1. **Can we assume that the sex/age distribution in the population with missing data is the same as in the population with available data?**

NO (see figure 3), origins of those without demo data are concentrated in Americas and Asia-Pacific, compared to MENA and Africa for those with data. Assuming origin has at least some relevance for a refugee population’s sex/age distribution (potentially through proxies such as length of displacement), missing data is non-ignorable.

1. **Is the demographic composition of refugee populations from the same country of origin similar in different countries of asylum?**

NO (figure 7), and some indication that neighbouring countries of asylum have more balanced male/female ratios than those further away (e.g. Rohingya in Bangladesh vs Malaysia, Somalis in Kenya and Ethiopia vs other countries of asylum).

Modelling implications:

1. Considerable variance over countries of asylum requires country of asylum as hierarchy element
2. Binary indicator of neighbouring country vs not neighbouring country is potentially a good covariate. Do we have a neighbouring country matrix available?
3. A more sophisticated gravity model style inclusion of covariates (physical distance, income differential) could help explain further variance
4. **Is the demographic composition of refugee populations in the same country of asylum similar over different countries of origin?**

NO (as expected, see figure 8). But – is there a lot less variance in the one high-income country (Germany) in the figure than in the others?

Modelling implications:

Two options: 1) one model per country of origin. Assumes that populations are isolated does not make use of demographic distribution of several origin populations to be similar in the same country of asylum (this is probably sensible for situations with recent displacement – however in protracted situations where refugee populations stay displaced in a country of asylum for decades, they might start assuming common fertility/mortality and become more similar?)

2) one model across the entire global population. Computationally challenging / impossible?

Process wise: start small, with one or a handful of origin countries and separate models.

1. **Are refugee populations from the same country of origin similar to each other in neighbouring countries of asylum and in countries in the same region?**

More similar in neighbouring countries, however there is little indication that region (UNHCR / SDG) alone is a good covariate for the demographic composition of an origin population once neighbouring country has been accounted for

Modelling implications:

No hierarchy on region of asylum

1. **Does the demographic distribution of a population from the same origin vary significantly across locations within the same country of asylum?**

Ad-hoc tests with Afghanistan and Syria origin: there is considerable variation sub-nationally with complete data, and we should consider including sub-national as hierarchy dimension. **Counterpoints**:

1. collection of sub-national data (i.e. the data structure itself) probably correlated with data availability and potentially demographics
2. when weighting average intra-country SDs by population size of country of asylum, intra-variance becomes much smaller

## Other comments:

* Fig 10-15: we should re-run these with complete data only to get a better idea of the significance of distance and neighbouring country

## Modelling key ideas

* Goal: predict the missing demographic distribution (counts or proportions of sex/age brackets within fixed total population counts) of refugee populations
* Instead of using the sex/age distribution of the global observed refugee population for the missing data, we try to predict the most plausible demographic composition of a refugee population for which we don’t have age/sex data by using additional information:
  + Which country of origin are they from
  + Which country of asylum are they in
  + Is the country of asylum a neighbouring country to the country of origin
  + Potentially other gravity-based variables such as income differential between countries of asylum and origin
* Total population counts are assumed to be observed without error
* Technical objective: joint model of missing data, observed data and parameters given covariates (hoping conditioning on covariates makes the missingness at random as opposed to not at random)

We want to predict the missing data values given observed data.

From <https://cran.r-project.org/web/packages/brms/vignettes/brms_missings.html#fn1>:   
*Actually, there is a third approach that only applies to* ***missings in response variables****. If we want to impute missing responses, we just fit the model using the observed responses and than impute the missings after fitting the model by means of posterior prediction. That is, we supply the predictor values corresponding to missing responses to the predict method.↩︎*

Since we have only missings in the response (demographic counts) variables, with the population total for rows with missing data available as offset in the binomial/multinomial model and potential covariates assumed to be fully observed, we will use this approach to explicitly predict the missing data values.

This means we will have two main steps:

1. Estimate posteriors of population parameters (e.g. probability vector of the multinomial given country of asylum and any covariates) with observed data only under prior assumptions
2. Predict missing data points by drawing from that posterior