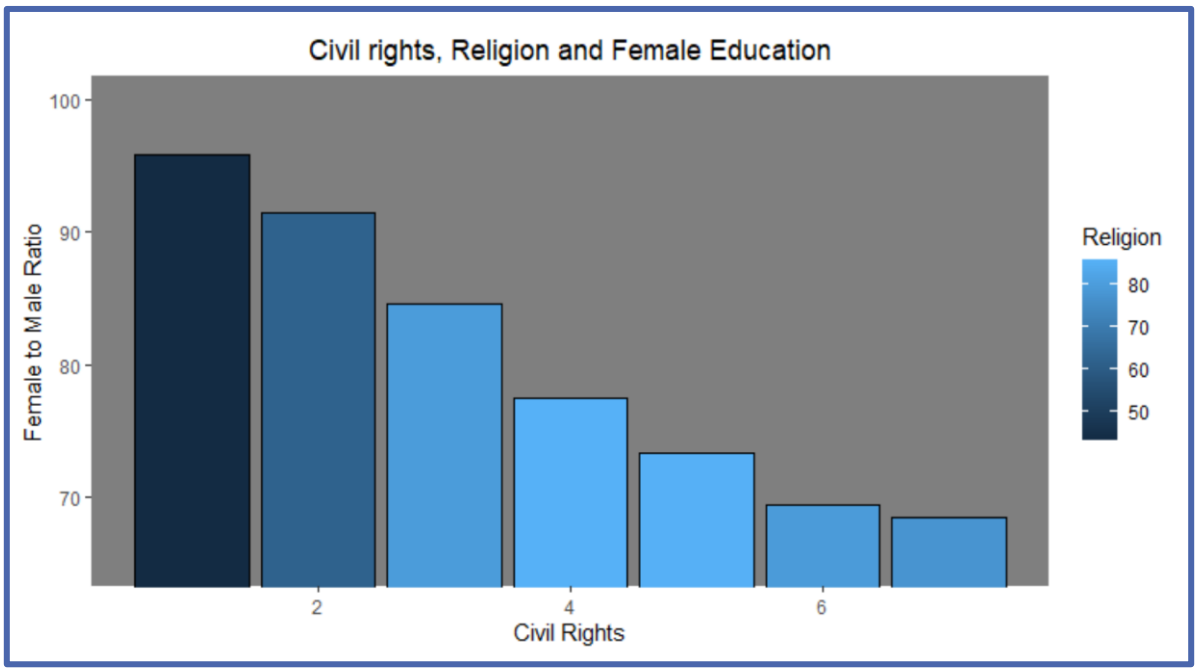


Plot 10: School Years vs Population Growth by Continent, Own Elaboration

Plot 11 portrays the correlation between the PISA (Programme for International Student Assessment) outcome and an education dimension gauged by the interplay of schooling years and education investment. The graphical representation exhibits an affirmative trend, though not a robust one, linking the two. This suggests that as the duration of schooling and investment in education augment, the academic returns also show an upward trajectory. Nevertheless, the constructive association is not markedly resilient, implying that other contributing factors may potentially exert a critical influence in determining PISA outcomes. Plot 12 illustrates two salient observations. Firstly, it highlights that economies with greater strictness in terms of religious identity tend to adopt a more conservative stance towards the enrolment of women in tertiary education. Secondly, the graph evinces that countries that uphold greater civil rights tend to exhibit a higher proportion of women pursuing higher education at the tertiary level. For the sake of our topic, the role of women in education, particularly in the context of fertility rates, assumes crucial significance, as research indicates that access to education has a direct bearing on women's reproductive health and decision-making.



Plot 11: Pisa Performance vs Education Interaction Term, Own Elaboration



Plot 12: Cultural Factors on Female Education Participation, Own Elaboration

Modelling

Regularized Elastic Net Regression

Accounting for the sizable database in terms of explanatory attributes which we had landed with as a result of our feature engineering process, our analysis demands the admission of a self-regulating mechanism to avoid collinearity issues and escape from potential overfitting. The difference between a linear regression and a regularization is that, in the case of linear regression we are minimizing the sum of squared residuals involved in our models whereas in regularization we will be adding an auxiliary term to account for variable inclusion penalization. This will prevent us from overfitting our model and reduce the variance of our predictions or sensibility to new data at the expense of introducing some bias.

The bias vs variance trade-off which was introduced at the beginning of the project is highlighted at this stage of our work when dealing with the inclusion-exclusion of a variable. Our aim will be to introduce some bias in our models in order to lower variance by incorporating some penalty on the model coefficients. Recall this mathematically makes sense since we are dropping the BLUE (best linear unbiased estimator) outcome to endow our research with a more original approach in terms of lowering the variability (Ng, 2004).

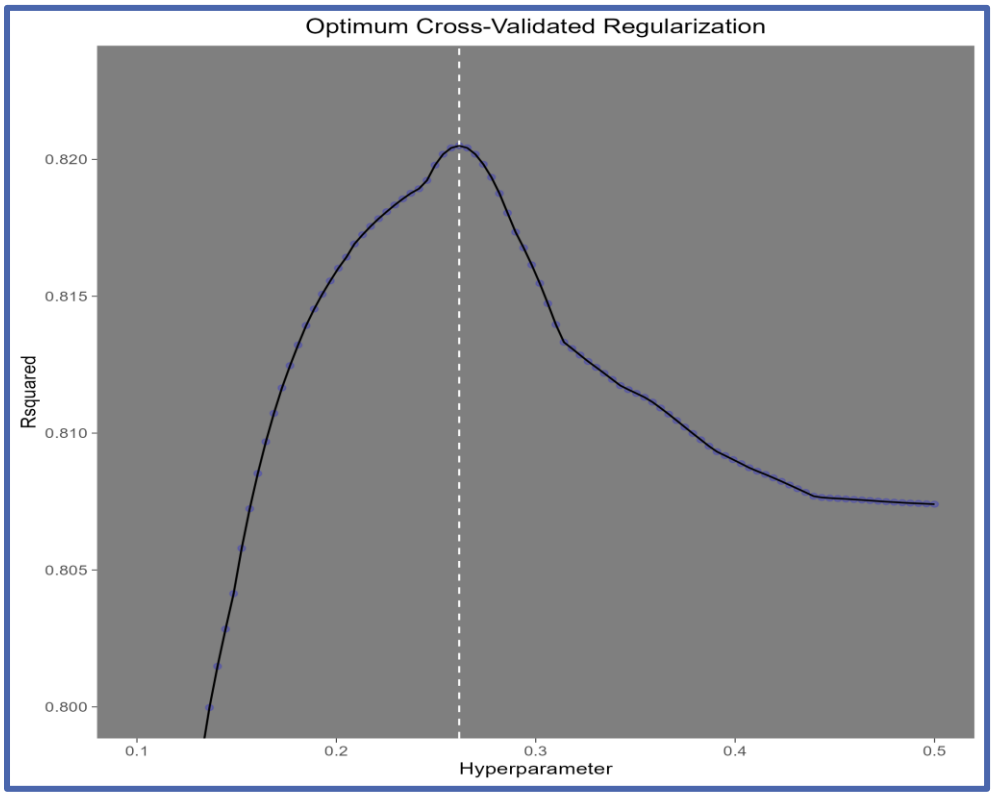
Regularization does not pertain to unbiased estimators; thus, it is possible to find an estimation with less variance. We summarize the penalization measure for a general elastic net model in the following equation:

$$Min\left[\sum(y-\hat{y})^2+\gamma_1\sum{\beta_i}^2+\gamma_2\sum|\beta_i|\right] \quad (\mathbf{5})$$

