Global Migration Determinants: Development Indicators that Motivate Migration

Haylee Ham

Abstract *In this study, I empirically test the determinants of international migration presented in three major migration theories: the neoclassical model stating that potential income and probability of securing employment in a new country are the major determinants of migration (Harris and Todaro, 1970), a study from Hugo (1998) stating that demographic shifts borne of changing fertility rates and aging populations drive migration to countries with a lack of available labor, and a theory that posits that the strength of a country's education system is an important determinant for families looking to migrate (Adepoju, 2010). In testing these theories, I use a new global bilateral data set that gives counts of total migration flows between every pair of 196 countries. I build a generalized additive model using the factors from these three theories and several controls and fixed effects in order to measure the effect that each factor has on migration. I find that the two factors presented in the Harris-Todaro model are the only significant motivators of migration in the generalized additive model and that the factors presented by Hugo and Adepoju do not appear to hold their significance in the presence of the Harris-Todaro factors.*

I. Introduction

In 2015, there were 244 million total people residing in a country other than their country of birth, the highest ever recorded, according to the International Organization for Migration. Scholars have formally attempted to uncover the causes for the relocation of vast amounts of people for over a century. While various theories and models have been developed and tested on subsections of the global migrant population, there has been a lack of scholarship concerning the entire population of migrants and the mechanism behind their choices. This lack of study is largely due to the past absence of global bilateral migration data sets, which consist of the migration flows between all combinations of origin and destination countries in the world. With the recent publication of these datasets, comparisons between individual country migration flows can be made in earnest. Using this newly available data along with country characteristics measured at the origin and destination country level, this study will explore the motivations behind the selection of destination countries by migrants. Specifically, this study will test the validity of the neoclassical Harris-Todaro theory of migration, the fertility-demographic shift theory presented by Hugo, and Adepoju's educational system theory. These are three main studies that present major social and economic characteristics that are often cited as reasons why migrants leave their country of origin and choose their country of destination. I hypothesize that the Harris-Todaro theory presents factors that will be the most significant indicators of migration.

Using a generalized additive model that combines the factors from each of these theories, I find that the two determinants presented in the Harris-Todaro theory, potential income and probability of obtaining employment, have the most significant effect on migration. The factors presented by Hugo, fertility rate and age demographic shifts, and Adepoju, the strength of the educational system in the origin country, were not significant drivers of migration.

This study is the first to use bilateral data to test major motivations of international migration on a global scale. Past studies have focused on only groups of destination or origin countries (Borjas, 1987; Adepoju, 2000) or have used global data to identify popular destination regions (Özden, 2011; Abel and Sander, 2014), rather than parsing out underlying motivations based on country characteristics.

II. Theory

Previous research in the area of migration determinants provides a framework of theory and empirical observations for this study. Literature in this field has focused on the neoclassical models of individual optimization and factors such as governmental amenities, border control, demographic makeup, geographic considerations, and social and network effects.

This study will particularly test the validity of three major migration theories at a global level. First, the early neoclassical theory put forth by Harris and Todaro (1970). This theory takes into account only the optimization of utility by the migrant, ignoring social and cultural factors. The neoclassical Harris-Todaro model states that a migrant will select a destination country by maximizing his or her income. The migrant considers his or her potential income in the destination country and the probability of securing employment as the main factors when selecting where to migrate.

Succeeding neoclassical thought, scholars began to look into migratory patterns and the characteristics of destination countries other than simply potential earning power. The economics of migration began to take into account political, social, geographic, and cultural factors. The second theory that will be tested considered how shifts in the demographic makeup of both origin and destination countries affect migration. This theory from Hugo (1998) stated that fertility rates and the age of a country's demographic play a large part in destination determination for migrants. As fertility rates slow in countries with growing economies, populations age and a shortage of working-age individuals appears in the country. Individuals from less developed countries with steady fertility rates, and therefore a shortage of jobs for those entering the workforce, will move from their origin country into these low fertility rate

countries. Hugo found this theory to be true in the context of migration within Asia between 1970 and 1990.

The third theory to be tested concerns the amenity of educational opportunities in the destination country. Adepoju (2000) found that families in Sub-Saharan Africa send children to a country with a strong education program so that the children can secure a higher wage and support the family through remittances. The pressure to choose a country with a very developed educational system outweighed many other concerns, with education being viewed as an investment in future income. Some in Sub-Saharan Africa have moved their entire family in search of a better education for their children, especially after a collapse of the education system in their origin country, as Adepoju points out was the case in Nigeria (1995b).

