**Calibration and open questions**

1. Calibration only, or calibration and validation
2. Choosing calibration metrics
3. Evaluating calibrations

**1) Calibration only, or calibration and validation**

While there is a range of possible calibration metrics at various scales (e.g., income diversity, income level, income source occupations) we are likely to focus on migration flow data, which we have from 2002-2011 from a sample generated by the Bangladesh Bureau of Statistics.

Validation involves calibrating the model on one part of the dataset, and testing the performance of the calibrated model on the other part of the dataset. K-fold cross-validation involves breaking the dataset into k parts, and one at a time removing one of the k parts, training the model on the remaining k-1 parts, and testing on the withheld part. The k different calibrations may be averaged or used as an ensemble.

In our case, the dataset includes 10 years of migration data – flows of migrants from each of 64 districts to each of the 63 other districts, disaggregated by gender and age groups. It SEEMS like a lot of data against which to both calibrate and validate. However, it is not a large number of independent data points. Migration flows within the same year are not truly independent from each other, nor are migration years independent samples of the same phenomenon – they are likely auto-correlated, as well as driven by different factors.

However, validation is an important test of the model’s predictive ability, in a way that calibration simply isn’t. It’s worth doing if we can. The open question is whether there is a meaningful way to split the dataset (e.g., train data on all years but one and test on the other; randomly sample flows from within a single year/across multiple years to train data, and test on a different random sample of flows).

At present, I have not done this. I have taken the average across all years of inter-district flows, aggregating across gender and age, and am using for calibration only. Open for discussion.

**2) Choosing calibration metrics**

Even focusing in on migration flows, there are a number of different measures we can choose as the basis for evaluating calibrations. In particular, we can choose to calibrate on i) relative migration rates (i.e., fraction of all migrations occurring between particular districts), which discards information on the total volume of migrations, or ii) absolute migration rates.

The reason this is a difficult decision is that while many people in a population may migrate, MOST people do not migrate – migrants are a very small fraction of a population (~1-2%), and MIDAS simulations include only a few thousand agents. Getting the absolute flows of migrants right means calibrating such that only a handful of agents migrate, making it difficult to capture all of the different inter-district flows. An alternative is to calibrate based on relative migration rates, allowing many more migrations to happen per agent in the simulation in order to achieve better resolution of differences across inter-district corridors, and consider the population of the MIDAS simulation to be a non-representative sample focused more closely on migrants.

Or, we can do both, or some weighted mix of both. Open for discussion.

An additional point of consideration is the appropriate unit of analysis for the calibration. Our data is organized by inter-district flow, making that an obvious choice, but the problem is that some flows are very small and likely to be noisy from year to year – unlike flows to and from Dhaka for instance, which will be more consistent. A workaround is to weight the flows by population – by the source population, destination population, or both. We can consider all of these options.