We will resort to cross-validation hyperparameter tuning by adopting the Lasso model such that we will divide our training set into k-folds, previously allocating a specific fold as validation and the rest serving as the learning sample. This process is iterated k times (each

fold used once for validation), and the performance of each model is evaluated based on our predefined performance metrics until the hyperparameters outputting the best realization are designated for the final process.

Plot 13 compiles all processed stages during the optimization iterations and the resulting R-squared corresponding to the weight yielding the best model behaviour. Notice this is introduced as a preceding feature selection algorithm in order to construct our multiple linear model in which the variables set to 0 by the Lasso model are dismissed.



Plot 13: Hyperparameter Tuning, Own Elaboration

Despite one´s concern with the development of the predictive linear model having grounds on the two-stage resolution of the previous dimensionality reduction task, table 2 sustains our premise by acknowledging the set of variables fully regularized during the process. This is, we propose the removal of those covariates holding a negligible weight according to Lasso but the re-training of model coefficients through ordinary least squares. This ensures not losing focus of our aim (education marginal effects) and avoids incurring in the omission of

meaningful attributes directly linked to our research question. For instance, according to table 2, education expenditure is barely contributing to explaining the variability in the dependent regressor, hence it is candidate for deletion. In such cases, it may be appropriate to keep the variable in the model, despite its small coefficient value. Removing it may lead to a loss of interpretability or explanatory power and could potentially alter the validity of the results.

Attribute	Regularized
(Intercept)	Fitting
Continent	Regularized
Gdppc_growth	Regularized
School_years	Fitting
Educ_expend	Regularized
Life_expec	Regularized
Health_expdpc	Fitting
Net_migration_rate	Fitting
Median_age	Fitting
Marriage_per_1000	Fitting
Civil_rights	Regularized
Relig_feel	Fitting
Ratio_f_m_educ	Fitting
Tert_educ	Fitting
Lg_Median_age	Fitting
Lg_health_expdpc	Regularized
Lg_Life_expec	Fitting
Lg_ratio_fm	Regularized
Lg_Educ_expend	Regularized
Lg_Gdppc_growth	Regularized
educexp_schlyrs	Regularized
Lifexp_schlyrs	Regularized
Ratio_fm_schlyrs	Fitting
Gdp_schlyrs	Fitting
Life_expec_sq	Regularized
Marriage_per_1000_sq	Regularized
ContinentAsia:civil_rights	Regularized
ContinentEurope:civil_rights	Fitting
ContinentNorth America:civil_rights	Fitting
ContinentOceania:civil_rights	Fitting
ContinentSouth America:civil_rights	Regularized
ContinentAsia:School_years	Regularized
ContinentEurope:School_years	Fitting
ContinentNorth America:School_years	Regularized

ContinentOceania:School_years	Fitting
ContinentSouth America:School_years	Fitting

Table 2: Regularized Feature Selection, Own Elaboration

Multiple Linear Model

Equation 6 provides the output of the multiple regression model in which we introduce the variables that survived the feature selection process arising from our previous elastic net. This framework yields statistically robust coefficients featuring education attributes covering logarithmic terms and interactions with school years, and outputs an R squared bordering on 81%. Our design meets the standards of global significance as pointed by the F-statistic rendering a residual standard error of half a percentage point, which is sparse. Nevertheless, the most salient output from the model is that the education variables holding the largest partial effect are the ones concerning the proportion of females in education (as a ratio to males) and schooling years. This reveals that education is a key driver of economic growth and cultural transformation such that institutional gender-equality investment in the matter leads economic resilience and sustainability (by demographic management). The full linear model table is included in the appendix at the end of the paper for one to inspect for further scrutiny.