The three theories described above were either developed as theoretical models (in the case of the Harris-Todaro model) or empirically tested on a subset of the global population (Asia, in the case of Hugo, and Sub-Saharan Africa, in the case of Adepoju). Indeed, most research in the area of migration has not been conducted at the global level, due to a paucity of comparable data across countries. This study will be able to test the three theories above at a global level for the first time and measure their accuracy for the world population, rather than on only a geographic subset of the earth.

Newly available cross-country migration data allows researchers to delve into the motivations behind international migration such as was not possible empirically before. There have been several studies that have attempted to test the above theories and their associated factors on an international level, although most have only been able to do so with limited scope. For instance, the paper by Mayda (2010) uses data from 14 OECD countries from 1980 to 1995. Mayda studied migration push and pull effects proxied economically using the per worker GDP in both the origin and destination countries (the same proxy I will use for potential income in the Harris-Todaro model). While, according to the theory of the international migration model, a strong push factor such as low per worker GDP in the origin country should cause high migration, Mayda found that this was not the case. She supposed that the social barrier of poverty caused by the low per worker GDP was enough to suppress migration, as individuals have neither the skills nor money to alter their country of residence. I will include Mayda's findings of the nonlinear relationship per worker GDP has with migration in my model.

III. Data

As stated, global bilateral migration data has been sparse until recently and therefore studies at a global scale have been limited. This paper will use the recently published dataset from the Wittgenstein Centre for Demography and Global Human Capital. This dataset contains

global bilateral flows between 196 countries from 1990 to 2010 in five-year increments. The authors of this dataset note that “migration flow data is often incomplete and not comparable across nations”. In order to overcome this, Abel and Sander have linked changes in migrant stock tables over time and used imputation to estimate the flows that are required to meet those measured changes. Abel and Sander have created a robust and novel dataset that is comparable across countries and accurately captures migrant origins and destinations for those who have permanently changed their country of residence over five-year periods (Abel and Sander, 2014).

Abel and Sander published this global bilateral migration dataset along with findings about the migration trends overtime from 1990 to 2010. The analysis in this paper did not include characteristics from the countries apart from their geographic location when identifying trends. In order to observe the changing preferences of the migrant population overtime, this paper will make use of the dataset published by Abel and Sander and also include development and economic indicators in order to test the validity of theories of migrant preferences at a global level that have been posited in the previous research discussed above.

The format of the data is dyadic, wherein each country is both the origin and destination country for every other country in a row of the data. Each row is thus a pair of countries, one destination and one origin, and the number of migrants who permanently changed their residence over a five-year period from the origin to the destination country in that row. In its raw form, the this dataset has a four variables for migration flows between all possible pairs of countries, one for each five-year period. For the purposes of this study, which is concerned with determinants of migration rather than time-series effects, I have reshaped the data so that all migration flow counts occur within one column and a separate column to indicate the five year interval has been appended. World development indicators from the World Bank have also been added to each row of the data. The world development indicators are observed yearly and have been averaged into the same five-year increments present in the bilateral migration flows dataset. The averages for all countries have then been matched to the bilateral migration flows dataset so that each migration flow observation has both origin and destination country characteristics associated with it. Finally, the control variables euclidean distance between each pair of countries and a boolean indicating common language were extracted from a dataset curated by the French research center CEPII and added to the dataset. The common language variable takes a value of 1 if at least 4% of the population in each country speaks the same language. (The links for each of these datasets can be found in the appendix.) This results in a dataset with 77,840 observations of bilateral migration flows and 7 characteristics per country for each observation.

The world development indicators used for this study are per worker GDP, labor participation rate, fertility rate, working age rate, and primary enrollment rate. Per worker GDP is the GDP for a country divided by total employment and measured in constant 2011 purchasing

power parity (PPP) international dollars, which has the same PPP as U.S. dollars in the United States. Labor participation rate is the percentage of the population ages 15 and older that is economically active, including those who are employed only part-time and those who are unemployed for a temporary reason. Fertility rate is measured as the number of children that a woman would birth if she were to live until the end of her childbearing years and birth the number of children indicated by age-specific fertility rates for the time period. These rates are developed by using registered live births or census data. Working age rate is a measure of the percentage of the population that is between 15 and 64 years old. This age range is generally considered to be of working age. Primary enrollment rate is the percentage of primary school age children enrolled in school.