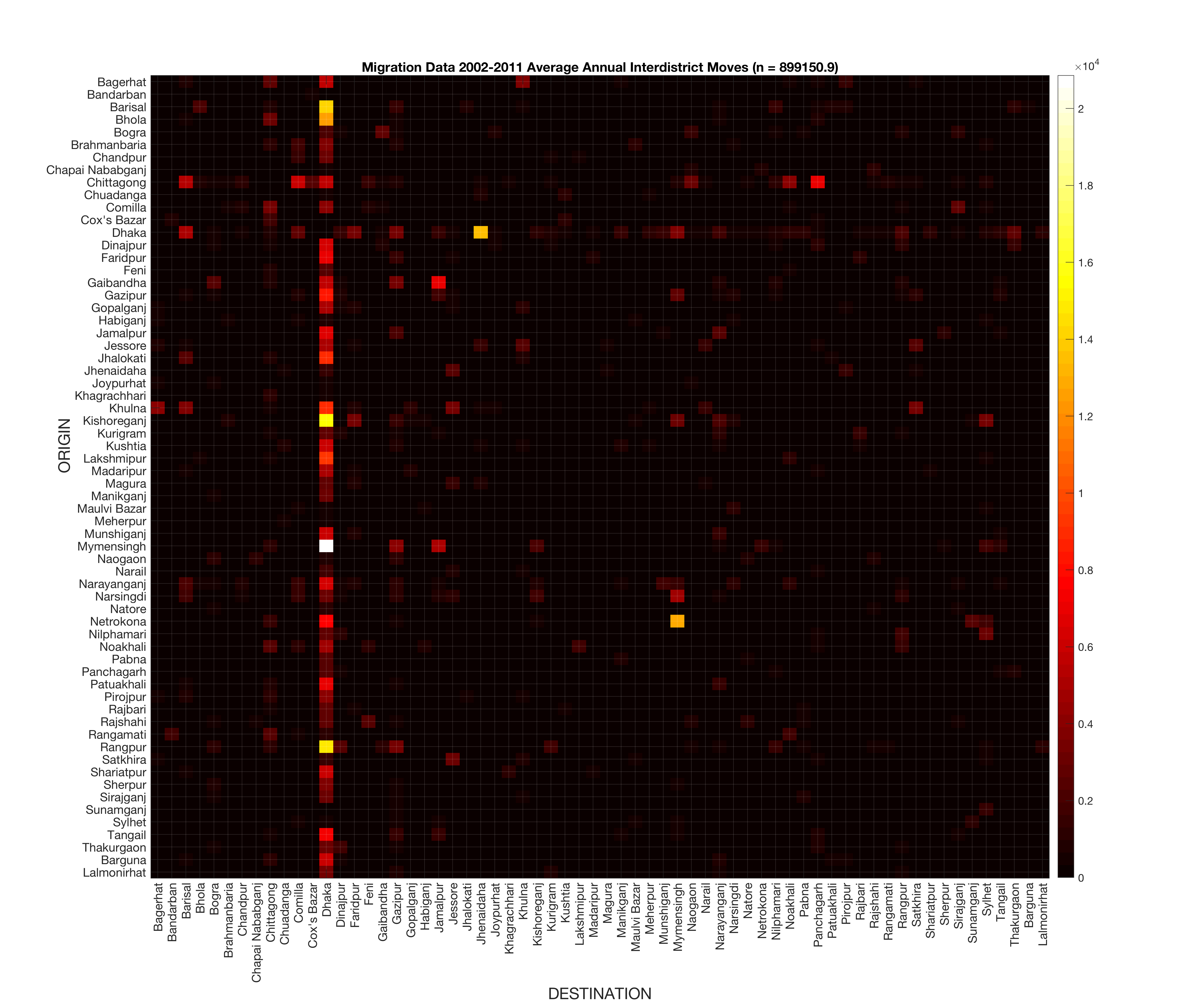
**3) Evaluating calibrations**

Once we have selected appropriate metrics and approaches, we have to decide how to evaluate calibrations.

The simplest approach is to select calibrations that minimize the sum of squared, weighted errors by interdistrict flow, for the chosen metric.

A related approach is to use the weighted Pearson’s r2, though it is worth highlighting that this metric picks up only correlation, ignoring differences in scale, and would be inappropriate for selecting calibrations to best match the absolute migration rate.

First, consider the dataset that is being used to evaluate calibrations – the average annual interdistrict flows from 2002-2011, aggregating across age groups and gender, as identified from the Bangladesh Bureau of Statistics sample (Figure 1).