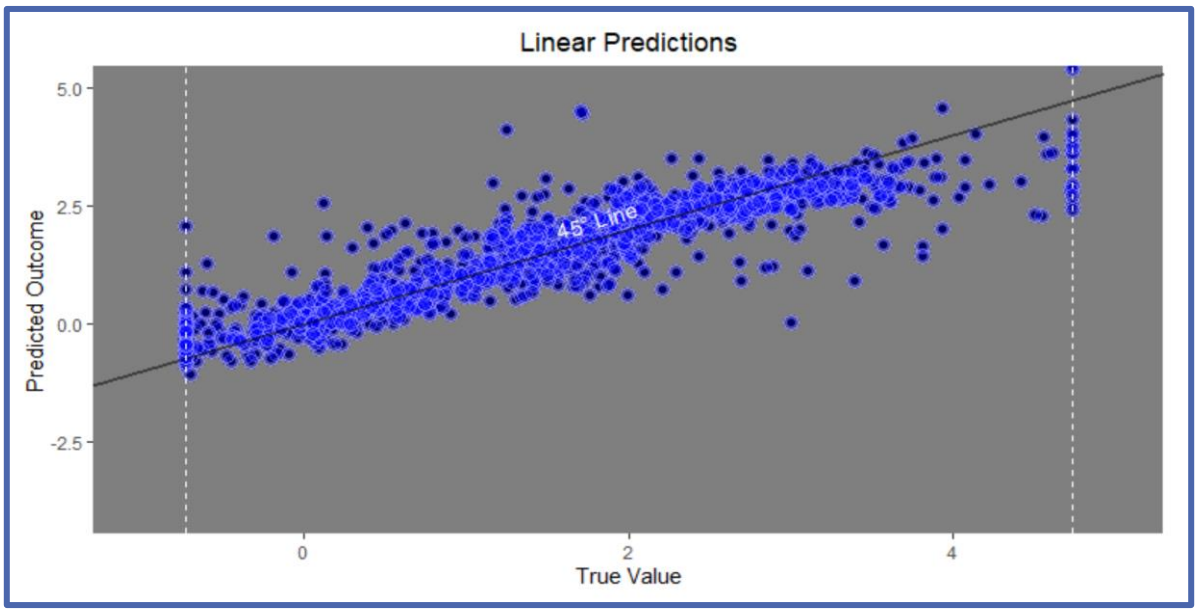
$$
\begin{aligned}
\mathrm{Pop}_{\mathrm{Growth}} = \\
1.56 \underset{***}{-0.04}\mathrm{Gdppc}_{\mathrm{growth}} \underset{*}{-0.31}\mathrm{School}_{\mathrm{years}} \underset{***}{-0.05}\mathrm{Educ}_{\mathrm{expend}} \underset{***}{+0.24}\mathrm{health}_{\mathrm{expdpc}} \underset{***}{+0.42}\mathrm{net}_{\mathrm{migration}_{\mathrm{rate}}} \\
\underset{***}{+0.02}\mathrm{marriage}_{\mathrm{per}_{1000}} \underset{***}{+0.07}\mathrm{relig}_{\mathrm{feel}} \underset{***}{-0.19}\mathrm{ratio}_{\mathrm{fm}_{\mathrm{educ}}} \underset{***}{+0.07}\mathrm{tert}_{\mathrm{educ}} \underset{***}{-1.26}\,\mathrm{lg}_{\mathrm{Median}_{\mathrm{age}}} \\
\underset{***}{+0.22}\,\mathrm{lg}_{\mathrm{Life}_{\mathrm{expec}}} \underset{***}{+0.39}\mathrm{ratio}_{\mathrm{fm}_{\mathrm{schlyrs}}} \underset{***}{+0.02}\mathrm{gdp}_{\mathrm{schlyrs}} \underset{*}{-0.06}\mathrm{ContinentAfrica{:}\,civil}_{\mathrm{rights}} \\
\underset{.}{-0.04}\mathrm{ContinentAsia{:}\,civil}_{\mathrm{rights}} \underset{***}{-0.09}\mathrm{ContinentEurope{:}\,civil}_{\mathrm{rights}} \\
\underset{***}{-0.13}\mathrm{ContinentNAmerica{:}\,civil}_{\mathrm{rights}} \underset{***}{+0.06}\mathrm{ContinentOceania{:}\,civil}_{\mathrm{rights}} \qquad (\mathbf{6}) \\
-0.03\mathrm{ContinentSAmerica{:}\,civil}_{\mathrm{rights}} \underset{*}{-0.03}\mathrm{ContinentAsia{:}\,School}_{\mathrm{years}} \underset{*}{-0.08}\mathrm{ContinentEurope{:}\,School}_{\mathrm{years}}
\end{aligned}
$$

$$+0.01\text{ContinentNAmerica}:\text{School}_{\text{years}} \quad -0.05\text{ContinentOceania}:\text{School}_{\text{years}} -0.041\text{ContinentSAmerica}:\text{School}_{\text{years}}$$

Our findings not only support our initial hypothesis, but also align with the theoretical frameworks proposed by Solow and Ramsey in the 20th century such that we encounter a negative relationship between the rates of population growth and GDP which confirms our expectations.

Furthermore, we have identified additional negative impacts on population dynamics related to education attributes such as school years, education investment, and the proportion of women in education. This finding supports our initial claim and suggests that investing in education and promoting gender equality in education can enable the control of demographic disparities. On the other hand, we have observed a positive association between health expenditure per capita and population evolution which also holds true for life expectancy. While healthcare spending appears to have a stronger influence on population growth trends, the marginal effect of school years reveals slightly larger. Specifically, our analysis indicates that each additional school year in Africa corresponds to a decrease of 0.31 percentage points in the population growth rate, being this effect even more pronounced when evaluating the relationship in Europe.

Plot 14 displays the fitted values of our linear model, which captures 81% of the variance in the response variable. Cut-off thresholds represented on the graph account for winsorization, which was introduced as an earlier stage to manage the behaviour of outliers in a way that top 2 percent deviations were replaced by quantiles 98 and 2, respectively. Such an R-squared indicates a strong relationship between the predictor and response variables, with the model accounting for a significant portion of the observed variation and suggests that the model's predictions are likely to be accurate, but further analysis will be applied to determine its relevance in quantifying marginal effects of education on population structure.



Plot 14: Actual vs Fitted values, Own Elaboration

Our analysis calls for the exploration of two crucial hypothesis testing. The initial enquiry concerns whether the marginal effect of school years essentially differs between European regions and African economies. We can evaluate this restriction by resorting to the coefficient of the interaction term between Europe and school years in equation 6 (recall reference category taken by our software is Africa) and examining the p-value. We notice that this element is statistically significant which involves that every additional school year in Europe is yielding a reduction of around 0.08% in population with respect to the effect in Africa. The second hypothesis requires the definition of our multiple restriction F-statistic as stated in equation (7) in which we will compare the residuals of the restricted model with the unbounded one.

$$F = \frac{({R^2}_{UR} - {R^2}_{R})(n-k-1)}{q(1-{R^2}_{UR})} \quad (7)$$

One would be interested in determining if the marginal effect of more women in tertiary education and the weight of boosting the aggregate proportion of people tertiary education are symmetrical. To test the latter, we need the following restriction to be met:

$$\frac{\partial PopGrowth}{\partial Ratio_{fm}} = \beta_1 + \beta_2 School_{yrs} = \frac{\partial PopGrowth}{\partial Tert_{educ}} = \beta_3 \rightarrow \beta_1 + \beta_2 School_{yrs} = \beta_3$$

Hence, when substituting the above expression of β_3, the corresponding terms in our equation become:

$$\beta_1 Ratio_{fm} + \beta_2 School_{yrs} * Ratio_{fm} + Tert_{educ}(\beta_1 + \beta_2 School_{yrs})$$

By factorizing and rearranging terms, our final model input is as follows:

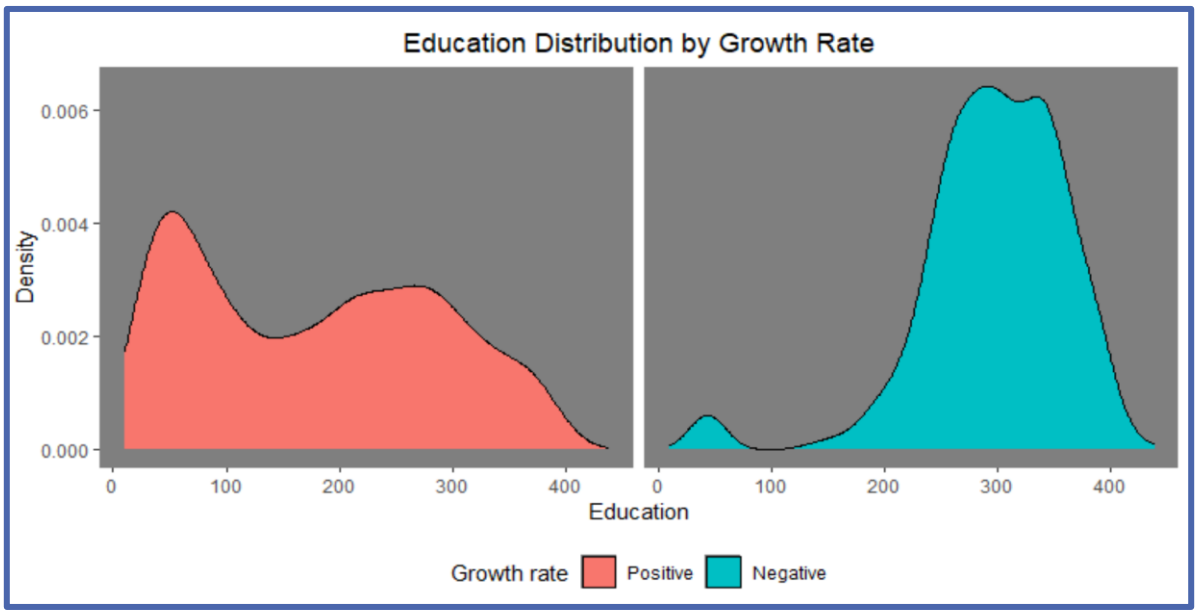
$$\beta_1(Ratio_{fm} + Tert_{educ}) + \beta_2 School_{yrs}(Ratio_{fm} + Tert_{educ}) \quad (\mathbf{8})$$

The F-statistic ($p<0.05$) lead the way to rejecting the null hypothesis such that the non-restricted model was yielding a better fitting in terms of squared residuals, hence it is established that the marginal effect of these two attributes remain unequal. To test for the one holding the greatest impact in population growth rate, we trained 3 additional models accounting for the absence of these two variables, the inclusion of one of them, and the admission of the other to monitor which one was being the major player in the R-squared of the model. Though the results for such analysis output minor distinctions, we notice that the fitting of the model is enhanced the most by capturing the influence of the ratio of females to males in education. Therefore, we conclude that the partial effect of introducing women to education is greater than the marginal impact of increasing the percentage of adults enrolled in tertiary education collectively. We identify that women's education results in broader socio-economic benefits that can indirectly influence population growth rates. The effect of female academic pursuits on demographic development is more sustainable in the long run compared to enrollment in universities. Education is a ongoing process that extends beyond formal schooling and can endow women with tools to settle about their personal lives, family structure, and communities.

Probabilistic Model

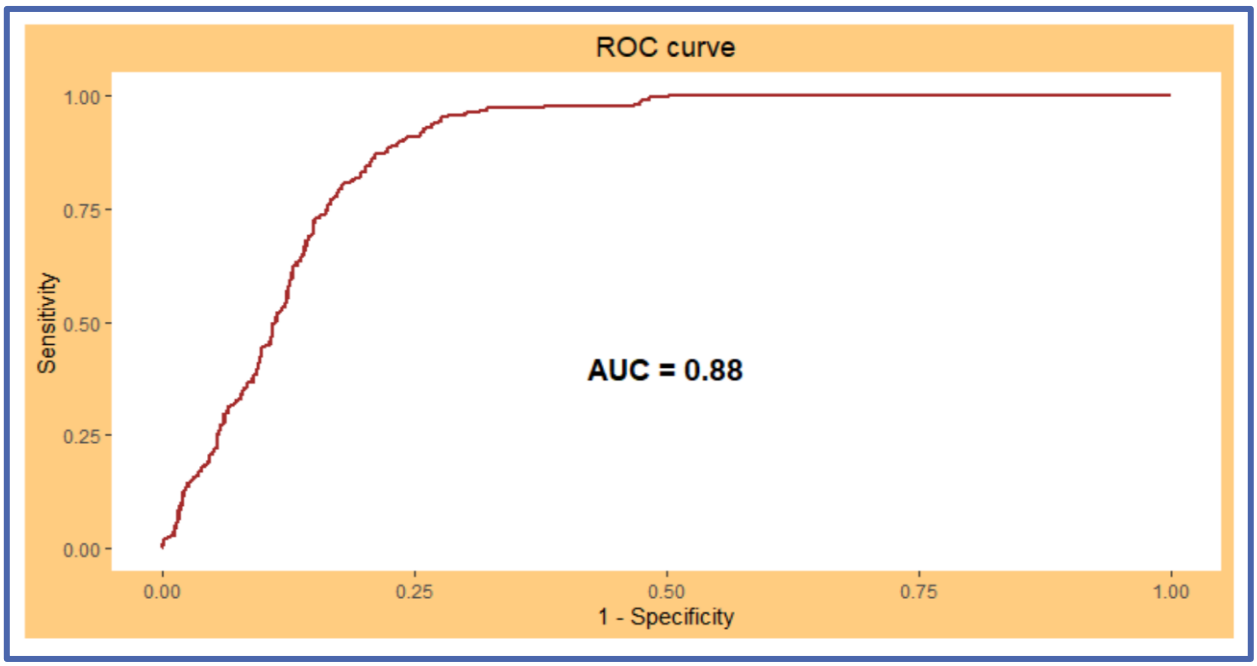
The output coefficients from our previous model are fitted into the generation of a single-weighted education attribute arguing for both the econometric omission of any potential confounders and the significance of each effect on population dynamics. This is, we will construct a binary classification model whose arguments exclusively take our designed aggregate education variable and the continental dimension (since one is interested in quantifying such impact across regions). Notice how the linear model enabled us a dual task, to obtain the statistically stainless weights accounting for all education aspects, and to test for the hypothesis of diverging slopes by continent.

We take advantage of a logistic regression or classification which allows the linear model to be related to the response by a logarithmic function. The advantage over linear regression is that we avoid providing probability outcomes larger than 1 or lower than 0 (unfeasible) and we obtain non-linear marginal effects (unrealistic) such that the partial effect on this model will weaken for extreme observations. The first step is to generate a new binary column which accounts for instances in which an economy face, for a given year, negative population rates. Plot 15 already characterizes how such economies are likely to hold enhanced education levels whereas the modal class for emerging population is contained within the bottom education values.