In testing the factors within the Harris-Todaro model, per worker GDP will be used as a proxy for potential income in a given country and labor participation rate will be used to measure the probability of securing employment in a country. For the second theory concerning demographic shifts, fertility rate will be used directly and working age rate will be used as a proxy to measure the age of the population in a given country. Finally, primary enrollment rate will be used to measure the strength of a country's education system. These variables used to estimate the validity of the three theories are summarized in the table below.

Table 1: Summary Statistics for Key Variables

Variable	Min	Median	Mean	Max
Migration flow (count)	0	0	1440	2,680,000
Fertility rate (count)	1.15	2.82	3.37	7.75
Labor participation rate (%)	32	58	58	88
Per worker GDP ($)	1068	20,200	34,708	205,327
Primary enrollment rate (%)	22	94	86	100
Working age rate (%)	47.9	61.7	60.7	83.8

Per worker GDP is of special interest concerning its importance in the theory that is hypothesized to be the best indicator of migration. In order to graphically show the the movement of migrants in terms of the per worker GDP in origin and destination countries, I have created the following circle plot.

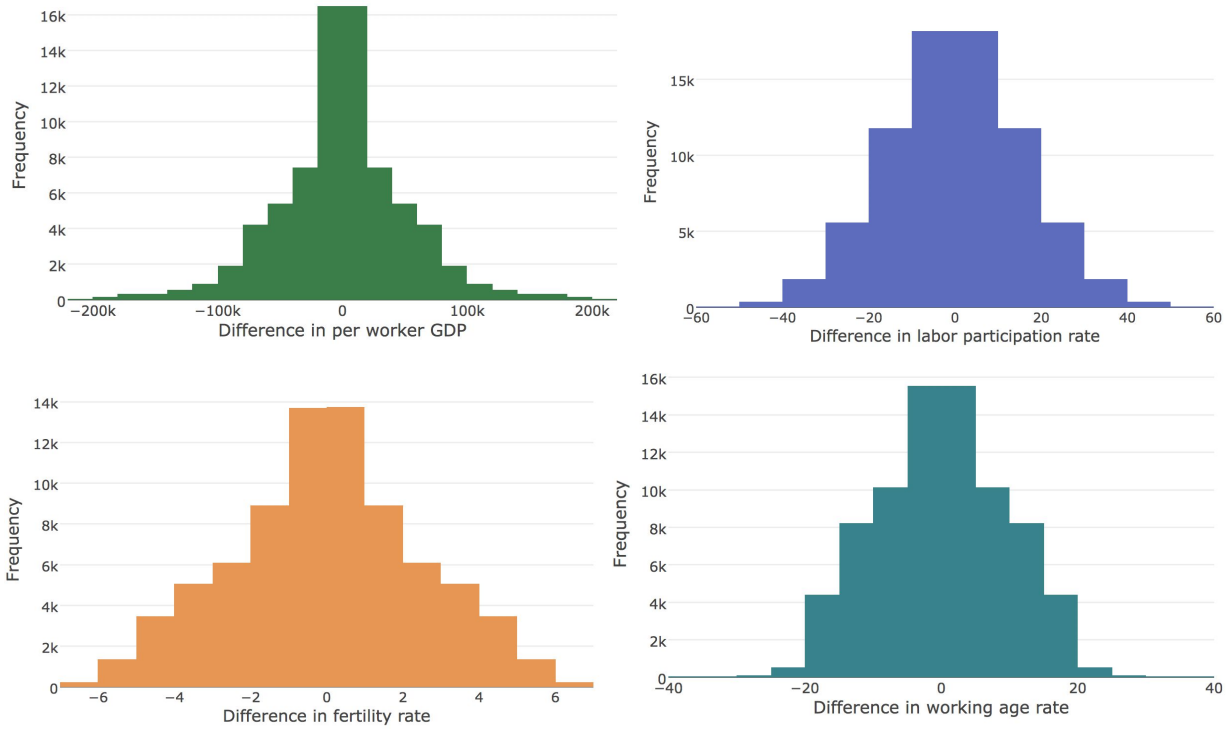
Figure 1: Migration in terms of per worker GDP quartiles