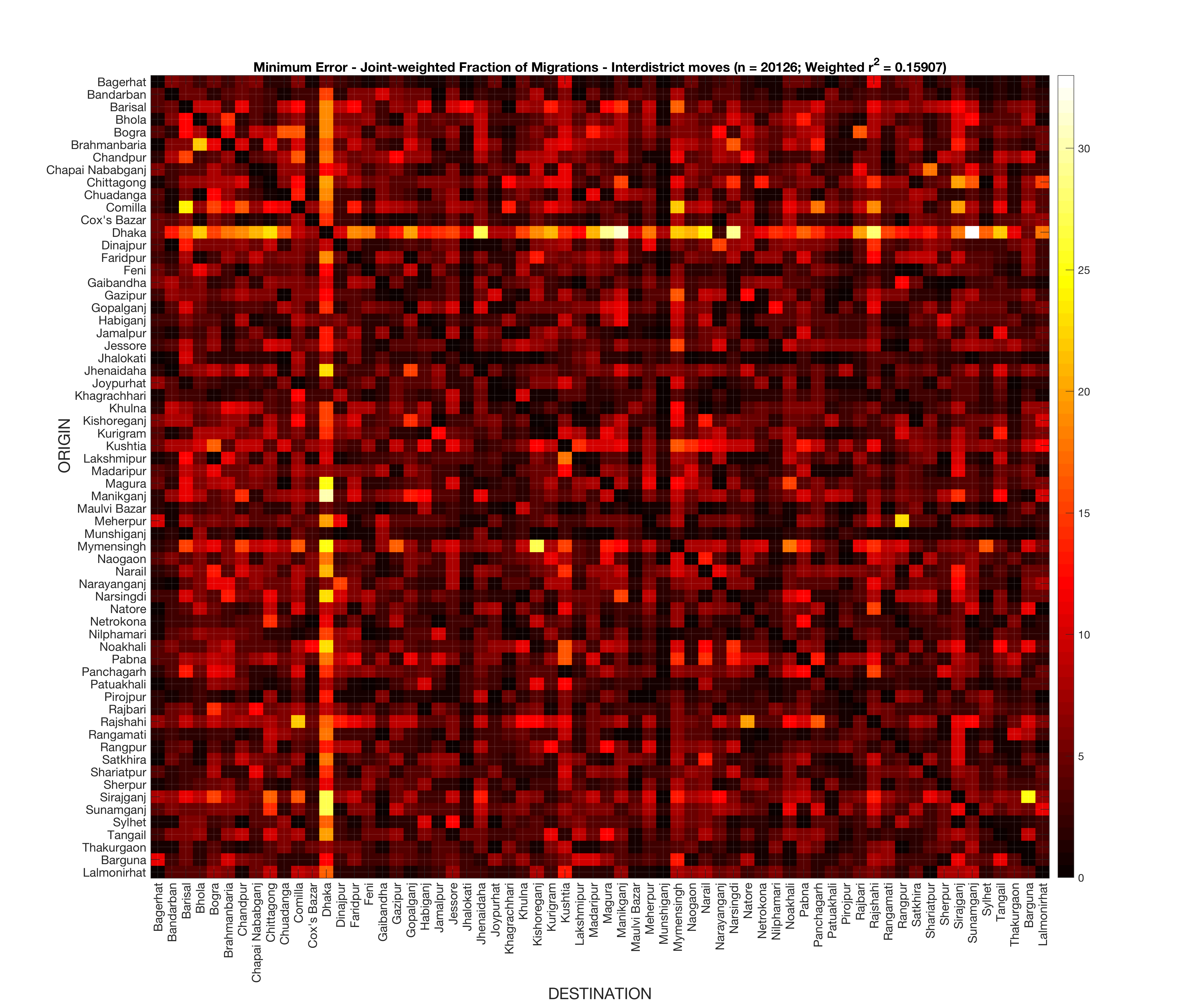


**Figure 1 – Observed average interdistrict migration rates from 2002-2011, aggregated across age and gender**

Note that these values are estimated by a sample and scaled up to the population, so that in many cases there are values of 0 for interdistrict flows that may be non-zero, only because no member of the sample participated in that flow.