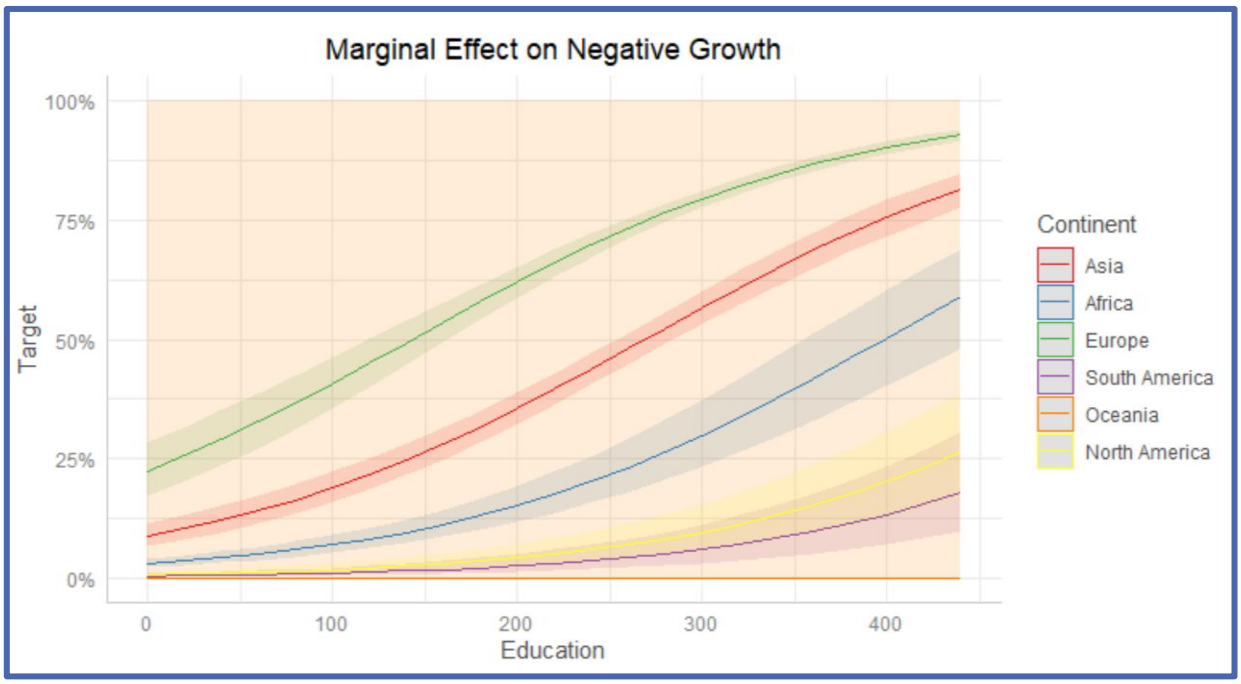
Plot 15: Education Distribution on Population Dynamics, Own Elaboration

The aim of this model is to assess the accuracy with which one can classify countries in terms of positive or negative growth rate in accordance with the input of a single variable, our aggregate education dimension. Since more than 90% of economies (each year) occupy positive growth rates, we resorted to an up-sampling algorithm to generate synthetic data and bias the model in order to weight more the minority class. This enables the elusion of our model to learn to predict the majority class distinctively at the expense of the minority class prompting overfitting, which may result in a narrow generalization on unseen observations. Plot 16 shows depicts the ROC curve at different thresholds of our binary classifier, yielding an area under the curve of 0.88, which indicates that the model has a favourable ability to distinguish between positive and negative classes, and improving the performance in contrast with random guessing.



Plot 16: ROC Curve of Logistic Model, Own Elaboration

We close our binary analysis by plotting the results of the marginal effect (plot 17) of our model as a function of our education aggregate and quantify these results by continent. This is, together with the results of the linear model and the hypothesis on the importance of gender parity, one of the most noteworthy findings from the study. One identifies not merely that the partial contribution of education to the likelihood of experiencing a negative population growth rate follows a concave functionality (with leveler marginal effects at the extremes), but additionally these results differ significantly across continents. The area embedded between the curves belonging to Europe and Asia hold relatively constant with a vertical difference close to 25 percentual points, implying that, conditional to having similar education variables, a European economy is 25% more likely than an Asian one to fall into this population contraction force. As a demonstrative case for the user, if we input the corresponding average values for Spain in terms of schooling years or the proportion of women in education, we obtain an overall aggregate result of 302 over the designated interval. This means that the probability of Spain holding a negative growth rate is close to 80% and it postulates as one of the riskiest countries in Europe with this connotation (it is indeed one of those that have already exceeded such a threshold).

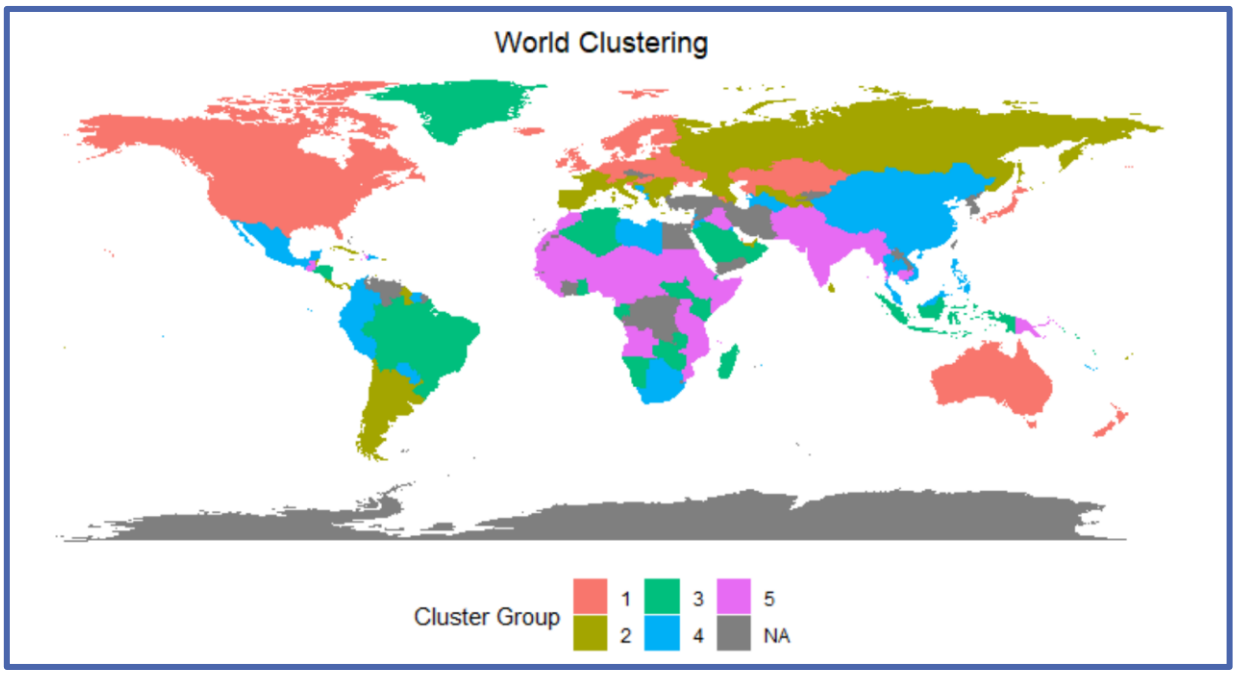


Plot 17: Logistic Partial Fixed Effects, Own Elaboration

Time Series Clustering Analysis

We will offer a final section to evaluate our population data from a time series perspective to compare the results from the previous models and audit for potential autoregressive components in demography dynamics. By resorting to our education aggregate computed in the former segment as the exclusive input, we will separate our countries into 5 distinct clusters based on the unsupervised k-means algorithm which groups the observations by minimizing the Euclidian distance to each center. Acknowledging that one should ensure observations within a cluster are similar and variability between collections should be maximized, this process will be iterated several times and the resulting combination yielding the minimum variation will be selected. We map the clustering results in plot 18 in which we can detect that southern European countries are grouped together with South American regions and Russia, pertaining to the second-best clan in terms of education.

