Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1080738
Assignment	Applied Analytical Statistics
Term	Michaelmas Term 2023
Title/Question	Bound by faith? Exploring the association between religious proximity and forced migrant count in a global sample
Word Count	5000

By placing a tick in this box ✓ I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.

Bound by faith? Exploring the association between religious proximity and forced migrant count in a global sample

1080738

Abstract

This paper investigates the association between religious proximity and forced migrant flows. Using a zero-inflated negative binomial regression model on an imputed sample of 3724 unique origin-destination dyads from 1977 to 2023, the study finds that a one-unit increase in religious proximity between two countries is associated with a 2.435 times increase in the count of forced migrants flowing between them. This result challenges existing scholarship, underscoring the need for further research to enhance the robustness of findings and deepen the understanding of religion as a potentially underestimated determinant in forced migrant flight patterns. The investigation responds to a critical context of global displacement, emphasising the urgency of evidence-based response infrastructures amid rising conflict, resource insecurity, and rapid globalisation. The results of this paper contribute to our understanding of what factors drive forced migrants go beyond their immediate neighbours in search of asylum.

Keywords: forced migration, religious proximity, zero-inflated negative binomial regression, demographic gravitation

1 Introduction

As of October 2023, there were 110 million forcibly displaced individuals around the world (UNHCR, 2023). From public health crises and devastating fires to disregard for human rights and safety, refugee camps evidence that current humanitarian infrastructures cannot cope (Laughon, Montalvo-Liendo, Eaton, & Bassett, 2023; Markham, 2022). As the world grapples with rising conflict (Nations, 2020), resource insecurity (Sofuoğlu & Ay, 2020) and rapid globalisation (Potrafke, 2015), forced migration is on the rise (Guo, Al Ariss, & Brewster, 2020). Thus, establishing stronger, evidence-based response infrastructures with reliable migration forecasting methods will be crucial (Pellandra & Henningsen, 2022).

Such forecasting methods rely on understanding what factors drive forced migration flight patterns. Traditionally, geographical distance has been considered the main determinant (Iqbal, 2007). And yet, scholars increasingly observe forced migrants traversing larger distances, beyond their immediate neighbours (Devictor, Do, & Levchenko, 2020; Neumayer, 2004). Figure 1 demonstrates this:

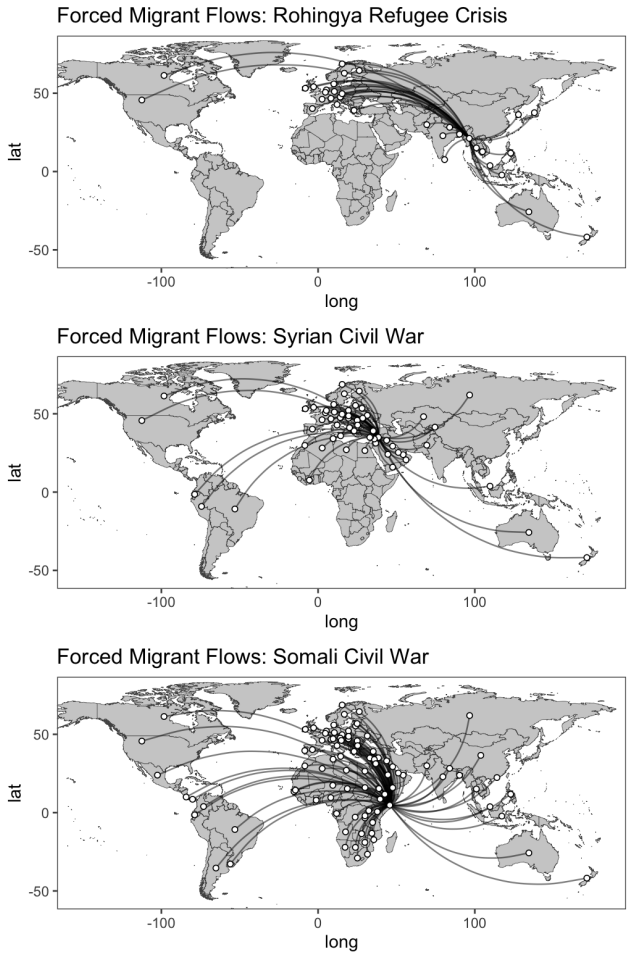

Figure 1: Diffusion of forced migrants in three historic refugee crises, showing that in all three cases forced migrants travelled way beyond their immediate neighbours.

Push-pull factor theory has been used to explain what drives migrants beyond contiguous destinations (Kang, 2020). Push factors, defined in this context as phenomena in the origin country that trigger outward migration, have been widely studied. For example, Hakovirta (1993) and Schmeidl (1997) associate violence and human rights abuses with forced migrant outflow volume. Others, such as Matsui and Raymer (2020) and Davenport et al. (2003) connect outflow volume to political terror, civil rights violations and threats to personal integrity, liberty or life. Outside political factors, scholars such as Van Hear et al. (2009) and Richmond (1993) argue that forced migrants are also pushed out by economic hardships and failures in the social system. Frameworks synthesising such diverse factors (Richmond, 1993) recognise that multiple drivers can co-exist in the decision process of asylum seeking.

On the other hand, pull factors are characteristics of destination countries which attract forced migrants. This is situated in the view of forced migrants as utility maximisers evaluating the net benefit of all possible asylum destinations (Neumayer, 2004). Scholars such as Kang (2020) claim that forced migrants are attracted to destinations with higher income and integration prospects. Similarly, Tucker (2018) finds that forced migrants consider prospects for citizenship in an effort to resolve statelessness, and others (Matsui & Raymer, 2020) find that policy features such as work bans and asylum claim recognition rates also play a role.

In reality, the interplay of origin and destination country characteristics are what determines flight patterns (Matsui & Raymer, 2020), while individual factors might influence micro-level migration decisions on an individual level. This concept holds explanatory power, as interactions between variables such as shared colonial history (Böcker & Havinga, 1997), shared language, or existing migrant diaspora networks (Neumayer, 2004), can be used to conceptualise a non-geographical proximity between countries in a latent dimension which makes forced migrant flows between them more likely (Xiao, Oppenheimer, He, & Mastrorillo, 2022). Such dyadic effects generally receive less attention in quantitative migration studies (Matsui & Raymer, 2020). This concept finds ground in Migration Systems Theory (MST) which views community feedback loops as primary drivers of destination choice (Leal & Harder, 2021).

A specific gap which emerges is whether religious ties can be used to approximate this non-geographical proximity. The Rohingya refugee crisis exemplifies the potential of religious ties as an explanatory factor, with persecuted Rohingya seeking asylum in contiguous Muslim-majority states such as Bangladesh, as well as further away Muslim-majority states such as Indonesia and Malaysia (Missbach & Stange, 2021). Scholars such as Neumayer (Neumayer, 2005) posit that forced migrants are attracted to destinations with religious kin because it reduces the cost of information, transport and integration by the same mechanisms attributed to ethnic and linguistic similarity. Figure 2 summarises this causal chain:

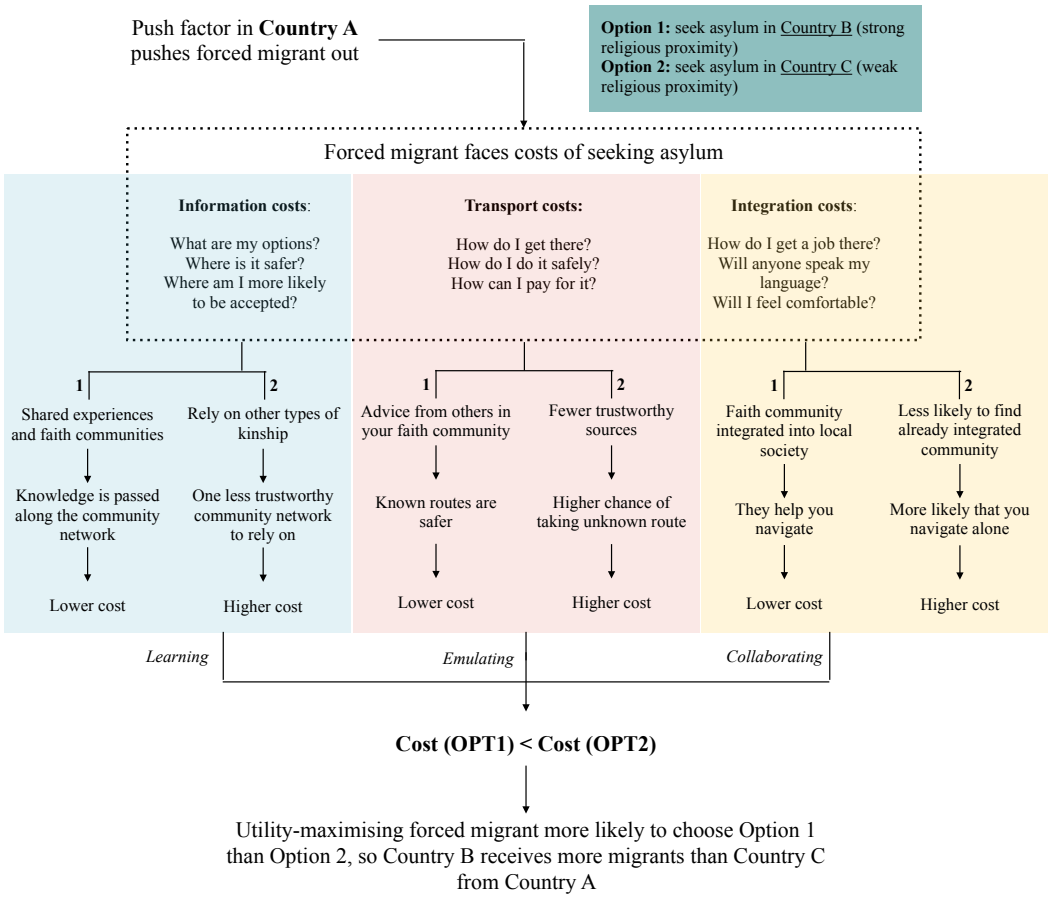

Figure 2: Causal chain explaining why a forced migrant from Country A is more likely to choose Country B (more religious kin) over Country C (less religious kin) via a cost reduction mechanism.

Investigating the impact of religious proximity on dyadic forced migrant flows stands to enrich our understanding of how forced migrants make decisions, empowering more reliable forecasting methods for strengthening humanitarian response frameworks. To this end, this paper asks the following question, and presents the following hypothesis based on the literature discussed above:

RQ: To what extent is religious proximity between two countries associated with the volume of forced migrant flows between them in a given point in time?

H1: There is a positive and statistically significant (95% confidence level) association between higher religious proximity and higher count of forced migrant flows for a given country-dyad-year.