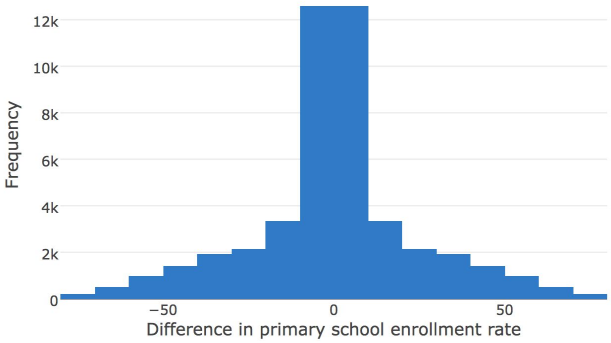
The circle plot can be read by following the flows of migrants from the place of origin to the place of destination. The color along the border of each quartile of the per worker GDP variable indicates the origin. Flows stemming from the very edge of the graph and that are the same color as their border are representative of migrants originating in that quartile and migrating to the quartile where the flow stops slightly further away from the border. From this

plot, it can be seen that migrants from the top quartile mostly migrate to other countries within the top quartile. The vast majority of migrants from the second quartile also migrate to countries within in the first quartile. Migrants from the third quartile are spread quite evenly among all four quartiles in their choice of destination. The fourth quartile sends roughly the same proportion of migrants to other countries within the fourth quartile and to countries within the first quartile.

For the purposes of this study, a simple difference of the values of the development indicators between destination and origin country will be taken for each migration flow observation. A positive difference indicates that the value is higher in the destination country than the origin country. These differences will allow the net effect of each of these indicators to be observed between all possible country pairings. Once the difference is performed, the development indicators are distributed symmetrically, with some distributions approximating the normal distribution and others showing higher peaks and heavier tails.

Figure 2: Histograms of World Development Indicator Differences





There exists missing data in several of the independent variables used for this study. Table 2 shows the variables that have missing values.

Table 2: Missingness of Variables

Variable	Number missing	Percent missing
Per worker GDP	2,216	2.8%
Labor participation rate	2,216	2.8%
Primary enrollment rate	31,558	40.5%

In order to account for this missingness, I considered whether the data is missing at random. While the data for the primary enrollment rate variable appears to be approximately missing at random, per worker GDP and labor participation rate appear to be missing based on region. The missingness for these two variables is clustered in Latin America. I employed the Bayesian multiple imputation method in order to impute the data and account for this missingness. I used world development indicators that were closely correlated with the variables with missing data, such as literacy rate and proportion employed, in order to build a strong predictive model to fill in the missingness. Five complete datasets were created by the predictive model and then used to estimate a linear version of the model. With the averaged coefficients and standard errors from the multiple imputations method, I noticed very little difference between these results and the results obtained from computing a linear model from the data obtained using the listwise deletion method. Due to the similarity of the results and the rigidity of the methods that can be used on the datasets resulting from multiple imputation, I moved forward with the listwise deletion version of the data. This final dataset contains 44,912 observations and a simple difference has been taken for all world development indicator variables between the values in the origin and destination country. The summary statistics are shown below.

Table 3: Summary Statistics for Key Variables from the Listwise Deletion Dataset

Variable	Min	Median	Mean	Max
Migration flow (count)	0	0	1660	2,680,000
Fertility rate (count) (difference)	-6.58	0	0	6.58
Labor participation rate (%) (difference)	-54.3	0	0	54.3
Per worker GDP ($) (difference)	-204,000	0	0	204,000
Primary enrollment rate (%) (difference)	-77.7	0	0	77.7
Working age rate (%) (difference)	-35.9	0	0	35.9

