Modeling Cultural Assimilation Using Global Migration, Economic Integration, and Cultural Diversity Metrics [*]

Connor Cruz
NCSSM Online
North Carolina School of Science and Mathematics
Durham, North Carolina

December 10, 2025

Abstract: Cultural assimilation is often assumed to be shaped by contemporary forces such as increasing globalization, international migration, and economic integration. This study investigates the extent to which demographic and economic variables can explain variation in cultural diversity. Using Mathematica's *CountryData* resource, we constructed a dataset of linguistic diversity, religious diversity, population, geographic size, GDP per capita, trade value, and net migration rates for over 200 countries. We applied logarithmic transformations to reduce skewed distributions and built a religious fractionalization index following Alesina et al. [1]. Multiple linear regression models were then used to assess the influence of migration and economic integration on cultural diversity while controlling for population and area. The results illustrate that migration and trade are not significant predictors of either linguistic diversity or religious fractionalization. Instead, the strongest predictor in both models was geographic area, suggesting that cultural diversity is shaped more by deep historical and geographic forces than by contemporary globalization. These findings reinforce the importance of computational methods in testing widely held assumptions about cultural change.

Key words: cultural assimilation, migration, religious fractionalization, linguistic diversity, computational social science, computational science, regression modeling, globalization

[*] Correspondence to: cruz26c@ncssm.edu

Introduction

Cultural assimilation is a central theme in the study of globalization. As nations become increasingly interconnected through trade, migration, and communication technologies, societies experience new forms of cultural contact that influence languages, religions, values, and social identity. Economic integration fosters shared markets and common communication systems, often diffusing cultural norms across borders [3]. For example, English has become the global language of business due to the economic prominence of the United States, shaping linguistic expectations even in non-English-speaking nations.

Despite this, cultural assimilation through economic forces is not universal. Some countries deliberately resist cultural homogenization. For instance, France has long adopted cultural protection policies to preserve linguistic and artistic traditions [5], demonstrating that globalization pressures can be met with both adoption and resistance.

Migration serves as another powerful mechanism of cultural change. When people cross borders for work, education, or refuge, they bring their languages, religions, and customs with them. Classic examples include the migration of Turkish laborers to Germany under the Gastarbeiter program and the growth of bilingual Hispanic communities in the United States. These cases illustrate that migration rarely produces complete assimilation; instead, it generates hybrid identities, a dynamic documented extensively in acculturation research [4].

Although both economic integration and migration appear to influence cultural assimilation, their impact varies widely across countries and remains difficult to quantify. Some of the most culturally diverse nations have minimal modern migration, while highly immigrant-receiving countries can remain relatively homogeneous. This mismatch raises an important empirical question: To what extent do contemporary forces such as migration and economic integration actually explain the cultural diversity observed across countries today?

To investigate this question, we use a computational approach combining global economic, demographic, and cultural data from Mathematica's *CountryData* resource. We analyze two dimensions of cultural diversity: linguistic diversity and a newly constructed religious fractionalization index. We then apply multiple linear regression to test whether migration rate, GDP per capita, and trade value significantly predict either form of diversity after controlling for population and geographic area. We then evaluate how these results align with common assumptions about globalization and assimilation.

Computational Approach

The data used in this study were gathered from Mathematica's *CountryData* resource [6]. We obtained the following variables for each country: GDP per capita, trade value, net migration rate, population, geographic area, languages spoken, religions, and religious population shares.

Certain variables exhibited strong positive skew, particularly trade value and population. To mitigate the influence of outliers, four variables were logarithmically transformed: GDP per capita, trade value, population, and geographic area. Since trade value may be zero for some countries, we defined

$$\text{LogTradeValue} = \log(1 + \text{TradeValue}).$$

To measure cultural diversity, and therefore cultural assimilation, two quantities were analyzed: linguistic diversity and religion.

Linguistic diversity was quantified simply as the number of languages listed for each country, an approach that aligns with prior large-scale cultural analyses [2]. Although this is a simple metric, this measure effectively captures significant differences between countries with only one or a few dominant national languages and those with potentially hundreds of indigenous languages.

Religious diversity was measured using the widely used religious fractionalization index (RFI), an adaptation of the Herfindahl concentration index [1]. The RFI represents the probability that two randomly selected citizens adhere to different religions. An RFI close to 0 would correspond to less religious diversity, while a higher RFI would imply more religious diversity among the population.