The rest of this paper proceeds as follows. Section 2 will detail the data, modelling strategy, and robustness checks undertaken to investigate RQ. Section 3 will detail the results, notably a statistically significant positive association between religious proximity and forced migrant count. Section 4 will situate the findings in relevant literature, discussing strengths, limitations, contributions and areas for future research before concluding.

2 Methods

2.1 Data

Investigating RQ requires a dataset containing a measure of bilateral forced migrant flows from origin country i into destination country j in a given year t. The dataset must also contain a measure of religious proximity between i and j in t. Moreover, the dataset must contain a vector of covariates reflecting the push, pull and interaction factors discussed in Section 1, including geographical distance between i and j as well as political, economic, historical socio-cultural linkage covariates. As no such dataset exists, I construct one from a range of reputable sources. Table i (see Appendix A in section 6.1) contains a summary of all data sources used, which variables they provide, limitations as well as justifications for their inclusion.

Prior to analysis, I expand the dataset to cover a complete time series (1977-2023) for all unique origin-destination pairs, addressing asymmetry issues in the UNHCR data which could lead to bias (Marbach, 2018; Silvestrin, Pantiskas, & Hoogendoorn, 2021). The expanded dataset naturally contains NaN values, the structure of which is summarised in Figure i (see Appendix B in section 6.2). I use iterative imputation to fill these gaps, wherein incomplete data points are predicted using available information from other variables through regression models iterated until convergence. Some may express concern that iterative imputation distorts pre-set scales of bounded variables, however implementing set minima and maxima solves this issue. As iterative imputation yields values that maximise the consistency with observed data (Kyner, 2021), I proceed.

Furthermore, I standardise all covariates using min-max normalisation, or feature scaling via the following proceduce:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Despite concerns regarding feature scaling constraining covariates' ranges and possibly masking outliers by reducing standard deviations, I implement it as it maintains relationships among original observations and does not allow negative values (Loukas, 2020). The final dataset includes information on 3724 unique orig-dest dyads, each for 1977-2023 - a more detailed table of dataset descriptive statistics can be found in Appendix 6.1 (Table ii).

2.2 Model

I employ a Zero-Inflated Negative Binomial (ZINB) regression to investigate RQ. I choose a ZINB model for three reasons. Firstly, my dependent variable (dyadic forced migrant count) is a count variable, meaning that it is likely severely skewed and non-normally distributed, bounded between [0, Inf], and sparse, making an Ordinary Least Squares (OLS) model unfit due to assumption violations (Aiken, Mistler, Coxe, & West, 2015). Some may argue that log transformation can remedy these issues, however count data is a special case where log-transforms do not remedy neither non-normality of residuals or heteroscedasticity, resulting in biased, inconsistent count scores (O'Hara & Kotze, 2010). Thus, a count regression is required, as confirmed by Figure 3 which evidences high-skew and overdispersion in the dependent variable (dyadic forced migrant count):

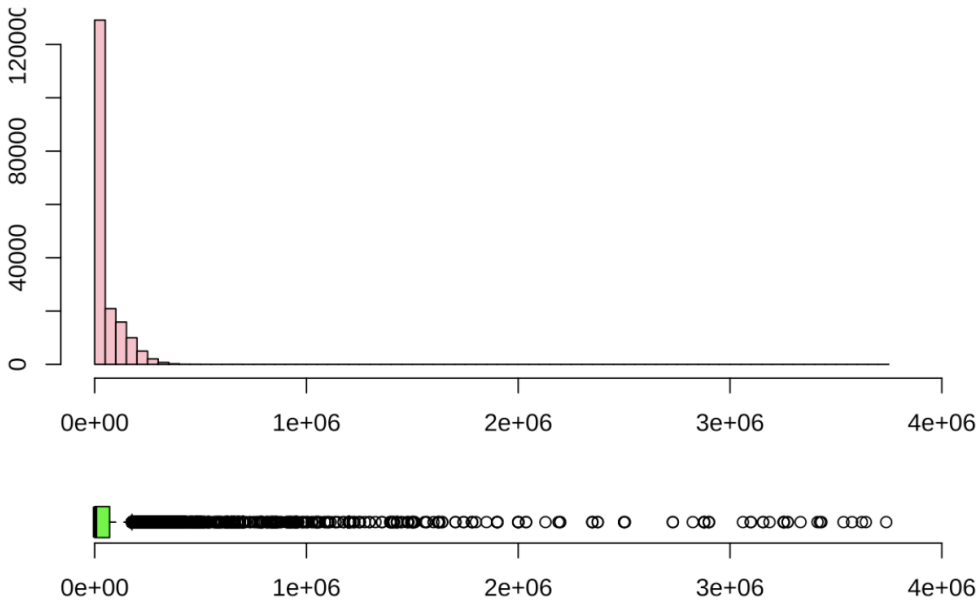

Figure 3: Combined histogram and box plot of dyadic forced migrant count, showing a severe right skew and overdispersion (variance bigger than mean).

Moreover, the high zero count as seen in Figure 3 is evidence of zero-inflation - a common problem with count data (Blasco-Moreno, Pérez-Casany, Puig, Morante, & Castells, 2019). It occurs when there are more zeros than anticipated due to a combination of *true* zeros (cases where forced migration actually did not occur) and *excess* zeros (additional zeros where the data collection process was flawed) (Heilbron, 1994). UNHCR's data collection process is prone to zero inflation (Marbach, 2018), likely due to conditions in the origin countries, such as ongoing conflict, interfering with research procedures (Deb & Baudais, 2022). Thus, the count regression must be able to deal with zero-inflation and overdispersion, for which a ZINB regression is best (Long, 1997) as it outperforms other count models, such as Poisson regression (He & Huang, 2023; Yau, Wang, & Lee, 2003).

ZINB regression combines two components: a negative binomial (NB) regression component to model the non-zero count data, and a binary logistic regression component to model excess zero occurrence. Thus, the expected count of forced migrants between i and j is:

$$E\left[\text{ flow }_{i,j,t}=k\right]=\underbrace{P(\text{ uncounted })*0}_{\text{excess zeros}}+\underbrace{P(\text{ counted })*E[y=k\mid\text{ counted }]}_{\text{true zeros when k=0}} \tag{2}$$

The NB regression component uses the NB2 parameterisation of the NB probability density function (NB-PDF), given by the following:

$$PDF_{\text{nb}}(y\mid r,p)=\begin{cases}0 & y<0\\ \frac{\Gamma(r+y)p^r(1-p)^y}{\Gamma(r)\Gamma(y+1)} & y\geqslant 0\end{cases} \tag{3}$$

Where r is the number of trials until the experiment is stopped, p is probability of success in each trial, and y is the outcome variable (Crowley, 2012). The NB-PDF is used to model the number of trials needed for r successes to occur in a sequence of independent and identically distributed Bernoulli trials (Long, 1997). The probability mass function (NB-PMF), expressed as such (Korosteleva, 2018):

$$\mathbb{P}(Y=y)=\begin{cases}\pi+(1-\pi)\left(\frac{r}{r+\lambda}\right)^r, & \text{if } y=0,\\ (1-\pi)\left(\frac{r}{r+\lambda}\right)^r\frac{\Gamma(r+y)}{y!\Gamma(r)}\left(\frac{\lambda}{r+\lambda}\right)^y, & \text{if } y=1,2,\ldots,\end{cases} \tag{4}$$

where r is a positive constant and

$$\begin{aligned}\pi&=\underbrace{\frac{\exp\{\beta_0+\beta_1x_1+\cdots+\beta_mx_m\}}{1+\exp\{\beta_0+\beta_1x_1+\cdots+\beta_mx_m\}}}_{\text{binary logistic regression component}}\\ \lambda&=\underbrace{\exp\{\gamma_0+\gamma_1x_{m+1}+\ldots+\gamma_{k-m}x_k\}}_{\text{NB regression component}}\end{aligned} \tag{5}$$

is used to compute the estimated parameters of a fitted ZINB regression as such:

$$\begin{aligned}\widehat{\pi}&=\frac{\exp\left\{\widehat{\beta}_0+\widehat{\beta}_1x_1+\cdots+\widehat{\beta}_mx_m\right\}}{1+\exp\left\{\widehat{\beta}_0+\widehat{\beta}_1x_1+\cdots+\widehat{\beta}_mx_m\right\}}\\ \widehat{\lambda}&=\exp\{\widehat{\gamma}_0+\widehat{\gamma}_1x_{m+1}+\cdots+\widehat{\gamma}_{k-m}x_k\}\end{aligned} \tag{6}$$

In the fitted form, the coefficients $\beta_0,\beta_1,...,\beta_m$ apply to the covariates of the logistic model, while the coefficients $\gamma_0,\gamma_1,...,\gamma_m$ apply to the NB model. Applied to my RQ, I calculate $\mathbb{P}(Y=y)$, where y is dyadic forced migrant count, with a logit model of the form

$$\text{logit}(\widehat{\pi}) = \beta_0 + \beta_1 \mathbf{X}_{i,t} + \beta_2 \mathbf{Y}_{j,t} + \beta_3 \mathbf{Z}_{i,j,t} + \beta_4 \mathbf{Q}_{i,j,t} + \epsilon_{i,j,t} \tag{7}$$

to estimate $\widehat{\pi}$ and a linear model with a log link function of the form

$$\log(\widehat{\lambda}) = \gamma_0 + \gamma_1 \mathbf{X}_{i,t} + \gamma_2 \mathbf{Y}_{j,t} + \gamma_3 \mathbf{Z}_{i,j,t} + \theta i, j, t \tag{8}$$

to estimate $\widehat{\lambda}$, where $\mathbf{X}$ contains *origin*-level covariates, where $\mathbf{Y}$ contains *destination*-level covariates, where $\mathbf{Z}$ contains *dyad*-level covariates, and where $\mathbf{Q}$ contains an additional binary triadic covariance *imputed* denoting whether the dyadic forced migrant count corresponding to the specific *orig-dest-year* triad has been imputed or not. Eac h covariate vector expands into its component variables, each with their own β and γ coefficients.

The specific covariates included in each of the four vectors are narrowed down using a Pearson correlation matrix to assess and filter out multicollinearity in my covariates, and Bayesian Model Averaging - a widely used method for feature selection (Wang, 2018). The rationale and procedure behind BMA, the Pearson correlation matrix, as well as the resultant chosen covariates, can be found in Appendix C (section 6.3).

2.3 Assumption and Robustness Checks

Before fitting the ZINB model, as an initial check that the my model improves upon a null model, I fit an intercept-only model and compare whether there is a statistically significant improvement in model fit through a Chi-squared Difference Test (Long, 1997).