IV. Computational Results

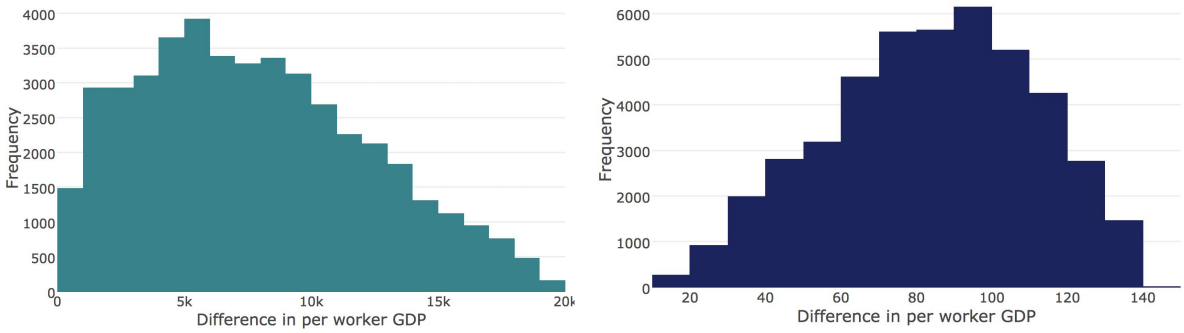
In order to explain the variation in migration flows between countries, I employed a generalized additive model with count of migrants between each pairing of countries as the dependent variable. As independent variables, the model contains differences of the world development indicators between the destination and origin country of each migration flow. Specifically, per worker GDP and labor participation rate are used to measure the neoclassical theory of migration introduced by Harris and Todaro. Per worker GDP, as theorized by Mayda, has a nonlinear relationship with migration flow. The expectation is that a discrepancy between per worker GDP in two countries increases migration to the country with a higher per worker GDP until a certain point, after which migration will fall. For this reason, I have fitted a 5-knot spline to the estimation of the per worker GDP variable's relationship with the dependent variable.

Fertility rate and working age rate are the variables used in this study to approximate the effect of the demographic shift theory posited by Hugo. Primary school enrollment rate will be used as a measure of the current strength of the education system in a country, an amenity which families consider very important when considering migrating. No transformations were needed with these three variables; they are all approximately normally distributed and expected to have a linear relationship with the dependent variable.

The model will control for the Euclidean distance between the origin and destination countries as well as a variable indicating whether the origin and destination country have a

common language. The distance variable was found to be right-skewed and therefore non-normally distributed. As a result, a square root transformation was applied to help somewhat normalize the variable. Figure 3 shows the distance variable's distribution before transformation on the left and the distribution after the square root transformation can be seen on the right.

Figure 3: Histograms of Distance Variable before and after square root transformation



Finally, the model controls for all 196 origin country fixed effects and all four time period fixed effects using dummy variables. Since the independent variables are measures of the difference between the origin and destination country, controlling for fixed effects in the origin country will not control for the variation present in the differences derived from the indicator variables, it will however control for origin country shocks that would drive migration. The following is the generalized additive model used for this study:

$$migration\_flow_{i,j} = \beta_0 + f(per\_worker\_GDP_{i,j}) + \beta_1\, labor\_participation\_rate_{i,j} + \beta_2\, fertility\_rate_{i,j} + \beta_3\, working\_age\_rate_{i,j} + \beta_4\, primary\_enrollment\_rate_{i,j} + \beta_5 sqrt(distance_{i,j}) + \beta_6\, common\_language_{i,j} + country\_fixed\_effects_i + time\_fixed\_effects_t + \varepsilon$$

As an initial step, I performed simple linear factor regressions to test the individual effects each of the indicators has on migration flows. In these regressions, the migration flow was the dependent variable and each indicator became the independent variable. In order to isolate the effect of the indicator, I also controlled for origin country and time fixed effects. The results of the factor regressions are presented in Table 4 below.

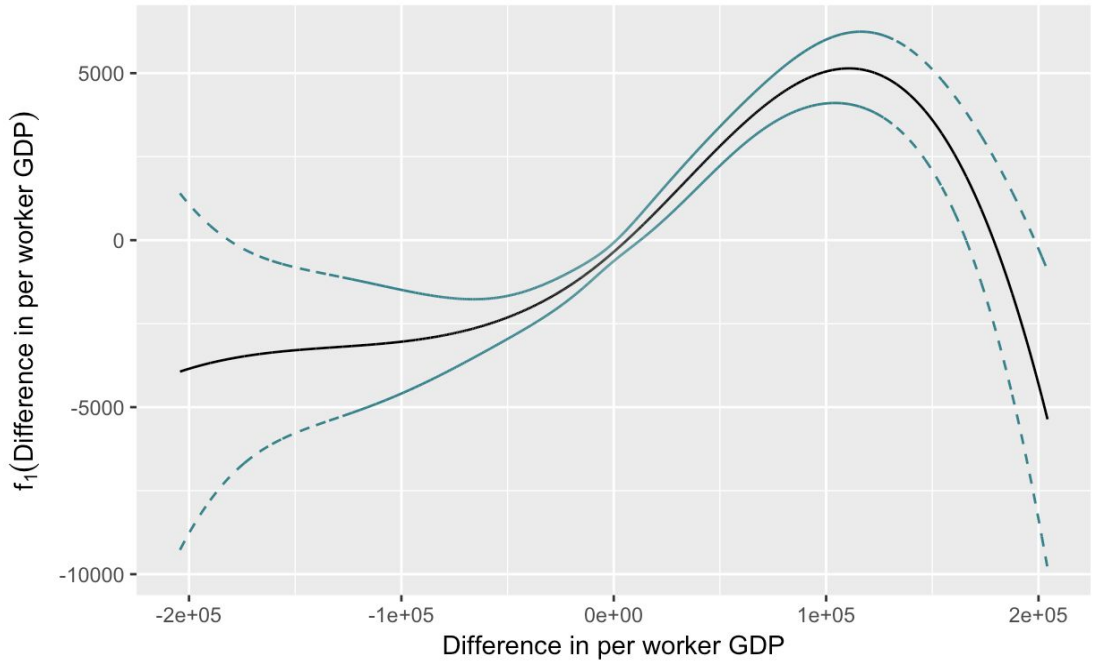
Table 4: Factor Regressions of World Development Indicators