The strongest cluster encompasses Nordic economies, Australia, and North America whereas the least advanced one is occupied by most African economies in combination with India. In connection to women's education, India has made significant strides in recent years, with a female literacy rate of around 69% compared to the 65% of Sub-Saharan Africa. However, female access to education in the country is still hindered by multiple forces such as poverty, social institutions, and gender-based discrimination. In some African districts, women also face significant barriers to study, ranging from early marriage and poverty to cultural norms. Recall our aggregate education variable is composed of various factors such as education investment, schooling years or ratio of women in education.



Plot 18: World Clustering by Education

We select 10 random economies from each cluster and compute the mean change over time in terms of population growth rate to input the information into our ARIMA series modelling. The primary requirement for our models is the stationarity of the data, which we will address by taking the first difference on each cluster and evaluating the optimal autoregressive and moving average components for each of our groups (Box and Jenkins, 1970). We will operate under the assumption of multivariate normality and weak stationarity such that we condition on the fact that our series arises from an homogenous mean and variance, and an ergodicity assumption in which we accept that the covariance between two time periods is a function of the number of lags between them ($\lim(x \rightarrow \infty)\ 0$).

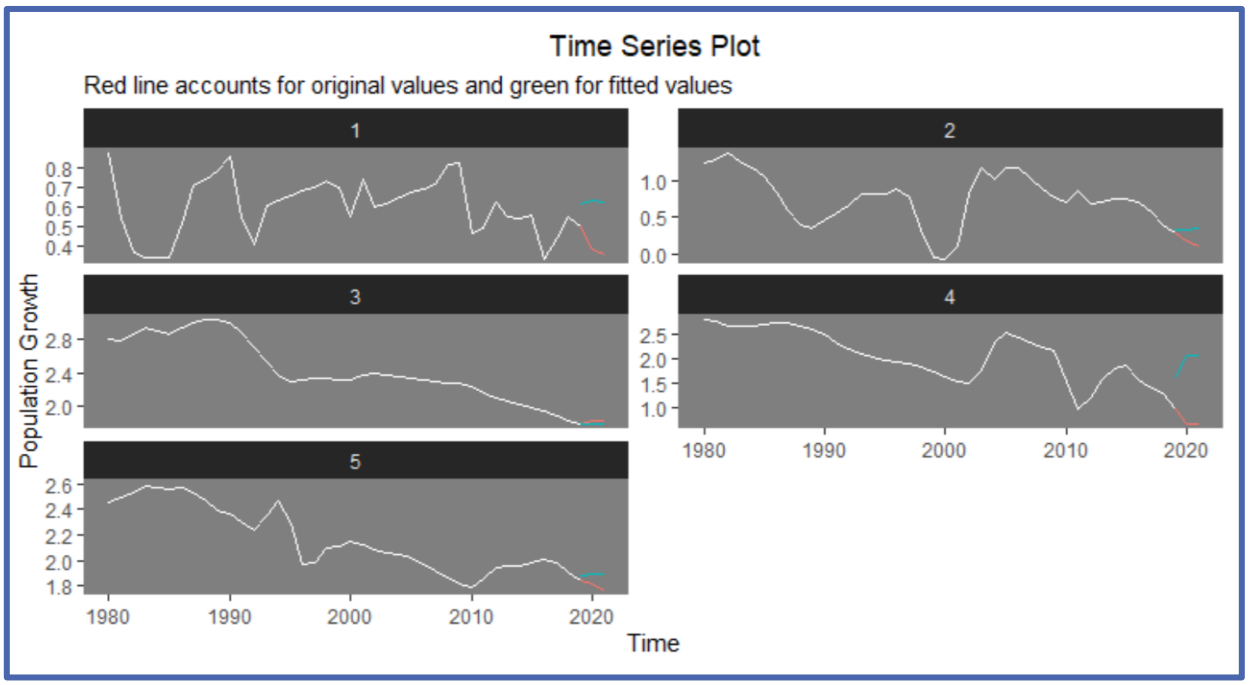
The appraisal of the set of Dickey fuller tests (unit root) to evaluate the premise of stationarity in mean can be resorted to by the user in the appendix in which the null hypothesis states that the coefficient of such autoregressive component behaves like a random walk. The "AR" component will model the relationship between the current value of the temporal data and its past values and will assume that the latest observation of the series can be explained by a linear combination of its previous values. The "MA" element represents the association between the current measurement of the series and earlier shocks. This considers that the

current growth rate depends on a linear combination of the error terms from previous time periods:

$$\dot{y}_t = c + \emptyset_1 \dot{y}_{t-1} + \emptyset_2 \dot{y}_{t-2} + \cdots \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \cdots \quad (9)$$

where $\dot{y} : Population\ Growth\ Rate$
$\emptyset_n : Autoregressive\ Component\ Coefficient$
$\theta_n : Moving\ Average\ Coefficient$

By estimating a different model for each of our clusters, the results from our time series template serve as a basis in order to compare the quality of predictions we can obtain from our previous educational analysis. The results point to a very significant autoregressive component of population growth rate and, in the last 3 clusters, a very marked moving average ingredient driving population dynamics. An elaborated finding of the models constructed for each group can be found in the appendix together with their corresponding Akaike evaluation and the specific coefficients accounting for expression (9). Since ARIMA models are suitable for short-term forecasting, and for the sake of splitting our analysis into a teaching and a testing set, we first fitted the most appropriate one for each of the series and reserved the most recent 3 years of data for predictive purposes:



Plot 19: Time Series Forecasting

We obtain accurate predictions such that the standard error of our model is around 0.55%, suggesting that, on average, the estimated population growth rate is expected to deviate from the true magnitude by approximately half a percentage point. ARIMA models are based on historical patterns and relationships within the data, hence as we forecast further into the future, the uncertainty for unforeseen external factors not captured in the previous data increase. If the reader recalls, the standard error for the linear model was approximately 0.5%, indicating a high level of precision in predicting short-term average population trends through our education analysis compared to ordinary time series evaluation, which is often more appealing and widely favored. This implies that our approach taking advantage of education as a factor in demographic analysis enables one to achieve remarkable proximity in forecasting short-term population trajectories.