To confirm that an OLS model is indeed unfit for my analysis, I fit a multivariate linear regression model and assess its residuals for any violations of the linearity and homoscedasticity assumption, which would make the conditions under which OLS estinated coefficients are the best linear unbiased estimates (BLUE) (Shaffer, 1991). Specifically, I use a residuals vs. fitted values plot and a Ramsey Regression Equation Specification Error test (RESET) to spot linearity violations, and a Breusch-Pagan test to check for heteroscedasticity (University of Wisconsin–Madison, 2021).

To justify the choice of NB regression over Poisson regression, I use an Ord plot to see if the dyadic forced migrant count variable follows a Poisson, binomial, negative binomial, or logarithmic series distribution (Friendly, 2011; National Institute of Standards and Technology, 2015). Moreover, I fit a Poisson model and perform an overdispersion test to check for any violations of equidispersion (Cameron & Trivedi, 1990). Finally, to check whether a zero-inflated NB model is a better fit to the data than a standard NB model, I perform a Chi-squared Difference Test between the two.

After fitting the ZINB model, I perform further diagnostic procedures. To examine goodness-of-fit of the model, a histogram of standardised residuals is used. Then, the data are examined for the presence of high-leverage points via an influence plot of standardised residuals versus hat-values, as well as a plot of Cook's distance for each observation. For an elaboration of rationale and procedure behind the aforementioned diagnostic operations, see Appendix D, section 6.4.

3 Results

Table 1 presents the results of the NB regression component of the ZINB model.

Table 1: Negative Binomial Regression Component

Dependent variable:	Forced migrant flows
Religious Proximity Index	0.890^{***} (0.029)
Social group equality (orig)	-0.927^{***} (0.041)
Population (orig)	-2.122^{***} (0.063)
Armed conflict (international) (orig)	-0.086^{***} (0.023)
Social group equality (dest)	-0.494^{***} (0.050)
Armed conflict (international) (dest)	-0.433^{***} (0.025)
Armed conflict (internal) (dest)	-0.959^{***} (0.021)
Contiguous	3.167^{***} (0.028)
Distance	0.189^{***} (0.033)
GDP (orig)	10.439^{***} (0.186)
GDP per capita (dest)	0.032 (0.064)
Shared official language	0.766^{***} (0.025)
Common coloniser post-1945	0.524^{***} (0.032)
Dependency relationship	2.977^{***} (0.042)
Common coloniser pre-1945	1.420^{***} (0.025)
Social Connectivity Index in 2021	6.159^{***} (0.104)
Labour force participation rate	-1.275^{***} (0.086)
Constant	8.984^{***} (0.055)
Observations	180,629
Log Likelihood	−1,764,535.000

Note: *p<0.1; **p<0.05; ***p<0.01

The coefficient on the Religious Proximity Index is 0.890, and given that it is a log-count, it can be interpreted multiplicatively following an exponential transformation (Hennigan, 2021; Yoshida, 2013). As such, a one unit increase in the Religious Proximity Index between i and j is associated with an increase in the count of forced migrants between i and j by a factor of 2.435. With a p-value below $2 * 10^{-16}$, this result is statistically significant at the 95% confidence level. Table *iii* (see Appendix E in section 6.5) shows the results of the logistic part of the ZINB model. Notable is the coefficient on Armed conflict (international) (orig) due to its congruence with aforementioned literature (Marbach, 2018; Silvestrin et al., 2021). If an origin country is affected by armed international conflict, on average it is 1.764 times more likely that a zero count reported from that country is not a true zero (i.e. arose from data collection issues and not an actual absence of migrants). With a p-value of 0.00542, this result is statistically significant at the 95% confidence level and beyond.

I now present the results from the assumption and robustness checks performed pre-analysis on alternative models to justify the selection of ZINB regression. The χ^2_{diff} parameter from the Chi-squared difference test between the ZINB and intercept-only models is statistically significant with a p-value below $2 * 10^{-16}$.

Figure 4 shows the Residuals vs Fitted Values plot for the multivariate OLS model:

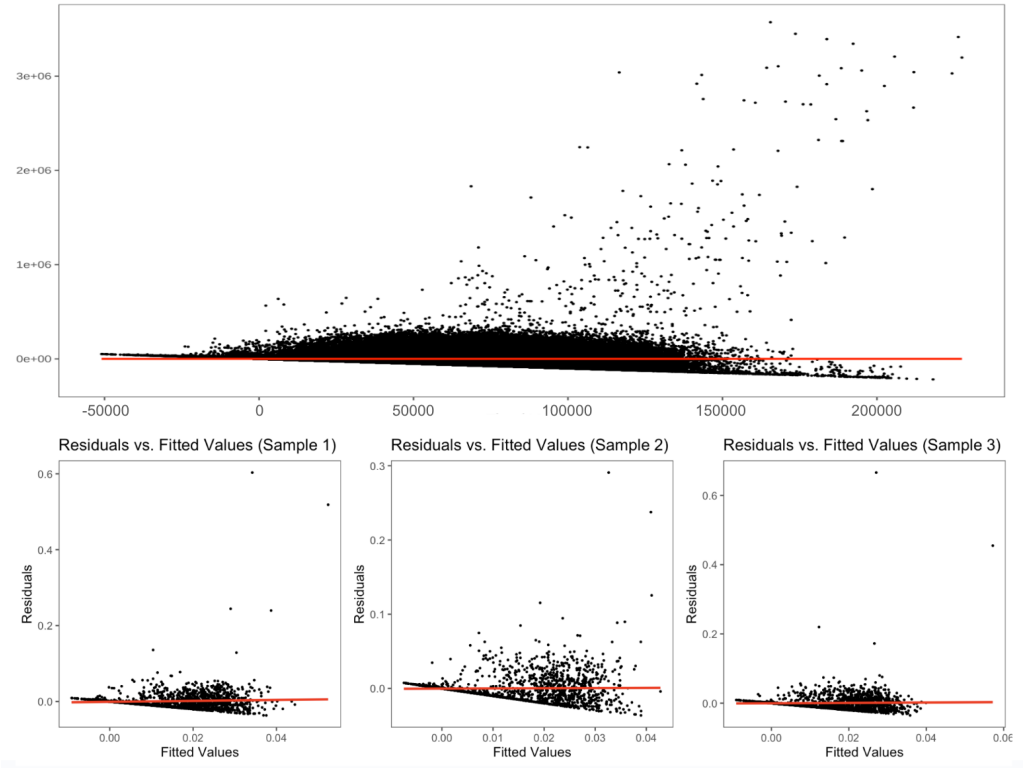

Figure 4: Residuals vs Fitted Values from the fitted OLS model (to visually assess the residuals for OLS assumptions' violations). The top plot captures the whole data, but given the clustering at the bottom due to the high number of observations, I plot three random samples (n = 2000) from the vector of residuals.

The residuals of the linear model exhibit an odd-cut off point at the bottom left. This tends to occur when there are many zeros in the data, and also due to the difference in support - the actual dyadic forced migrant count variable has support 0, Inf, while the support of the OLS fitted values is -Inf, Inf. With a Ramsay RESET f-statistic of 1342.742, statistically significant at all conventional levels given a p-value of $2.675 * 10^{-292}$, there is evidence of a linearity violation. The result of the Breusch-Pagan test is also statistically significant at all conventional levels given a p-value of $2.2 * 10^{-16}$, also evidencing heteroscedasticity. Both results evidence than an OLS model is unfit for my data.

Figure 5 shows the Ord plot of dyadic forced migrant count, showing that the variable's distribution is closest to a negative binomial distribution. Fitting a Poisson model with the same specifications as the ZINB model and subsequently performing an overdispersion test produces a dispersion parameter of 93473.14, statistically significant at all conventional levels with a p-value below $2.2 * 10^{-16}$, evidencing a violation of the equidispersion violation. The results of the Chi-Squared difference test between the fitted ZINB model and a standard negative binomial regression model of the same specification produces a p-value of $1.5 * 10^{-284}$, which makes the χ^2_{diff} parameter statistically significant and thus evidencing that the zero-inflated model is a better fit than the standard model.

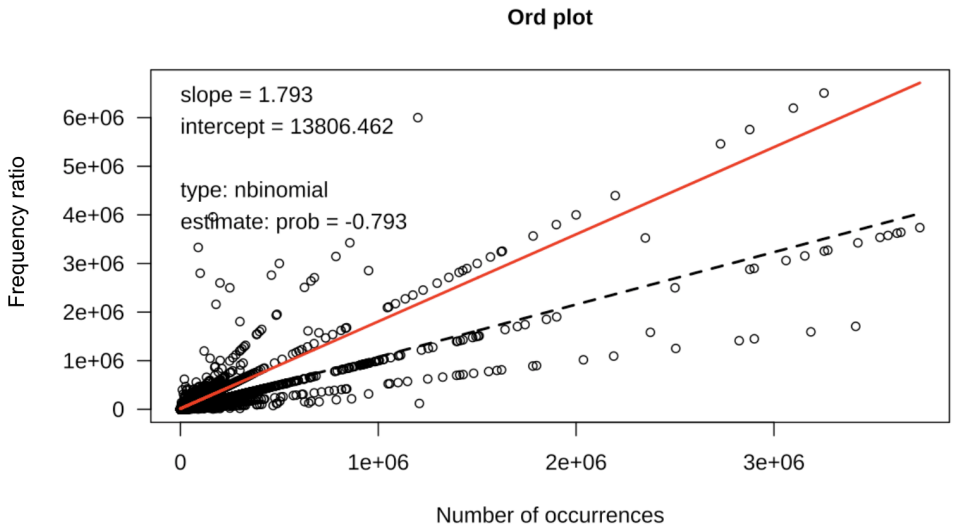

Figure 5: Ord plot of forced migrant count. Both the slope and the intercept are positive, implying that the forced migrant count variable's distribution is closest to a negative binomial distribution.

Moving on to post-model diagnostics of the ZINB model, Figure 6 shows the histogram of standardised residuals from the fitted ZINB model:

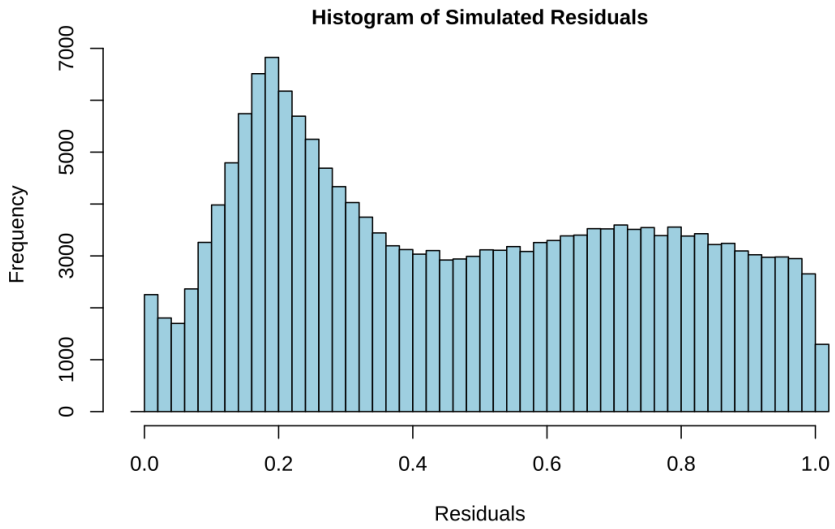

Figure 6: Histogram of simulated residuals from the fitted ZINB model.