Variable	Estimate	Std. Error	p-value
Fertility rate difference	-590.049	6.90e+01	1.24e-17*
Labor participation rate difference	17.999	1.04e+01	8.23e-02
Per worker GDP difference	4.22e-02	3.03e-03	4.46e-44*
Primary enrollment rate difference	50.887	6.55	7.85e-15*
Working age rate difference	179.003	1.73e+01	5.65e-25*

** Significant at a 0.001 level*

From the table above, it can be seen that fertility rate, per worker GDP, primary enrollment rate, and working age rate are all significant at at least a 0.001 level. Except for labor participation rate, each of the variables appears to be significant in their own right when not in a model with any other explanatory variables. However, factor regressions only estimate the effect of each variable when not accounting for other theorized predictors of migration. By estimating the generalized additive model, the relative effect of each factor can be assessed. After estimating the generalized additive model, the relationship of per worker GDP to migration is fitted by a cubic spline, as shown in Figure 4.

Figure 4: Marginal Effect of Per Worker GDP on Migration



The above spline fit of the per worker GDP variable shows that migration between two countries increases rapidly as the per worker GDP in the destination country becomes increasingly greater than the per worker GDP in the origin country. After the per worker GDP in the destination country is approximately $100,000 greater than in the origin country, migration falls sharply. This graph supports the theory presented by Mayda that migration will increase to countries with higher per worker GDP differentials until a critical point where the social barrier of poverty overcomes a migrant's ability to emigrate. The spline graph suggests that that critical point is when the differential hits approximately $100,000.

The other independent variables were all estimated parametrically and their estimates can be found in Table 5 below. Additionally, a coefficient plot for the parameter estimates can be found in the appendix. While the control variables of distance and common language are strongly significant, the only explanatory variable that is significant is the labor participation rate difference. Interestingly, this is the variable that was not significant in the factor regressions. This seems to indicate that when the labor participation rate is included in a larger model, another regressor absorbs some of the residual variability present in the dependent variable and therefore lends the labor participation rate greater statistical significance.

Table 5: Parameter Estimates

Variable	Estimate	Std. Error	p-value
Labor participation difference	58.83	11.45	2.8e-07*
Fertility rate difference	5.53	219.81	0.979
Working age rate difference	29.22	51.82	0.573
Primary enrollment rate difference	12.23	10.08	0.225
Distance, kilometers (square root)	-65.38	4.95	<2e-16*
Common language	1750.36	374.16	2.9e-06*

** Significant at a 0.001 level*

Overall, it appears that when the regressors from each of the three theories are included in the generalized additive model, the only regressors that are significant are those stipulated in the Harris-Todaro theory: per worker GDP and labor participation rate. Labor participation rate has a positive relationship with migration. For each one percentage point increase in the labor participation rate, migration to the destination country increases by 58.83 migrants. The factors found by Hugo to be of most importance within Asia, fertility rate and an age demographic shift

(proxied by the working age rate), were not significant predictors of migration once estimated within a model that included per worker GDP and labor participation rate. Similarly, Adepoju's findings within Sub-Saharan Africa that the strength of the education system in the destination country (proxied by primary enrollment rate) was a strong predictor is not supported by this study.