Figure 1 shows the religious fractionalization index, where n represents the total amount of religions in a population and π_i is the amount of the population which practices religion i:

$$RFI = 1 - \sum_{i=1}^{n} (\pi_i)^2$$

Figure 1: Religious Fractionalization Index Formula

In addition to constructing diversity measures, several preprocessing steps were also required to ensure that the data were able to be statistically modeled. Many values obtained from the *CountryData* are expressed as *Quantity* objects with units attached, which cannot be directly analyzed in numerical computations. To address this issue, all quantities were converted to unit-free numerical values using the *QuantityMagnitude* function. Additionally, missing or incomplete entries were removed to prevent inaccurate data points, allowing relationships to be expressed more factually.

The data were finally combined into association structures, which allowed the regression models to identify each country's data as observations and ensured consistent analysis of a country's provided data.

Two multiple linear regression models were constructed. The first used the number of languages as the dependent variable, while the second used RFI. Both models used the same predictors: migration rate, LogGDPPerCapita, LogTradeValue, LogPopulation, and LogArea.

A multiple linear regression framework was selected for this study because it provides an interpretable method for analyzing the relationship between cultural diversity metrics and their respective predictors. Linear models allow the unique contribution of each variable to be isolated while the others remain constant, thus allowing for multiple correlative factors to be analyzed. Although cultural occurrences are often complex and shaped by nonlinear processes, linear regression still offers a useful approximation of relationships and is widely used in studies on cultural and economic diversity [1, 2]. The logarithmic transformations applied to skewed variables helped to satisfy the assumption of linearity by mitigating the effects of outliers. Logarithmic transformations also lend to the assumption of homoscedasticity. Moreover, because each country represents an independent geopolitical unit, the assumption of independent observations is reasonably justified. While the models created do not explicitly test for the other criteria for linear regression to apply, the consistency of the results suggests that the same conclusions will be obtained when faced with minor deviations from these assumptions. Nevertheless, future work could explore alternative approaches in modeling this relationship, such as through hierarchical models or nonlinear regression, to capture more complex relationships which may not fit a linear pattern.

Results and Discussion

Migration and Cultural Diversity

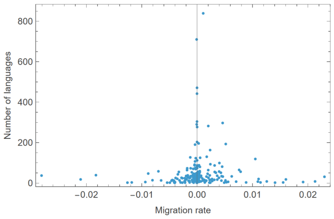


Figure 2: Migration Rate vs. Number of Religions Scatterplot

To visualize the relationship between migration rate and cultural diversity, we created two scatterplots. Figure 2 displays migration rate versus the number of languages spoken in each selected country. As shown by the figure, the points cluster tightly around a migration rate of zero. However, the number of languages ranges from one to over 800. There does not appear to be any linear trend, nor does there seem to be any trend in general.

Countries with high linguistic diversity do not seem to exhibit unusual or significantly high migration rates. Moreover, countries with low linguistic diversity occur across all migration levels.

A similar pattern is evident in Figure 3, which plots migration rate against religious fractionalization (obtained via the religious fractionalization index). Migration rates share the same pattern as above, being centered around zero, while religious diversity spans nearly the entire range of 0 to 1. Countries of high fractionalization, as well as countries with relatively homogeneous fractionalization, mostly appear with a migration rate near 0. In general, there is not much clustering nor a significant relationship between the data.

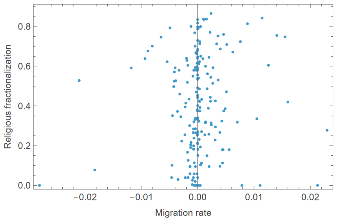


Figure 3: Migration Rate vs. Number of Religions Scatterplot

These scatterplots ultimately support the conclusion that modern migration rates do not meaningfully correspond to levels of cultural diversity. Any variation in diversity, whether in languages or religion, appears independent of contemporary migration.

Linguistic Diversity Regression Results

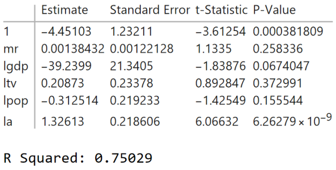
Figure 4 summarizes the results of a multiple linear regression model predicting the number of languages from migration rate (mr), log GDP per capita (lgdp), log trade value (ltv), log population (lpop), and log area (la).