The non-normal distribution of the residuals in Figure 6 shows that the model does not fully capture the underlying relationships in the data, as there is asymmetry around the zero line. Figure 7 shows numerous high-leverage points present in the data.

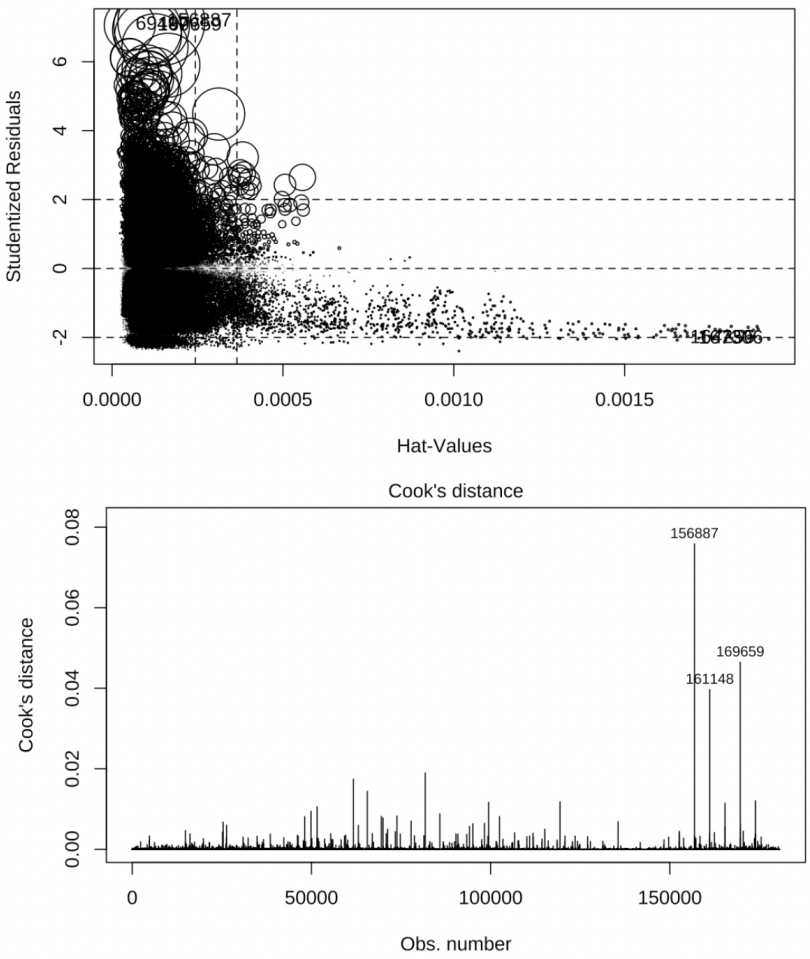

Figure 7: Combined plot: top plot is an influence plot of Studentised Residuals vs Hat-Values, with circle size proportional to Cook's distance, and bottom plot shows Cook's distance for each observation.

4 Discussion

4.1 Principal Findings

Recall the research question: *to what extent is religious proximity between two countries associated with the volume of forced migrant flows between them in a given point in time*? The principal finding from investigating this research question is that there is a positive and statistically significant association between religious ties and dyadic forced migrant count, enough to reject the null hypothesis that no association exists.

This finding is congruent with *qualitative* studies on the topic. Hagan and Ebaugh (2003) posit that the role of religion is often overlooked in the decision-making behind forced migration. With reference to Figure 2, a clear causal chain explains this role, especially strong when flight is catalysed by religious persecution, as is the case with the Rohingya (Czaika & Reinprecht, 2022). Scholars such as Wuthnow and Offutt (2008) highlight that in the age of globalisation, religion increasingly creates kinship networks that transcend borders.

And yet, *quantitative* studies disagree. For instance, Böcker and Havinga (1997) and Neumayer (2004) find no significant evidence that religious ties determine destination choice, in their studies of the EU context. Neumayer's (2005) later study probes further using a fixed effects model, and finds that religious proximity does not significantly influence volume of forced migrant flows. This is puzzling, as many of the same scholars find significant associations between forced migrant flows and other facets of ethno-cultural identity, such as language and ethnicity (Böcker & Havinga, 1997; Barthel & Neumayer, 2014; Rüegger & Bohnet, 2018), which this paper also observes in Table 1. Possible reasons for these differences include model miss-specification, omitted variable bias, as well as differences in data sources, given that migrant count data is very volatile.

4.2 Strengths and Limitations

A core strength of this paper is its dyadic research design incorporating both push and pull factors along with interaction covariates determining forced migrant flows. Many studies in the field of quantitative migration studies, such as Thielemann (2005) and Schneider and Holzer (2002) only consider pull factors, neglecting influential push factors such as political violence and oppression, instability and conflict (Neumayer, 2004, 163). Studies that *do* explore these factors often exclude pull factors, leading to the same omission biases (Neumayer, 2004). It is even rarer for studies to consider how push and pull factors interact to jointly drive migration patterns (Matsui & Raymer, 2020). My analysis draws on these existing studies and combines their insights to produce a more comprehensive analysis to highlight the role of an understudied determinant of forced migration flight patterns.

My paper makes a second contribution in considering the problem of zero-inflation, which only a handful of studies do. While studies such as Neumayer (2004) and Böcker and Havinga (1997) also employ a dyadic design, however their analysis is limited to the EU context, revealing a third contribution of this paper - testing existing theory on a *global* sample. While it would be a valid criticism to say that forced migration patterns

vary greatly by region, and thus more focused analyses are more perceptive to region-specific trends which a broader model may miss (Carammia, Iacus, & Wilkin, 2022), I hold that both approaches are required to further our understanding of the drivers of forced migration.

A further strength of my paper comes from the assumption and robustness checks performed. The first Chi-squared difference test validated that my chosen model was thankfully better than an intercept-only model. The residuals plot and the Ramsat RESET test from the fitted multivariate OLS alternative to the chosen model confirmed that a linear model would not have been suitable due to violations of linearity and homoscedasticity, justifying count regression (Suzuki, 2020). The dispersion test on the subsequent Poisson regression showed significant overdispersion, validating my choice of a negative binomial regression which is more suitable for an overdispersed response variable. Finally, the use of a zero-inflated NB model over a standard one is justified by a second Chi-squared difference test. Overall, the various checks allowed this paper to confidently apply the ZINB model.

However, post-model diagnostic checks reveal that the ZINB model is far from perfect. As seen in Figure 6, the non-normal distribution of standardised residuals signifies asymmetry around the zero residual line, which indicates that the model has not captured all underlying relationships in the data. The substantial amount of high-leverage points in Figure 7 further indicates that the model is likely very sensitive to slight changes. This is a model stability challenge, which could weaken the interpretability, reliability and replicability of my findings (Yu, 2013).

Another weakness of my approach is blindness to time variation. Scholars such as Carammia et al. (2022) highlight the detriments of not recognising that forced migration patterns vary largely across space and *time*. My study only includes cross-sectional covariates and does not account for time effects, which scholars such as Beyer et al. (2022) deploy as a strong criticism of similar models to mine.

Finally, my analysis is potentially sensitive to imputation method. Figure i (see Appendix B, section 6.2 reveals that 21.1% of my data was missing and required imputing. Especially given the high number of high-leverage points present, it is possible that changing the method of imputation from iterative imputation to another method, such as a multiple imputation algorithm like Classification and Regression Trees (CART) imputation, would introduce changes to high-leverage points and thus significantly alter the resultant coefficients from the fitted model.

4.3 Implications and Future Research

Despite limitations of the model, this paper furthers our understanding of how religious ties shape forced migration flows, challenging existing scholarship directly and suggesting a need for methodological review of older findings. Studies like mine are what underpin migrant flow forecasting frameworks which are used to inform policies as well as humanitarian action. As such, the policy implications of my findings manifest in evidence to revise what factors current forecasting and nowcasting frameworks rely on. In an ideal world where humanitarian actors and policymakers have the 'perfect' model to predict forced migrant flows between two countries given information about the countries and

about time, it would be statistical inference studies exploring questions like mine which inform the model's structure.

Future research stands to enrich my findings through incorporating time variation, considering techniques like autoregressive terms (i.e. lagged versions of covariates), moving averages (to account for long-term trends in forced migrant count), differencing (to account for short-run changes in the forced migrant count) and seasonality modelling (to account for repetitive patterns in forced migrant count through time) (Hoffmann Pham & Luengo-Oroz, 2022). The benefits can be seen in the work of scholars such as (Dreher, Fuchs, & Langlotz, 2019), whose implementation of time lags revealed that a migration driver they first found to be insignificant - total aid inflows - revealed their effect with an *eleven* year lag. This demonstrates that future research can reduce omitted variable bias by allowing for effects through time.

A further avenue for future research is to perform additional robustness checks on my findings that are out of scope for this paper. To investigate issues with model stability, further modelling should include the separation of high-leverage points above and below a given Cook's distance threshold, as well as the exclusion of outliers, and a subsequent analysis of standardised residuals for both sub-samples separately. Shifting this threshold and repeating the modelling would also reveal how/if coefficients change as different tiers of outliers and influential observations are removed. Moreover, the modelling strategy should be repeated on an unimputed version of the dataset, where only complete cases are kept. Comparing those results to the original fitted model results would help reveal whether imputation method sensitivity is exerting influence over the model coefficients, and whether the number of high-leverage points goes down.

Finally, future research should probe further into the coefficients on the zero-component of the fitted ZINB model in Table *iii* (Appendix E, section 6.5). The finding related to conflict and its effect on the propensity of a zero being excessive stands to reveal additional sources of bias, namely due to zero inflation, which are not traditionally accounted for in asylum migration studies quantitatively. Figure *iv* (Appendix E, section 6.5) also reveals some odd coefficients, notably on GDP (origin), imputation status and labour participation rate, which future research should investigate.