In addition to per worker GDP and labor participation rate, the control variables distance and common language are shown to be very important indicators of migration destination selection, with migration decreasing as distance increases and a substantial increase in migration occurring between countries with a common language.

V. Conclusion

In testing the validity of three major migration theories at a global scale, this study found that the factors introduced in the Harris-Todaro model were the only significant indicators of migration. Higher per worker GDPs and labor participation rates in a country both drive migration to that country and were used as proxies for potential income that a migrant might make and the probability of securing employment in a new country, respectively. These two factors were significant in a model that also included the factors presented in the Hugo study concerning demographic shifts and the Adepoju study concerning strength of a country's educational system. The fertility rate and the rate of individuals who are between 15 and 65 (generally considered to be of working age) in a country were used to measure the notion that countries with a surplus of labor will migrate to countries that are experiencing a paucity of available labor. Both fertility rate and working age rate were found to be insignificant predictors of migration once included in a model with the more significant regressors from the Harris-Todaro model. The variable used as a proxy to measure the strength of a country's education system, the rate of enrollment of appropriately-aged children in primary school, was also found to be insignificant in the presence of the factors presented by the Harris-Todaro model.

In future development of this work, I would like to include more controls for political and social factors such as how tightly controlled a country's borders are, which countries have strong trade relationships, how generous a country's welfare program is, and the size of the diaspora from the origin country that has already migrated to the destination country (measured more directly than the common language boolean). These factors would allow me further test migration theories at a global level and determine whether the results presented in this study suffer from omitted-variable bias. I would also like to search for proxies with less missingness and find a method more robust than listwise deletion for dealing with missingness that will allow

for models with more complicated functional forms (which I was unable to accomplish with the statistical software packages *Amelia* and *Mice*).

In conclusion, this study has furthered the work of identifying major determinants of international migration at a global scale. For the first time, a global bilateral migration dataset was used to test the validity of migration determinants that, up to this point, had only been tested using a limited scope. With the emergence of these novel global bilateral datasets, inferential studies and predictive models alike can be developed at a global level to assist countries and international organizations to better understand and prepare for changing migration trends.

References

Abel, Guy J., and Nikola Sander. “Quantifying Global International Migration Flows.” *Science*, vol. 343, no. 6178, 2014, pp. 1520-1522., http://science.sciencemag.org/content/343/6178/1520.

Adepoju, A. "Issues and Recent Trends in International Migration in Sub-Saharan Africa." *International Social Science Journal*, vol. 52, no. 165, 2000, pp. 383-394.

Adepoju, A. “Emigration dynamics in Sub-Saharan Africa.” International Migration Special Issue: Emigration Dynamics in Developing Countries, vol. 33, nos3/4, 1995b.

Borjas, George J. “Self-Selection and the Earnings of Immigrants.” *The American Economic Review*, vol. 77, no. 4, 1987, pp. 531–553., www.jstor.org/stable/1814529.

Harris, John R., and Michael P. Todaro. “Migration, Unemployment and Development: A Two-Sector Analysis.” *The American Economic Review*, vol. 60, no. 1, 1970, pp. 126–142., www.jstor.org/stable/1807860.

Hugo, G. "The Demographic Underpinnings of Current and Future International Migration in Asia." *Asian and Pacific Migration Journal*, vol. 7, no. 1, 1998, pp. 1-25.

Mayda, A. M. "International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows." *Journal of Population Economics*, vol. 23, no. 4, 2010, pp. 1249-1274, doi:10.1007/s00148-009-0251-x.

Özden, C., et al. "Where on Earth is Everybody? the Evolution of Global Bilateral Migration 1960-2000." *World Bank Economic Review*, vol. 25, no. 1, 2011, pp. 12-56, doi:10.1093/wber/lhr024.

Appendix

The code used to clean the data and the cleaned dataset used for this study are available at the following link: https://github.com/hayleefay/Migration-Trends

The bilateral migration flows data set can be accessed on the following website: http://www.global-migration.info/.

The world development indicators can be accessed at the World Bank's website: http://data.worldbank.org/data-catalog/world-development-indicators

The distances between countries and common language data can be accessed at CPEII's website: http://www.cepii.fr/cepii/en/bdd_modele/bdd.asp

The code used to make the D3 Circle Plot can be found at the following link: https://github.com/null2/globalmigration

Additional Figures

Coefficient Plot for Parameter Estimates