It is shown through this table that LogArea has a heavily significant p-value of roughly 6.26×10^{-9}. As such, geographic area is found to have a high correlation with migration. This can be attributed to larger countries having a higher likelihood of containing more languages, which is consistent with patterns exemplified by geographic isolation in the past.

Conversely, migration rate does not have a significant p-value, which supports the data found in Figure 2. Trade value and population size also do not have significant p-values.

GDP per capita is almost significant, assuming $\alpha = 0.05$, suggesting that there may be a correlation between GDP per capita and number of languages, but this result is weak and inconclusive.

Finally, $R^2 \approx 0.75$, implying that the model explains roughly 75% of the variation in linguistic diversity. Most of this explanatory power comes from geographic area, rather than the studied migration and economic variables. Hence, it is reasonable to conclude that linguistic diversity is largely determined by historical and geographic factors, rather than modern patterns in globalization.



	Estimate	Standard Error	t-Statistic	P-Value
1	−4.45103	1.23211	−3.61254	0.000381809
mr	0.00138432	0.00122128	1.1335	0.258336
lgdp	−39.2399	21.3405	−1.83876	0.0674047
ltv	0.20873	0.23378	0.892847	0.372991
lpop	−0.312514	0.219233	−1.42549	0.155544
la	1.32613	0.218606	6.06632	6.26279×10^{-9}

```
R Squared: 0.75029
```

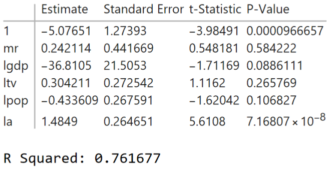
Figure 4: Multiple Regression on Number of Languages

Religious Fractionalization Regression Results

Figure 5 presents the regression results for the religious fractionalization index based on the same independent variables.

In this table, LogArea remains as the only heavily significant predictor, with $p \approx 7.17 \times 10^{-8}$, displaying a positive relationship between area and religious diversity. This is supported by the general notion that larger countries should contain more religious groups and thus more balanced religious distributions.

Similarly to the results in Figure 4, migration rate, GDP per capita, trade value, and population do not have a significant influence on religious diversity.



	Estimate	Standard Error	t-Statistic	P-Value
1	−5.07651	1.27393	−3.98491	0.0000966657
mr	0.242114	0.441669	0.548181	0.584222
lgdp	−36.8105	21.5053	−1.71169	0.0886111
ltv	0.304211	0.272542	1.1162	0.265769
lpop	−0.433609	0.267591	−1.62042	0.106827
la	1.4849	0.264651	5.6108	7.16807×10^{-8}

```
R Squared: 0.761677
```

Figure 5: Multiple Regression on Number of Languages

The model fit R^2 for this regression has a high degree of similarity to that of Figure 2, suggesting that geographic area has a similar positive influence on both language amount and religious diversity. Thus, the data suggest that geographical area and geography play a fundamental role in sustainable cultural diversification. They also undermine the assumption that modern demographic and economic variables have a significant effect on cultural diversity.

Conclusions

This study tested whether contemporary migration rates and economic integration predict linguistic and religious diversity across countries. Using a global dataset from Mathematica's *CountryData* resource [6] and diversity metrics grounded in established cultural theory [1, 2], we constructed two multiple linear regression models.

It was found that, across both of these regression models, migration rate was not a significant predictor of cultural diversity with respect to the assigned metrics. Similarly, economic measures such as GDP per capita and trade value displayed no significant relationship with either linguistic diversity or religious fractionalization. Instead, it was found in both models that the strongest and most consistent predictor was geographic area. Hence, it is concluded that larger countries exhibit greater cultural diversity.

These results suggest that cultural diversity in the present world is not heavily affected by modern globalization forces like migration and economy. Instead, cultural diversity reflects long-term historical events and processes that persist into the future regardless of recent migratory or economic changes, as geographic boundaries most often remain unchanged. These results moreover illustrate how computational methods are able to reveal the most likely forces causing varying cultural patterns.