5 Conclusion

This paper asked: to what extent is religious proximity between two countries associated with the volume of forced migrant flows between them? To answer this question, I employed a zero-inflated negative binomial regression model, holding it to be the best fit for a count regression problem where the data exhibits high zero-count and overdispersion. The model took dyadic forced migrant counts for a global sample of 3724 unique orig-dest dyads, each for 1977-2023 as the dependent count variable. The religious proximity index was used as the independent variable of interest, and a vector of covariates (deemed significant by past studies and selected through Bayesian Model Averaging) was added to account for variance not explained by the religious proximity variable. Statistically significant at the 99.9% confidence level, the model found that a one unit increase in the religious proximity between two countries is associated with a 2.435 times increase in the dyadic count of forced migrants. With the model selection deemed appropriate during pre-analysis robustness checks, and despite post-analysis robustness checks revealing issues with model stability and uncaptured relationships in the data, this finding contradicts existing scholarship on the association between religious ties and forced migration, inviting future research to make these findings more robust and enrich the field's understanding of religion as an underestimated determinant of forced migrant flight patterns.

References

Aiken, L. S., Mistler, S. A., Coxe, S., & West, S. G. (2015). Analyzing count variables in individuals and groups: Single level and multilevel models. *Group Processes & Intergroup Relations*, *18*(3), 290–314.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, *32*, 259-80.

Barthel, F., & Neumayer, E. (2014, 10). Spatial dependence in asylum migration. *Journal of Ethnic and Migration Studies*, *41*, 1131-1151. doi: 10.1080/1369183x.2014.967756

Beyer, R. M., Schewe, J., & Lotze-Campen, H. (2022, 02). Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications*, *9*(1), 56. Retrieved 2022-12, from `https://www.nature.com/articles/s41599-022-01067-x` doi: 10.1057/s41599-022-01067-x

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019, 05). What does a zero mean? understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*, 949-959. doi: 10.1111/2041-210x.13185

Böcker, A., & Havinga, T. (1997). Asylum migration to the european union: Patterns of origin and destination. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2633536

Cameron, A., & Trivedi, P. K. (1990, 12). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, *46*, 347-364. doi: 10.1016/0304-4076(90)90014-k

Carammia, M., Iacus, S. M., & Wilkin, T. (2022, 01). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, *12*. doi: 10.1038/s41598-022-05241-8

Conte, M., Cotterlaz, P., & Mayer, T. (2022). *Cepii working paper* (Vol. 2022). Retrieved from `http://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=8`

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Altman, D. A., ... Ziblatt, D. (2023). *V-dem codebook v13.* Varieties of Democracy (V-Dem) Project.

Crowley, S. (2012). *Maximum likelihood estimation of the negative binomial dis- tribution.* Retrieved 2024-01-07, from `https://vixra.org/pdf/1211.0113v1.pdf`

Czaika, M., & Reinprecht, C. (2022). Migration drivers: Why do people migrate? *IMISCOE Research Series*, 49-82. doi: 10.1007/978-3-030-92377-8_3

Davenport, C., Moore, W., & Poe, S. (2003, 01). Sometimes you just have to leave: Domestic threats and forced migration, 1964-1989. *International Interactions*, *29*(1), 27–55. Retrieved 2022-12, from `https://doi.org/10.1080/03050620304597` doi: 10.1080/03050620304597

Deb, S., & Baudais, V. (2022, 10). *The challenges of data collection in conflict-affected areas: A case study in the liptako-gourma region.* Stockholm International Peace Research Institute. doi: 10.55163/vwim3307

Devictor, X., Do, Q.-T., & Levchenko, A. A. (2020, 04). The globalization of refugee flows. *Social Science Research Network*. doi: 10.3386/w28332

Disdier, A.-C., & Mayer, T. (2007, 12). Je t'aime, moi non plus: Bilateral opinions and international trade. *European Journal of Political Economy*, *23*, 1140-1159. doi: 10.1016/j.ejpoleco.2006.09.021

Dreher, A., Fuchs, A., & Langlotz, S. (2019, 02). The effects of foreign aid on refugee flows. *European Economic Review*, *112*, 127-147. doi: 10.1016/j.euroecorev.2018.12.001

Dynan, K., & Sheiner, L. (2018). *Gdp as a measure of economic well-being* (Tech. Rep.). Hutchins Center Working Paper.

Feldkircher, M., & Zeugner, S. (2022). *Bayesian model averaging with bms for bms version 0.3.5.* CRAN. Retrieved from `https://cran.r-project.org/web/packages/BMS/vignettes/bmsmanual.pdf`

Friendly, M. (2011, 08). Visualizing categorical data. *SpringerReference*. doi: 10.1007/springerreference_64265

Guo, C. G., Al Ariss, A., & Brewster, C. (2020, 04). Understanding the global refugee crisis: Managerial consequences and policy implications. *Academy of Management Perspectives*, *34*. doi: 10.5465/amp.2019.0013

Hagan, J., & Ebaugh, H. R. (2003, 12). Calling upon the sacred: Migrants' use of religion in the migration process. *International Migration Review*, *37*, 1145-1162. doi: 10.1111/j.1747-7379.2003.tb00173.x

Hakovirta, H. (1993). The global refugee problem: A model and its application. *International Political Science Review / Revue internationale de science politique*, *14*(1), 35–57. Retrieved 2022-12, from `https://www.jstor.org/stable/1601374`

He, Q., & Huang, H.-H. (2023, 03). A framework of zero-inflated bayesian negative binomial regression models for spatiotemporal data. *Journal of Statistical Planning and Inference*, *229*, 106098-106098. doi: 10.1016/j.jspi.2023.106098

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, *36*, 531-547. doi: 10.1002/bimj.4710360505

Hennigan, P. (2021). *Negative binomial regression guide.* University of New Hampshire. Retrieved from `https://www.researchgate.net/publication/349573943_Negative_Binomial_Regression_Guide`

Hoffmann Pham, K., & Luengo-Oroz, M. (2022, 08). Predictive modelling of movements of refugees and internally displaced people: towards a computational framework. *Journal of Ethnic and Migration Studies*, 1-37. doi: 10.1080/1369183x.2022.2100546

Iqbal, Z. (2007, 04). The geo-politics of forced migration in africa, 1992—2001. *Conflict Management and Peace Science*, *24*, 105-119. doi: 10.1080/07388940701257515

Kang, Y.-D. (2020, 01). Refugee crisis in europe: determinants of asylum seeking in european countries from 2008–2014. *Journal of European Integration*, 1-16. doi: 10.1080/07036337.2020.1718673

Korosteleva, O. (2018). *Advanced regression models with sas and r.* Chapman and Hall/CRC. doi: 10.1201/9781315169828

Kyner, T. J. (2021, 08). *Iterative imputation with scikit-learn.* Retrieved from `https://towardsdatascience.com/iterative-imputation-with-scikit-learn-8f3eb22b1a38`

Laughon, K., Montalvo-Liendo, N., Eaton, S., & Bassett, L. (2023). Health and safety concerns of female asylum seekers living in an informal migrant camp in matamoros, mexico. *Journal of Advanced Nursing*, *79*(5), 1830–1839.

Leal, D. F., & Harder, N. L. (2021). Global dynamics of international migration systems across south–south, north–north, and north–south flows, 1990–2015. *Applied Network Science*, *6*, 1–27.

Long, J. S. (1997). Regression models for categorical and limited dependent variables. Sage Publications.

Loukas, S. (2020, 06). *Everything you need to know about min-max normalization in python.* Retrieved from `https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79`

Marbach, M. (2018). On imputing unhcr data. *Research & Politics*, *5*(4),

2053168018803239.

Markham, L. (2022, 04). 'a disaster waiting to happen': who was really responsible for the fire at moria refugee camp? *The Guardian*. Retrieved from https://www.theguardian.com/world/2022/apr/21/disaster-waiting-to-happen-moria-refugee-camp-fire-greece-lesbos

Matsui, N., & Raymer, J. (2020, 06). The push and pull factors contributing towards asylum migration from developing countries to developed countries since 2000. *International Migration*. doi: 10.1111/imig.12708

Minora, U., Belmonte, M., Bosco, C., Johnston, D., Giraudy, E., Iacus, S., & Sermi, F. (2023). *Migration patterns, friendship networks, and the diaspora: the potential of facebook's social connectedness index to anticipate displacement patterns induced by russia's invasion of ukraine in the european union.* International Organization for Migration. Retrieved 2023-11-02, from https://publications.iom.int/system/files/pdf/MRS-73.pdf

Missbach, A., & Stange, G. (2021). Muslim solidarity and the lack of effective protection for rohingya refugees in southeast asia. *Social Sciences*, *10*(5), 166.

National Institute of Standards and Technology. (2015). *Ord plot.* Retrieved from https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/ordplot.htm

Nations, U. (2020). *A new era of conflict and violence.* Retrieved from https://www.un.org/en/un75/new-era-conflict-and-violence

Neumayer, E. (2004, 06). Asylum destination choice: What makes some west european countries more attractive than others? *European Union Politics*, *5*, 155-180. doi: 10.1177/1465116504042444

Neumayer, E. (2005, 09). Bogus refugees? the determinants of asylum migration to western europe. *International Studies Quarterly*, *49*, 389–409. Retrieved from https://academic.oup.com/isq/article-abstract/49/3/389/1934647 doi: 10.1111/j.1468-2478.2005.00370.x

O'Hara, R., & Kotze, J. (2010). Do not log-transform count data. *Nature Precedings*, 1–1.