References

StudySmarter UK. (n.d.). Population and economic growth. Retrieved from https://www.studysmarter.co.uk/explanations/macroeconomics/economic-performance/population-and-economic-growth/

Nizamul Islam, M. (2013). The optimal population growth rate in Diamond (1965) model: The role of demographic dividend. Journal of Economic Studies, 40(6), 744-757. doi: 10.1108/JES-04-2012-0049

Sustainable Development Report Index and Dashboards. (n.d.). SDG 4: Quality Education. Retrieved from https://dashboards.sdgindex.org/map/goals/SDG4

World Bank. (n.d.). Home. Retrieved from https://www.worldbank.org/en/home

Hlavac, M. (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. Retrieved from https://CRAN.R-project.org/package=stargazer

United Nations. (n.d.). Population. Retrieved from https://www.un.org/es/global-issues/population

World Economic Forum. (2015). How education can moderate population growth. Retrieved from https://www.weforum.org/agenda/2015/07/how-education-can-moderate-population-growth/#:~:text=%E2%80%9CEducation%20leads%20to%20lower%20birth,economic%20growth%20easier%20to%20achieve.%E2%80%9D

Albert Wang. (n.d.). albertyw/avenews: Avenews 1.0.0. [GitHub repository]. Retrieved from https://github.com/albertyw/avenews

Pew Research Center. (2016, December 13). How Religion May Affect Educational Attainment: Scholarly theories and historical background. Retrieved from https://www.pewresearch.org/religion/2016/12/13/how-religion-may-affect-educational-attainment-scholarly-theories-and-historical-background/

Solow, R. M. (1956). A Contribution to the Theory of Economic Growth. The Quarterly Journal of Economics, 70(1), 65-94. Retrieved from http://piketty.pse.ens.fr/files/Solow1956.pdf

National Institute on Aging (NIA). (n.d.). Education and gender inequality may explain why India's women have worse late-life cognition. Retrieved from https://www.nia.nih.gov/news/education-and-gender-inequality-may-explain-why-indias-women-have-worse-late-life-cognition#:~:text=Overall%2C%20men%20had%20much%20more,oldest%20adults%20in%20the%20sample.

Ng, A. Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. Computer Science Department, Stanford University, Stanford

Box, G.E.P. and Jenkins, G.M. 1970. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day.

Appendix

This is the original output table from the linear model corresponding to equation 6 encompassing the level of significance for each variable along with the associated metrics of fitting and standard errors. These results are visually represented in plot 14.

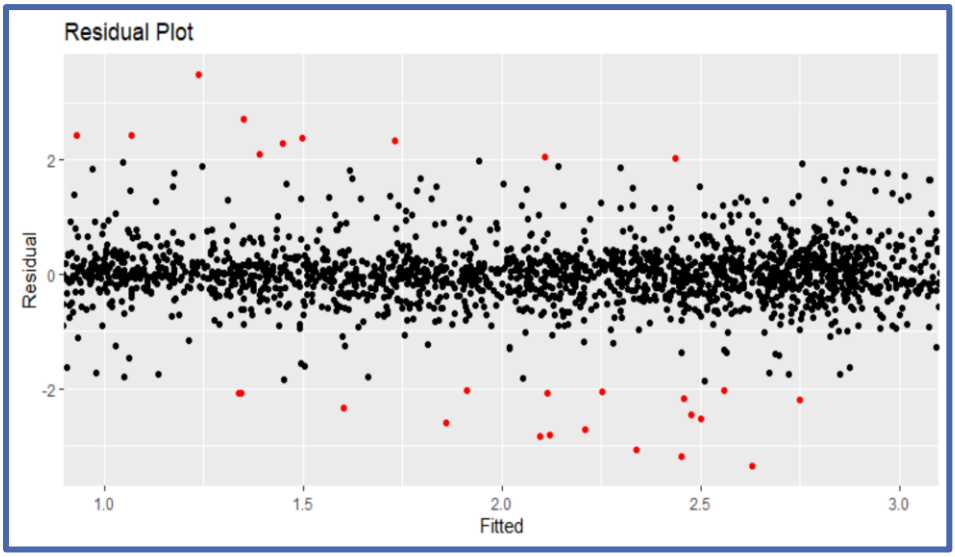
Table 1: Multiple Linear Model

	Dependent variable:
	. outcome
Gdppc_growth	−0.045**
	(0.021)
School_years	−0.309***
	(0.063)
Educ_expend	−0.048***
	(0.011)
health_expdpc	0.235***
	(0.018)
net_migration_rate	0.416***
	(0.011)
marriage_per_1000	0.015
	(0.012)
relig_feel	0.070***
	(0.016)
ratio_f_m_educ	−0.194***
	(0.033)

tert_educ	0.065***
	(0.014)
lg_Median_age	−1.264***
	(0.029)
lg_Life_expec	0.219***
	(0.022)
ratio_fm_schlyrs	0.394***
	(0.089)
gdp_schlyrs	0.017
	(0.021)
'ContinentAfrica:civil_rights'	−0.065**
	(0.027)
'ContinentAsia:civil_rights'	−0.045*
	(0.025)
'ContinentEurope:civil_rights'	−0.087***
	(0.017)
'ContinentNorth America:civil_rights'	−0.128***
	(0.018)
'ContinentOceania:civil_rights'	0.055
	(0.111)

'ContinentSouth America:civil_rights'	−0.030
	(0.021)
'ContinentAsia:School_years'	−0.025
	(0.033)
'ContinentEurope:School_years'	−0.077*
	(0.046)
'ContinentNorth America:School_years'	0.013
	(0.024)
'ContinentOceania:School_years'	−0.046
	(0.112)
'ContinentSouth America:School_years'	−0.041
	(0.027)
Constant	1.563***
	(0.010)
Observations	3,067
R^2	0.818
Adjusted R^2	0.817
Residual Std. Error	0.550 (df = 3042)
F Statistic	569.699*** (df = 24; 3042)
Note:	*p<0.1; **p<0.05; ***p<0.01

Residual plot arising from linear model in equation 6:

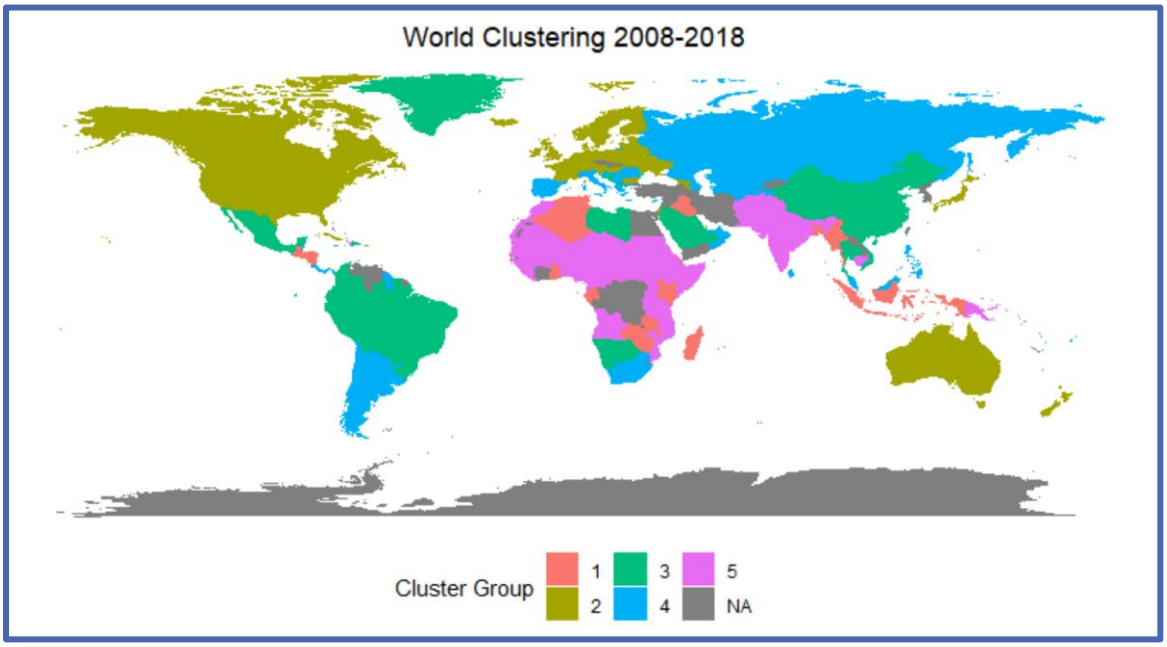


We provide the reader with the confusion matrix and performance metrics of the logistic model in which selected the threshold such that we optimized the sensitivity, this is, the value which accounts for the correct classification of the minority class. Recall we introduced a reweighting of our data to account for the imbalance which yielded an AUC of 0.88 (Plot 16).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1165   19
         1  372  188

               Accuracy : 0.7758
                 95% CI : (0.7555, 0.7952)
    No Information Rate : 0.8813
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3833

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9082
            Specificity : 0.7580
         Pos Pred Value : 0.3357
         Neg Pred Value : 0.9840
             Prevalence : 0.1187
         Detection Rate : 0.1078
   Detection Prevalence : 0.3211
      Balanced Accuracy : 0.8331

       'Positive' Class : 1
```

In the context of the clustering analysis, we present the reader with the aggregated grouping of countries based on their education indicators for the period spanning from 2008 to 2018. We observe that Europe, North America, and Asia form a distinct cluster, indicating similarities in their education profiles. Conversely, Spain is grouped together with other regions in Southern Europe, such as Italy, which suggests that Spain's education characteristics are more aligned with those of the Southern European economies rather than with the broader European cluster. Moreover, our analysis reveals that Spain's positioning

within this cluster is influenced by its relatively low investment in education compared to other economies.



In our analysis, we incorporate the stationarity Dickey-Fuller tests for each of the clusters, with the dependent variable being the population growth rate. It is important to note that the null hypothesis in this test assumes the coefficient to be equal to one, implying the absence of stationarity in the data, resembling a random walk pattern. By conducting these tests, we aim to examine the stationarity properties of the population growth rate within each group. By assessing the stationarity of population dynamics in each cluster, we gain insights into the temporal properties of the variable and its potential implications for modeling and analysis.

```
        Augmented Dickey-Fuller Test

data:  time_series$`1`
Dickey-Fuller = -2.9519, Lag order = 3, p-value =
0.1984
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  time_series$`2`
Dickey-Fuller = -3.1446, Lag order = 3, p-value =
0.1225
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  time_series$`3`
Dickey-Fuller = -4.0729, Lag order = 3, p-value =
0.01633
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  time_series$`4`
Dickey-Fuller = -3.2148, Lag order = 3, p-value =
0.09802
alternative hypothesis: stationary
```

```
        Augmented Dickey-Fuller Test

data:  time_series$`5`
Dickey-Fuller = -2.5916, Lag order = 3, p-value =
0.3405
alternative hypothesis: stationary
```

In our study, we incorporate the analysis derived from ARIMA models, specifically examining the coefficients obtained for each of the clusters. Though the determination of the number of differences to be applied in the ARIMA models is directly informed by the preceding Dickey-Fuller analysis, it is important to note that over-differencing, while potentially introducing some distortion, is considered preferable to under-differencing due to its lower cost. We are implicitly evaluating the correlogram plots generated from the

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). The ACF plot represents the correlation between the current observation and the preceding observations at different lag intervals. It helps identify the potential presence of serial correlation or seasonality in the data. The PACF plot, on the other hand, shows the correlation between the current observation and its past values, while accounting for the influence of intermediate observations. It helps identify the appropriate order of the autoregressive component in the ARIMA model, as significant spikes indicate direct relationships between the current observation and previous measurements.

```
Series: first_years[, i]
ARIMA(2,0,0) with non-zero mean

Coefficients:
         ar1      ar2    mean
      1.3594  -0.6244  0.7509
s.e.  0.1220   0.1241  0.0883

sigma^2 = 0.02356:  log likelihood = 18.21
AIC=-28.43   AICc=-27.25   BIC=-21.77
```

```
Series: first_years[, i]
ARIMA(2,1,0) with drift

Coefficients:
         ar1      ar2    drift
      0.4931  -0.5826  -0.0153
s.e.  0.1298   0.1255   0.0090

sigma^2 = 0.003803:  log likelihood = 53.05
AIC=-98.09   AICc=-96.88   BIC=-91.54
```

```
Series: first_years[, i]
ARIMA(2,1,1)

Coefficients:
         ar1      ar2     ma1
      0.1880  -0.1682  0.9396
s.e.  0.1757   0.1700  0.1433

sigma^2 = 0.02151:  log likelihood = 19.2
AIC=-30.4   AICc=-29.19   BIC=-23.85
```

```
Series: first_years[, i]
ARIMA(0,0,1) with non-zero mean

Coefficients:
         ma1   mean
      0.6503  0.603
s.e.  0.1003  0.031

sigma^2 = 0.01476:  log likelihood = 27.62
AIC=-49.23   AICc=-48.55   BIC=-44.24
```

```
Series: first_years[, i]
ARIMA(0,1,2) with drift

Coefficients:
         ma1     ma2    drift
      1.5608  0.6107  -0.0267
s.e.  0.1472  0.1432   0.0150

sigma^2 = 0.0009631:  log likelihood = 77.74
AIC=-147.49   AICc=-146.28   BIC=-140.94
```