Future work in computational analysis of cultural assimilation may incorporate additional predictors such as urbanization, politics, and internal migration. A metric could also be developed to account for geographic area when analyzing the number of languages and religious diversity of a population. Nevertheless, our findings demonstrate that computational modeling can uncover the deeper structural forces shaping cultural patterns and challenge assumptions about globalization and cultural assimilation.

Beyond these findings, this study highlights a broader insight on analyzing cultural dynamics: that quantitative and statistical approaches are able to reveal when widely assumed cultural relationships lack empirical evidence. Globalization is often portrayed as an indomitable force which produces uniformity, yet the results suggest that contemporary demographic and economic forces exert much less influence on cultural assimilation and diversity than what is commonly believed. In this manner, the study reinforces the importance of integrating computational modeling into various fields of study, including political and social science. Computational science allows researchers to revisit longstanding assumptions, detect hidden causal forces, and ultimately challenge assumptions which are only supported by plausibility. Thus, understanding assimilation in a data-driven perspective not only provides insight on present affairs but also provides a foundation for predicting how populations might evolve under future demographic or geopolitical changes.

Acknowledgments

The author thanks Mr. Robert Gotwals for assistance with this work.

References

[1] Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg. “Fractionalization.” Journal of Economic Growth, vol. 8, no. 2, 2003, pp. 155–194.

[2] Fearon, James D. “Ethnic and Cultural Diversity by Country.” Journal of Economic Growth, vol. 8, no. 2, 2003, pp. 195–222.

[3] Guiso, Luigi, Paola Sapienza, and Luigi Zingales. “Does Culture Affect Economic

Outcomes?" Journal of Economic Perspectives, vol. 20, no. 2, 2006, pp. 23–48.

[4] Berry, John W. "Immigration, Acculturation, and Adaptation." Applied Psychology: An International Review, vol. 46, no. 1, 1997, pp. 5–34.

[5] Hobsbawm, Eric. *Nations and Nationalism Since 1780: Programme, Myth, Reality*. Cambridge University Press, 1990.

[6] Wolfram Research, Inc. *Wolfram Mathematica, Version 14.3*. Champaign, IL: Wolfram Research, Inc., 2024.

Figures and Tables

The following code was used for data acquisition, cleaning, transformation, visualization, and model fitting. All computations were conducted in Wolfram Mathematica 14.3.

Data Acquisition and Cleaning

```
(* Gets all countries *)
allCountries = CountryData[];

(* Economic and migration information *)
getEconMig[ct_] := <|"Country" -> ct, "Name" -> CountryData[ct, "Name"],
    "GDPPerCapita" -> CountryData[ct, "GDPPerCapita"],
   "TradeValue" -> CountryData[ct, "TradeValueAdded"],
   "MigrationRate" -> CountryData[ct, "MigrationRateFraction"],
   "Population" -> CountryData[ct, "Population"],
   "Area" -> CountryData[ct, "Area"]|>;

(* Cultural information: languages and religions *)
getCultural[ct_] := Module[
   {languages = CountryData[ct, "Languages"],
    religions = CountryData[ct, "Religions"]},
   <|"NumLanguages" -> Length[languages],
    "NumReligions" -> Length[religions]|>];

(* Combine economic/migration and cultural data for each country *)
combinedData =
  Table[Join[getEconMig[ct], getCultural[ct]], {ct, allCountries}];

(*Remove entries with missing core values*)
keys = {"GDPPerCapita", "TradeValue", "MigrationRate", "NumLanguages",
    "NumReligions", "Population", "Area"};

cleanData = Select[combinedData, FreeQ[Lookup[#, keys], _Missing] &];

dataset = Dataset[cleanData];
```

Log Transformation

```
(* Function to get magnitude given a string *)
num[x_] := If[QuantityQ[x], QuantityMagnitude[x], x];

(* Add log-transformed variables to each country association *)
```

```
modelData = Map[Function[assoc, Module[
     {
      gdp = num[assoc["GDPPerCapita"]], trd = num[assoc["TradeValue"]],
      pop = num[assoc["Population"]],
      area = num[assoc["Area"]],
      mig = assoc["MigrationRate"]
      },
     Association[assoc,
      "LogGDPPerCapita" -> If[gdp > 0, Log[gdp], Missing["NotPositive"]],
       "LogTradeValue" ->
       If[trd >= 0, Log[1 + trd], Missing["NotPositive"]],
      "LogPopulation" -> If[pop > 0, Log[pop], Missing["NotPositive"]],
       "LogArea" -> If[area > 0, Log[area], Missing["NotPositive"]],
      "MigrationRateNumber" -> num[mig]]
     ]
    ]
   , cleanData];

modelDataset = Dataset[modelData];
```