Pellandra, A., & Henningsen, G. (2022). *Predicting refugee flows with big data: a new opportunity or a pipe dream?* UNHCR. Retrieved from https://www.unhcr.org/blogs/predicting-refugee-flows-with-big-data-a-new-opportunity-or-a-pipe-dream/

Potrafke, N. (2015, 05). The evidence on globalisation. *The World Economy*, *38*, 509-552. doi: 10.1111/twec.12174

Richmond, A. H. (1993). Reactive migration: Sociological perspectives on refugee movements. *Journal of Refugee Studies*, *6*, 7-24. doi: 10.1093/jrs/6.1.7

Rüegger, S., & Bohnet, H. (2018, 01). The ethnicity of refugees (er): A new dataset for understanding flight patterns. *Conflict Management and Peace Science*, *35*(1), 65–88. Retrieved 2022-12, from http://journals.sagepub.com/doi/10.1177/0738894215611865 doi: 10.1177/0738894215611865

Schmeidl, S. (1997). Exploring the causes of forced migration: A pooled time-series analysis, 1971-1990. *Social Science Quarterly*, *78*(2), 284–308. Retrieved 2022-12, from https://www.jstor.org/stable/42864338

Schneider, G., & Holzer, T. (2002). *Asylpolitik auf abwegen : nationalstaatliche und europäische reaktionen auf die globalisierung der flüchtlingsströme.* Opladen : Leske + Budrich. Retrieved 2024-01-09, from http://nbn-resolving.de/urn:nbn:de:bsz:352-134507

Shaffer, J. P. (1991, 11). The gauss—markov theorem and random regressors. *The American Statistician*, *45*, 269-273. doi: 10.1080/00031305.1991.10475819

Silvestrin, L. P., Pantiskas, L., & Hoogendoorn, M. (2021, 07). A framework for imbalanced time-series forecasting. *arXiv (Cornell University)*. doi: 10.48550/arxiv.2107.10709

Sofuoğlu, E., & Ay, A. (2020, 02). The relationship between climate change and political instability: the case of mena countries (1985:01–2016:12). *Environmental Science and Pollution Research*, *27*. doi: 10.1007/s11356-020-07937-8

Suzuki, T. (2020, 12). Destination choice of asylum applicants in europe from three conflict-affected countries. *Migration and Development*, 1-13. doi: 10.1080/21632324.2020.1855738

Thielemann, E. (2005). Does policy matter? on governments' attempts to control unwanted migration. *SSRN Electronic Journal*. doi: 10.2139/ssrn.495631

Tucker, J. (2018, 10). Why here? factors influencing palestinian refugees from syria in choosing germany or sweden as asylum destinations. *Comparative Migration Studies*, *6*(1), 29. Retrieved 2022-12, from `https://doi.org/10.1186/s40878-018-0094-2` doi: 10.1186/s40878-018-0094-2

UNHCR. (2023, 10). *Refugee statistics.* Author. Retrieved from `https://www.unhcr.org/refugee-statistics/`

University of Wisconsin–Madison. (2021). *Regression diagnostics with stata.* Social Science Computing Cooperative. Retrieved from `https://sscc.wisc.edu/sscc/pubs/RegDiag-Stata/`

Van Hear, N., Brubaker, R., & Bessa, T. (2009, 06). Managing mobility for human development: the growing salience of mixed migration. *Human Development Research Paper (HDRP) Series*, *20*. Retrieved 2023-11-07, from `https://mpra.ub.uni-muenchen.de/19202/`

Wang, R. (2018). *Understanding multicollinearity in bayesian model averaging with bic approximation* (Doctoral dissertation). Retrieved from `https://summit.sfu.ca/item/17968`

Werner, C., & Schermelleh-Engel, K. (2010). *Introduction to structural equation modeling with lisrel.* Goethe University.

Wilson, P. (2015, 02). The misuse of the vuong test for non-nested models to test for zero-inflation. *Economics Letters*, *127*, 51-53. Retrieved from `https://www.sciencedirect.com/science/article/abs/pii/S016517651400490X` doi: 10.1016/j.econlet.2014.12.029

Wuthnow, R., & Offutt, S. (2008, 06). Transnational religious connections*. *Sociology of Religion*, *69*, 209-232. doi: 10.1093/socrel/69.2.209

Xiao, T., Oppenheimer, M., He, X., & Mastrorillo, M. (2022, 01). Complex climate and network effects on internal migration in south africa revealed by a network model. *Population and Environment*, *43*, 289-318. doi: 10.1007/s11111-021-00392-8

Yau, K. K. W., Wang, K., & Lee, A. H. (2003, 06). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, *45*, 437-452. doi: 10.1002/bimj.200390024

Yoshida, K. (2013). *Models for excess zeros using pscl package (hurdle and zero-inflated regression models) and their interpretations.* RPubs. Retrieved from `https://rpubs.com/kaz_yos/pscl-2`

Yu, B. (2013, 09). Stability. *Bernoulli*, *19*. doi: 10.3150/13-bejsp14

6 Appendices

6.1 Appendix A: Summary of Data

Source	Variables	Limitations	Justification
UNHCR (1978 - 2023, 8626 unique orig-dest pairs)	**Forced migrant flows** (count of people) (*forced_mig*) constructed by adding together Refugees under UNHCR's mandate and Asylum-seekers	UNHCR only records non-zero flows where both origin and destination are known, some missing values are not actually zero, but instead are migrants of unknown origin (Marbach, 2018)	This is the case in approximately 7% of the data, so through imputation the data is still reliable (Marbach, 2018).
Varieties of Democracy (V-DEM) dataset (Coppedge et al., 2023)	**Freedom from torture** (0-4) (*freedom_torture*) **Social group equality** in respect for civil liberties (0-4) (*group_equality_lib*) **Armed conflict** (international/internal) (binary) **Equal protection index** (0-1) (*equal_protection*)	As this dataset is itself a compilation of different sources, data gaps appear unevenly across variables, meaning that for most orig-dest dyads at least one of them will be missing. Moreover, the scale reported in the codebook is at times different to the actual values in the dataset, leading to potential interpretability issues.	Prior to modelling, all variables will be standardised to avoid challenges from differing scales. Moreover, thorough iterative imputation, the data gaps can be filled in a way that reflects non-missing information for a given row, reducing issues from uneven data gaps.
Labour Force Statistics (LFS)(ILO, 2023)	**Labour force participation rate** (%) (*labour_part_rate*)	It may not be relevant to this analysis, as many asylum seekers are subjected to working bans, meaning that their job prospects are completely distinct to those of the local people.	Migrants may still use this as an indication of overall job market health as part of their evaluation of their prospects should they be granted asylum.
Meta (Bailey, Cao, Kuchler, Stroebel, & Wong, 2018)	**Social Connectivity Index** (1-1 billion) (*scaled_sci_2021*)	It allows for very limited analysis as data is only available for 2021.	Existing studies find a significant association between the SCI and refugee flows (Minora et al., 2023).

Table i: Summary of variables included in data

Source	Variables	Limitations	Justification
CEPII Gravity Database (Conte, Cotterlaz, & Mayer, 2022)	**Distance** (km) (*dist*) **Contiguity** (1 if share border, 0 if not) (*contig*) **Population** (count, i and j separate) (*pop*) **GDP** (current thousands USD, i and j separate) (*gdp*) **GDP per capita** (current thousands USD, i and j separate) (*gdp_cap*) **Shared official language** (binary) (*comlang_off*) **Shared spoken language** (binary) (*comlang_ethno*) **Common coloniser post/pre 1945** (binary) (*comcol, sibling_ever*) **Dependency** (binary) (*col_dep_ever*) **Religious Proximity Index** (0-1, sum of products of shares of different religions in i and j (Disdier & Mayer, 2007)) (*comrelig*)	There is more than one way to measure distance (e.g. between centroids, capital cities, minimum distance), so the chosen metric may be sensitive to measurement bias. Some argue that GDP is neither accurate nor fair, and is a limited measure of economic wellbeing in the context of justifying social policy (Dynan & Sheiner, 2018). The bilateral variables are mostly binary (except for religious proximity), which fails to capture nuance in the strength of the linkage between i and j.	Prior to modelling, all variables will be standardised to avoid challenges from differing scales. Moreover, thorough iterative imputation, the data gaps can be filled in a way that reflects non-missing information for a given row, reducing issues from uneven data gaps.

Table i (cont): Summary of variables included in data

k	min	median	mean	max	n
col_dep_ever	0	0.0000000	1.060800e − 01	6.931472e − 01	184353
comcol	0	0.0430897	1.618834e − 01	6.931472e − 01	184353
comlang_ethno	0	0.2644810	2.810436e − 01	6.931472e − 01	184353
comlang_off	0	0.2863168	2.927109e − 01	6.931472e − 01	184353
comrelig	0	0.2053868	2.323424e − 01	6.931472e − 01	184353
conflict_internal_dest	0	0.0000000	1.403050e − 01	6.931472e − 01	184353
conflict_internal_orig	0	0.1673949	2.489148e − 01	6.931472e − 01	184353
conflict_international_dest	0	0.0000000	1.101227e − 01	6.931472e − 01	184353
conflict_international_orig	0	0.0000000	1.287391e − 01	6.931472e − 01	184353
contig	0	0.1103388	1.950539e − 01	6.931472e − 01	184353
dist	0	0.2246159	2.393132e − 01	6.931472e − 01	184353
equal_protection_dest	0	0.6112546	5.582800e − 01	6.931472e − 01	184353
equal_protection_orig	0	0.4546392	4.062346e − 01	6.931472e − 01	184353
forced_mig	0	1347.1568953	4.530997e + 04	3.737369e + 06	184353
freedom_torture_dest	0	0.5449107	5.035797e − 01	6.931472e − 01	184353
freedom_torture_orig	0	0.3443435	3.453983e − 01	6.931472e − 01	184353
gdp_d	0	0.0313656	7.961660e − 02	6.931472e − 01	184353
gdp_o	0	0.0106070	2.628920e − 02	6.931472e − 01	184353
gdpcap_d	0	0.1110901	1.206802e − 01	6.931472e − 01	184353
gdpcap_o	0	0.0373097	5.981390e − 02	6.931472e − 01	184353
group_equality_lib_dest	0	0.5054577	4.876153e − 01	6.931472e − 01	184353
group_equality_lib_orig	0	0.4194242	3.897410e − 01	6.931472e − 01	184353
labour_part_rate_total_dest	0	0.5065334	4.985541e − 01	6.931472e − 01	184353
labour_part_rate_total_orig	0	0.4488335	4.443596e − 01	6.931472e − 01	184353
population_dest	0	0.0140641	3.673700e − 02	6.931472e − 01	184353
population_orig	0	0.0126562	4.200040e − 02	6.931472e − 01	184353
scaled_sci_2021	0	0.0312380	5.280480e − 02	6.931472e − 01	184353
sibling_ever	0	0.3001607	2.997979e − 01	6.931472e − 01	184353

Table ii: Descriptive statistics

6.2 Appendix B: Missing Values

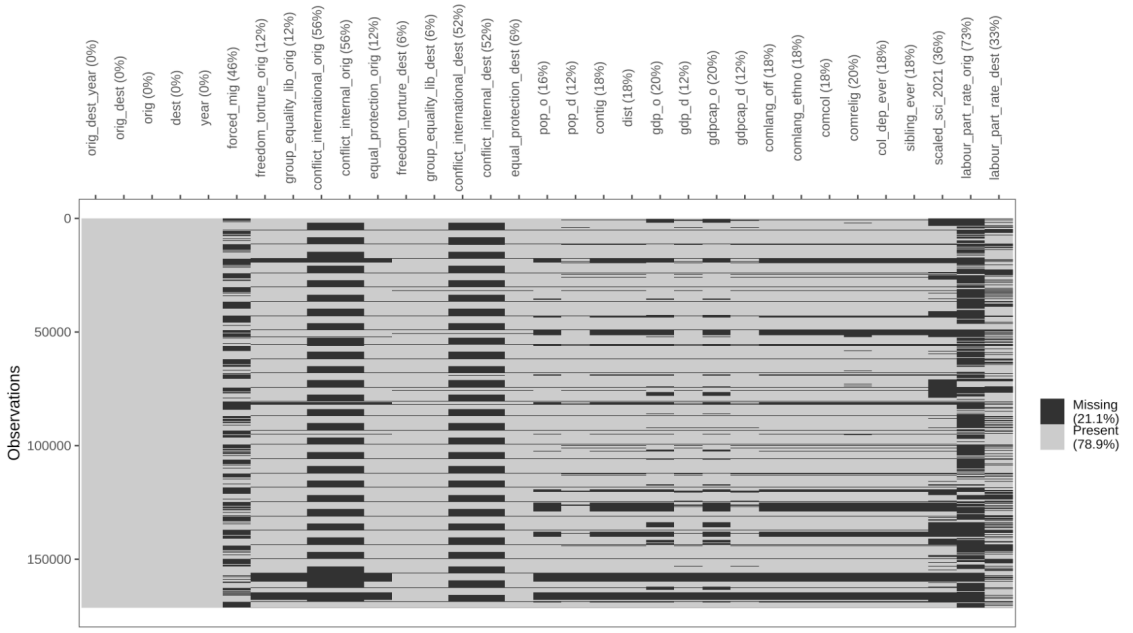

Figure i: Visual matrix of missing vs present values in the expanded, unimputed dataset.

6.3 Appendix C: Feature Selection

As my dataset was built by including covariates that existing literature suggests are relevant, with no guarantee of relevance or safeguard against multicollinearity, I use a Pearson correlation matrix to identify any multicollinearity issues, and I employ a Bayesian Model Averaging (BMA) approach to filter out non-influential covariates. BMA is a widely used method for feature selection (Wang, 2018). As an alternative to inefficiently including all possible covariates in a regression model, BMA estimates models for all possible combinations of covariates in a covariate space $\mathbf{X}$, such that for k covariates, $2k$ models are fit and stored in a model space $\mathbf{M}$. Each model is then assigned a weight - the *Posterior Model Probability* (PMP) - reflecting the probability that it best predicts the observed data (Feldkircher & Zeugner, 2022). Through renormalisation, the PMP calculation is used to estimate the probability of an individual covariate θ being included in the 'best model' - referred to as the *Posterior Inclusion Probability* (PIP) (ibid). To determine which covariates to include in my ZINB model, I set a PIP threshold of 0.5 and exclude any variables that have a PIP below that, as they can be deemed not influential enough and would thus complicate the model unnecessarily. I choose this as BMA has been argued to deal better with model uncertainty and to be more comprehensive than other approaches, such as a Least Absolute Shrinkage and Selection Operator (LASSO) approach (ibid).

The results from the BMA approach are summarised in Figure *ii*:

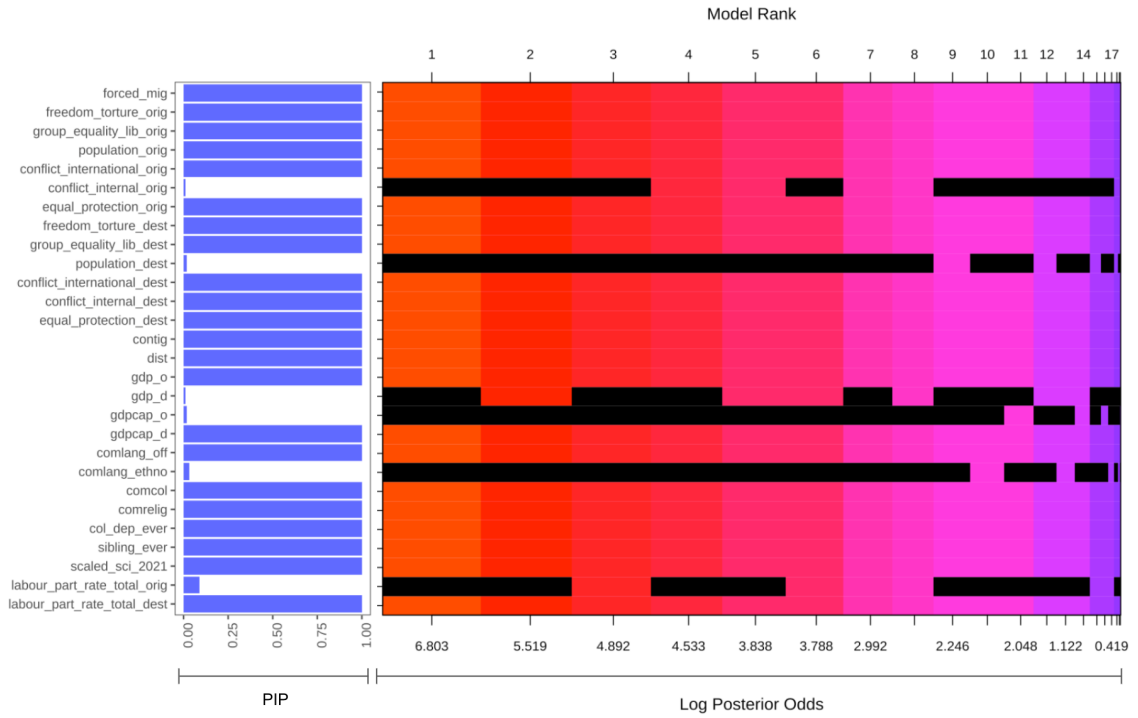

Figure ii: Combined plot showing the PIPs for each covariate considered, as well as a plot of the log posterior odds of their inclusion in the model that best fits the observed data.

According to the PIP exclusion threshold and Figure *ii*, the following covariate will no longer be considered in the final model specifications: *conflict_internal_orig*, *population_dest*, gdp_d, *gdpcap_o*, *comlang_ethno* and *labour_part_rate_total_orig*. Next, I present the Pearson correlations matrix of the remaining covariates in Figure *iii*, inspecting the data for any multicollinearity (i.e. pairwise correlation coefficient above 0.7).

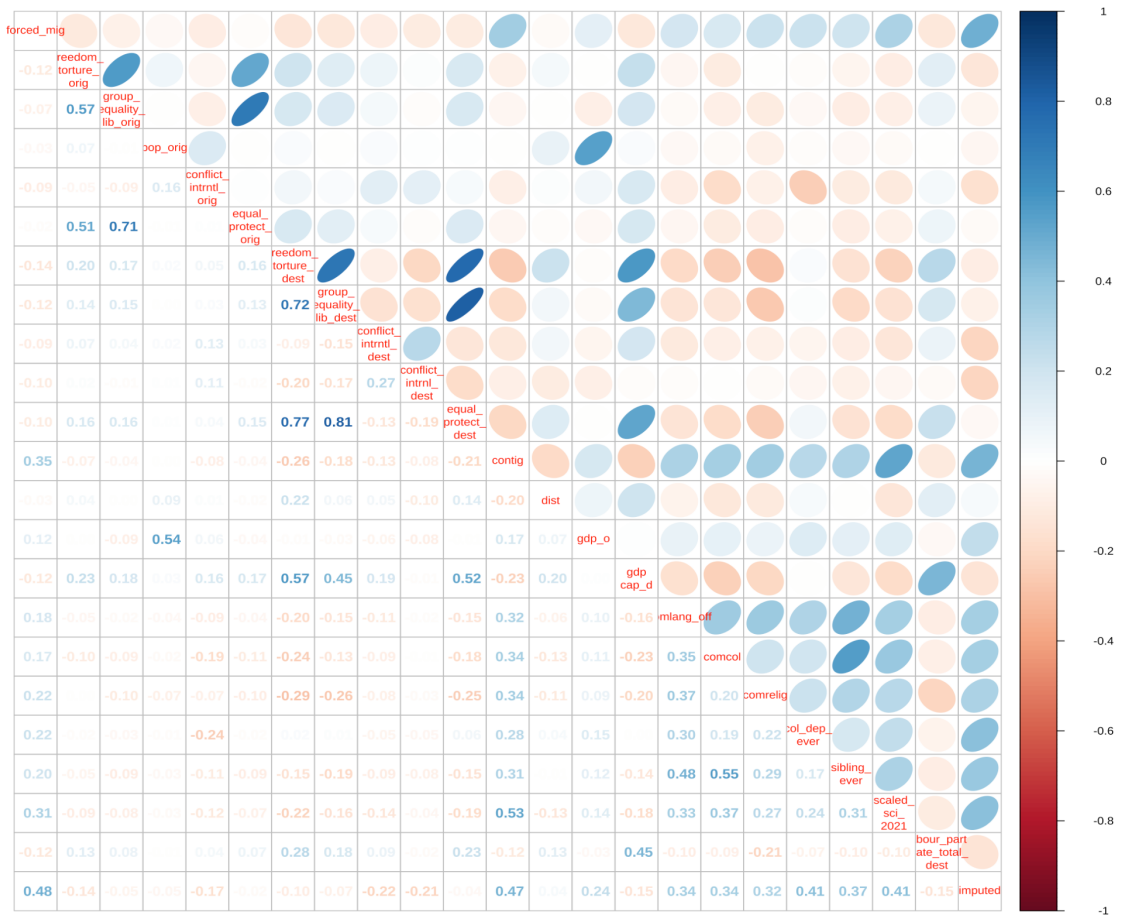

Figure iii: Pairwise Pearson correlation matrix showing correlations between all pairs of considered covariates.

Figure *iii* shows that the majority of pairwise correlations are weak, thus indicating no multicollinearity. However, there are a few correlations above 0.7, namely between *group_equality_lib_orig* and *equal_protect_orig*, *freedom_torture_dest* and *group_equality_lib_dest*, *freedom_torture_dest* and *equal_protect_dest*, and *group_equality_lib_dest* and *equal_protect_dest*. Given that the *group_equality_lib* variable for both origin and destination countries is most highly correlated with the other political covariates mentioned, and since the other covariates do not have any high correlations with covariates other than the ones mentioned, I remove the *freedom_torture* and *equal_protect* variables for both origin and destination countries, only keeping the *group_equality_lib* variable under the assumption that it captures the variation from the others.

With reference to equations 7 and 8, the following covariates are the final chosen components of each of the four covariate vectors:

- **X**: *group_equality_lib_orig*, *population_orig*,*gdp_o*.
- **Y**: *group_equality_lib_dest*, *conflict_international_dest*, *conflict_internal_dest*, *gdpcap_d*, *labour_part_rate_total_dest*.
- **Z**: *dist*, *contig*, *comlang_off*, *comcol*, *comrelig*, *col_dep_ever*, *sibling_ever*, *scaled_sci_2021*.
- **Q**: *imputed*.