Religion Fractionalization Index Calculation

```
(* Religious fractionalization index calculation *)
relFractionalization[ct_] := Module[
   {fracs, vals, norm},
   (*Fractions are usually rules:religion->fraction*)
   fracs = CountryData[ct, "ReligionsFractions"];
   (* Checks if data is not available *)
   If[fracs === Missing["NotAvailable"] || fracs === {} ||
     fracs === Null,
    Return[Missing["NotAvailable"]];
    ];
   (* Extract numeric fractions and clean data *)
   vals = Cases[fracs, (_ -> p_?NumericQ) :> p, Infinity];
   vals = DeleteMissing[vals];
   (* No usable fractions *)
   If[vals === {}, Return[Missing["NotAvailable"]];];
   (* Ensures 0 diversity for 1 religion *)
   If[Length[vals] == 1, Return[0];];
   (*Normalize and compute RFI *)
   norm = vals/Total[vals];
   1 - Total[norm^2]];
```

```
(* Add religious fractionalization index to data *)
dataWithRFI =
  Map[Function[assoc,
    Association[assoc,
     "RelFracIndex" -> relFractionalization[assoc["Country"]]]],
   modelData];

dataRFI = Dataset[dataWithRFI];
```

Scatterplot Generation

```
(* Points for migration vs. number of languages *)
migLangPoints =
  Cases[dataWithRFI,
   KeyValuePattern[{"MigrationRateNumber" -> mr_?NumericQ,
      "NumLanguages" -> nl_?NumericQ}] :> {mr, nl}];

ListPlot[migLangPoints,
 FrameLabel -> {"Migration rate", "Number of languages"},
 PlotRange -> All, Frame -> True]

(*Migration vs religious fractionalization*)
ptsMigRFI =
  Cases[dataWithRFI,
   KeyValuePattern[{"MigrationRateNumber" -> mr_?NumericQ,
      "RelFracIndex" -> rfi_?NumericQ}] :> {mr, rfi}];

ListPlot[ptsMigRFI,
 FrameLabel -> {"Migration rate", "Religious fractionalization"},
 PlotRange -> All, Frame -> True]
```

Regression Models

```
(* Data for language diversity model *)
lingData =
  Cases[dataWithRFI,
   KeyValuePattern[{"NumLanguages" -> nl_?NumericQ,
      "MigrationRateNumber" -> mr_?NumericQ,
      "LogGDPPerCapita" -> lgdp_?NumericQ,
      "LogTradeValue" -> ltv_?NumericQ,
      "LogPopulation" -> lpop_?NumericQ,
      "LogArea" -> la_?NumericQ}]
```

```
    :> {nl, mr, lgdp, ltv, lpop, la}];

lmLang =
  LinearModelFit[
   lingData, {mr, lgdp, ltv, lpop, la}, {mr, lgdp, ltv, lpop, la}];

lmLang["ParameterTable"]

rSquared1 = lmLang["RSquared"];

Print[StringJoin["R Squared: ", ToString[rSquared1]]]


(* Data for religious fractionalization model *)
relFracData =
  Cases[dataWithRFI,
   KeyValuePattern[{"RelFracIndex" -> rfi_?NumericQ,
      "MigrationRateNumber" -> mr_?NumericQ,
      "LogGDPPerCapita" -> lgdp_?NumericQ,
      "LogTradeValue" -> ltv_?NumericQ,
      "LogPopulation" -> lpop_?NumericQ,
      "LogArea" -> la_?NumericQ}] :> {rfi, mr, lgdp, ltv, lpop, la}];

lmRFI = LinearModelFit[
   relFracData, {mr, lgdp, ltv, lpop, la}, {mr, lgdp, ltv, lpop, la}];

lmRFI["ParameterTable"]

rSquared2 = lmRFI["RSquared"];
Print[StringJoin["R Squared: ", ToString[rSquared2]]]
```