6.4 Appendix D: About Assumptions and Robustness Checks

The following is a more detailed account of section 2.3

6.4.1 Chi-squared difference test

As an initial check that the ZINB model improves upon a null model, I fit an intercept-only model that removes all added covariates and compare whether there is a statistically significant improvement in how well the model fits the data (Long, 1997). To do this, I use a Chi-squared difference test, as such a test is only meaningful when the compared models are nested, meaning that one can be obtained by changing parameters in the other - a criterion which these models fit (Werner & Schermelleh-Engel, 2010). The Chi-squared difference test is computed as such:

$$\begin{aligned} \chi^2_{\text{diff}} &= \chi^2_{\text{null}} - \chi^2_{\text{full}} \\ df_{diff} &= df_{\text{null}} - df_{\text{full}} \end{aligned} \tag{9}$$

where χ^2_{null} is is the Chi-squared statistic of the intercept-only model, χ^2_{full} is the Chi-squared statistic of the main ZINB model, and df_{null} and df_{full} are the degrees of freedom of the intercept-only and ZINB models respectively. If the χ^2_{diff} parameter is significant, I can conclude that the full model is a better fit for the data than the null model (ibid).

6.4.2 Linearity and homoscedasticity assumptions

To ensure that a count regression is appropriate for my data than a linear model, I fit a multivariate linear regression model and assess its residuals for any violations of the linearity and homoscedasticity assumptions, either of which would indicate that an OLS model is not appropriate. The linear model is equivalent to the ZINB in terms of covariates included. Under the Gauss-Markov theorem, the OLS estimated regression coefficients will be the best linear unbiased estimates (BLUE) possible, provided that the following assumptions are met (Shaffer, 1991):

1. **Linearity**: model must fit a linear pattern without non-linear link functions.

2. **Randomness**:the input data must be a random sample of the population.
3. **Non-Collinearity**: there must be weak to no collinearity between covariates.
4. **Exogeneity**: covariates are not correlated with the error term.
5. **Homoscedasticity**: the error of the variance is constant.

As the feature selection stage already features a safeguard against multicollinearity, I will not test for that assumption, neither will I test for exogeneity or randomness as it is outside the scope of this paper. To test for linearity, I plot the residuals vs fitted values from the linear model - if any pattern emerges, this would imply non-linear relationships present in the data which the model has not accounted for. Moreover, I use a Ramsey Regression Equation Specification Error Test (RESET), which conducts a nested model comparison by adding polynomial terms to the current model, performs an F-test, and evaluates whether the inclusion of non-linear terms improves model fit, indicating potential specification errors like omitted curvilinear terms or linearity violations (University of Wisconsin–Madison, 2021). To test for homoscedasticity, I use a Breusch-Pagan Test, which regresses the residuals with the fitted values or predictors, and assesses whether they can explain any residual variance (ibid). If the resultant p-value is small, there is evidence that residual variance is non-constant, thus signalling a violation of homoscedasticity.

6.4.3 Overdispersion and zero-inflation

Beyond confirming that a count regression is more appropriate than a linear model, I must also ensure that the type of count regression chosen - negative binomial regression, is a better fit than Poisson regression. Moreover, I must also check if the zero-inflated negative binomial regression is better than a standard negative binomial regression. To check whether I have chosen the right distribution, I first use an Ord plot, which plots the ordered values of the data against the expected values from a given distribution (Friendly, 2011). This can be used to assess whether data follows a Poisson, binomial, negative binomial, or logarithmic series distribution (National Institute of Standards and Technology, 2015). This is determined by the following conditions of the fitted line (ibid):

Distribution	Slobe $\boldsymbol{\beta}$	Intercept $\boldsymbol{\alpha}$	Parameter estimate
Poisson	0	+	$\lambda = \alpha$
Binomial	-	+	$\mathrm{p} = \beta/(\beta - 1)$
Negative binomial	+	+	$\mathrm{p} = 1 - \beta$
Logarithmic series	+	-	$\theta = -\alpha$

As a further check, in line Cameron and Trivedi (1990), I fit a Poisson regression, and perform an Overdispersion test. The intuition is that if my dependent variable follows a Poisson distribution, then the equidispersion assumption - the equivalence between the mean and the variance - is met. This is treated as the null hypothesis scenario, tested against an alternative where:

$$\mathrm{Var}(Y) = \mu + c * f(\mu) \tag{10}$$

If c is above 0, and if the result is statistically significant, I can conclude that the equidispersion assumption is violated and that there is overdispersion in my data, which a negative binomial model is more fit to handle as previously discussed. To check whether a zero-inflated negative binomial model is better than a standard negative binomial model, I will again use a Chi-Squared test. One could argue that a more appropriate test specific for zero-inflation is required, such as a Vuong test, however the Vuong test has been shown to be inappropriate due to a misunderstanding of the term "non-nested model" (Wilson, 2015). In the absence of a better alternative, I proceed with this approach.

Finally, I perform two model diagnostic procedures on the fitted ZINB model. Firstly, I plot a histogram of simulated residuals, which would confirm goodness-of-fit of the model if the histogram approximates a normal distribution around 0. I then investigate influential points through a plot of Cook's distance for each observation, as well as an Influence Plot of standardised residuals versus hat-values. Both will confirm if there are many high-influence points, which would compromise the stability of the model, as it would imply that a small change in those points could significantly change the model's results.

6.5 Appendix E: ZIMB Results

Table iii: Binary Logistic Regression Component

Dependent variable:	forced_mig
Imputed	16.591 (65.794)
Religious Proximity Index	−0.063 (0.245)
Social group equality (orig)	6.363*** (0.432)
Population (orig)	1.664*** (0.462)
Armed conflict (international) (orig)	0.567*** (0.204)
Social group equality (dest)	8.792*** (0.709)
Armed conflict (international) (dest)	0.221 (0.229)
Armed conflict (internal) (dest)	1.136*** (0.187)
Contiguous	−3.485*** (0.277)
Distance	0.304 (0.274)
GDP (orig)	2.989** (1.312)
GDP per capita (dest)	1.602*** (0.505)
Shared official language	0.181 (0.224)
Common coloniser post-1945	0.728*** (0.239)
Dependency relationship	1.192*** (0.294)
Common coloniser pre-1945	1.795*** (0.258)
Social Connectivity Index in 2021	−0.691 (0.781)
Labour force participation rate	10.395*** (0.896)
Constant	−34.051 (65.797)
Observations	180,629
Log Likelihood	−1,764,535.000

Note: *p<0.1; **p<0.05; ***p<0.01

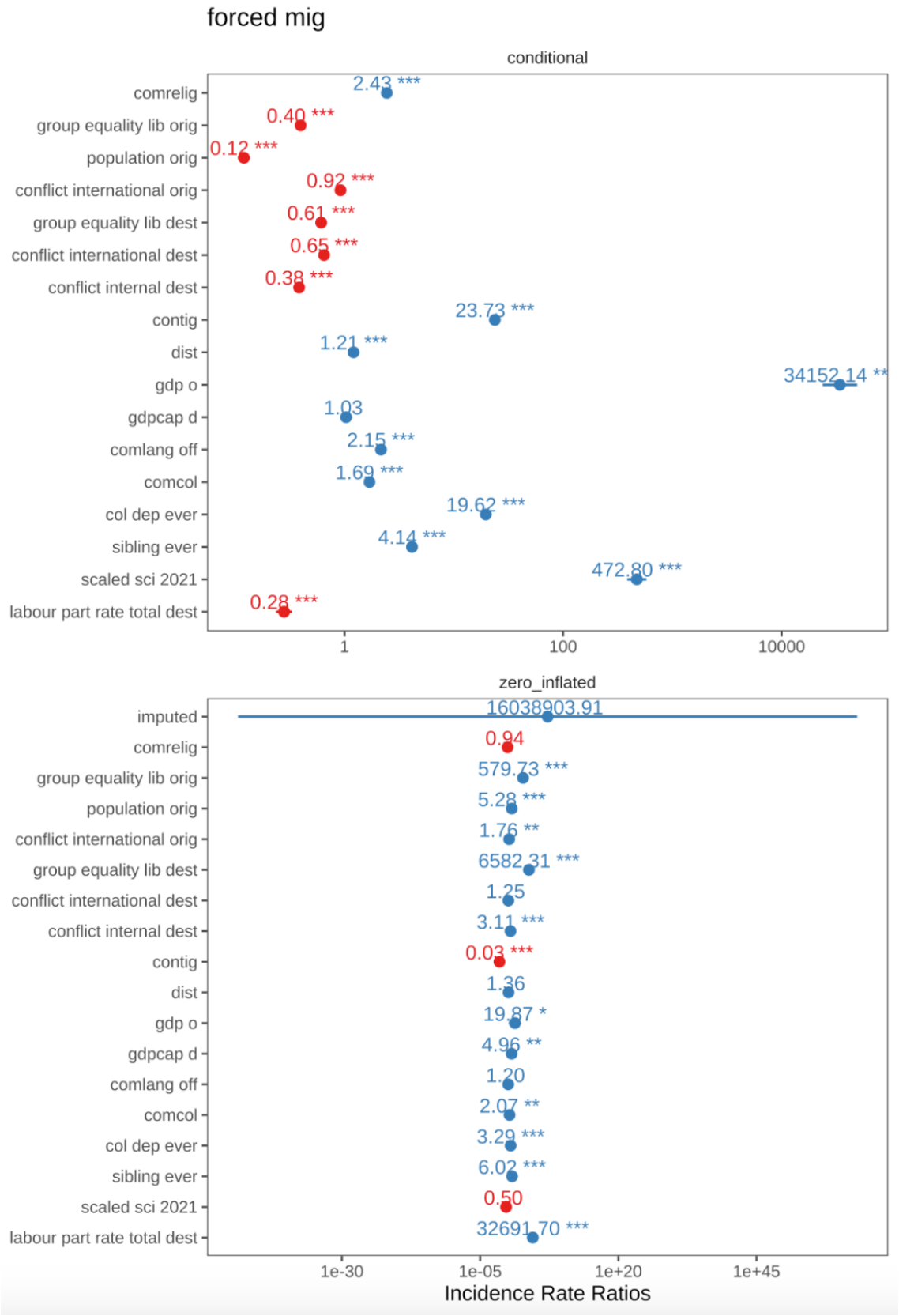

Figure iv: Forest plots showing the results of the ZINB model (top plot shows results from the NB component, bottom plot shows the results from the logistic component). Exponential transformations have already been applied for interpretability. Some coefficients, e.g. on the *imputed* and *gdp_o* variables, raise concern, however it is beyond the scope of this essay to investigate them.