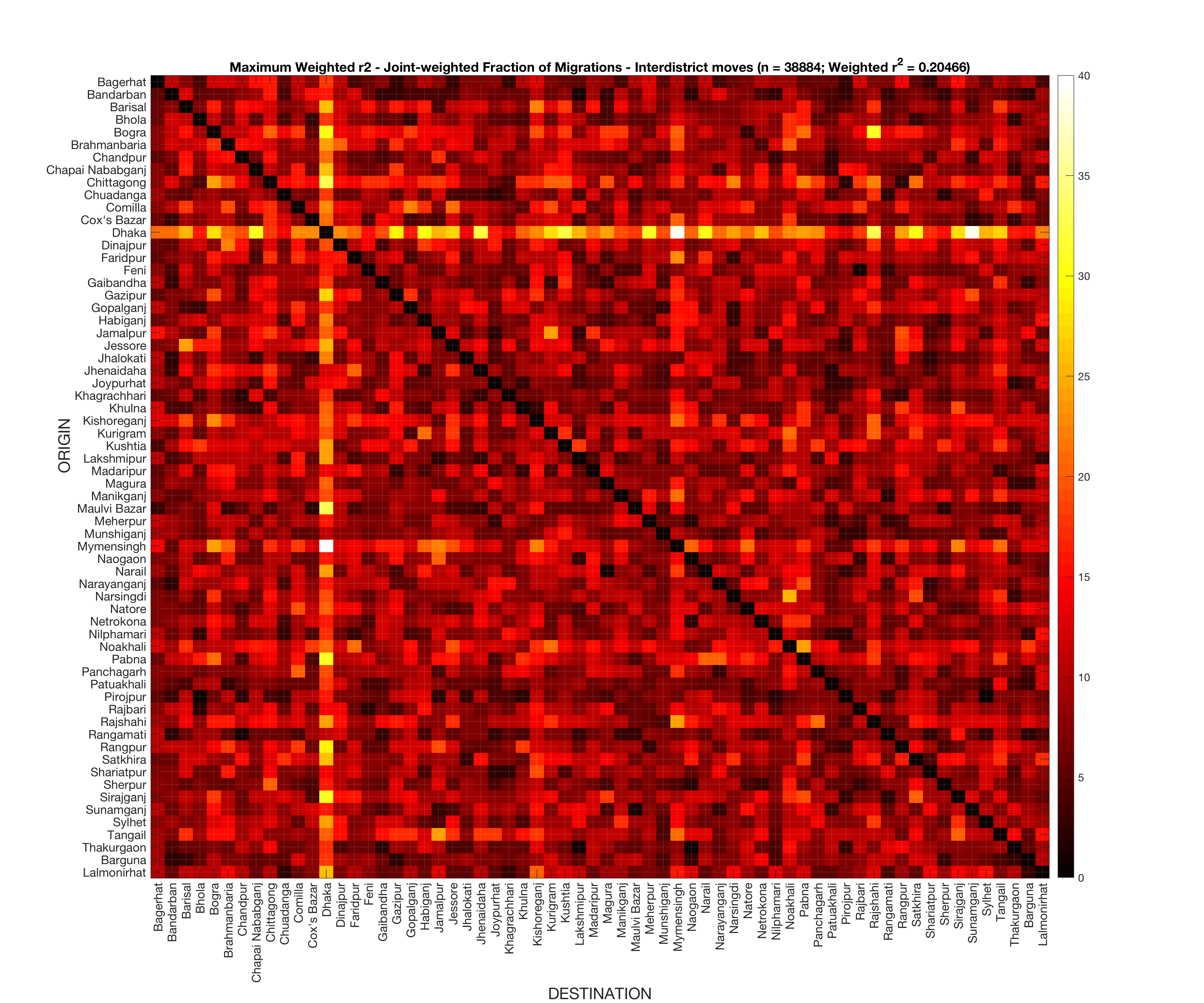
For any calibration metric A (relative migration rates or absolute migration rates, e.g.), we can estimate a weighted sum of squares error as:

From a pool of 3776 Monte Carlo calibrations developed from Nov 28-30, I have found the following calibration to minimize SSE for relative migration rates:



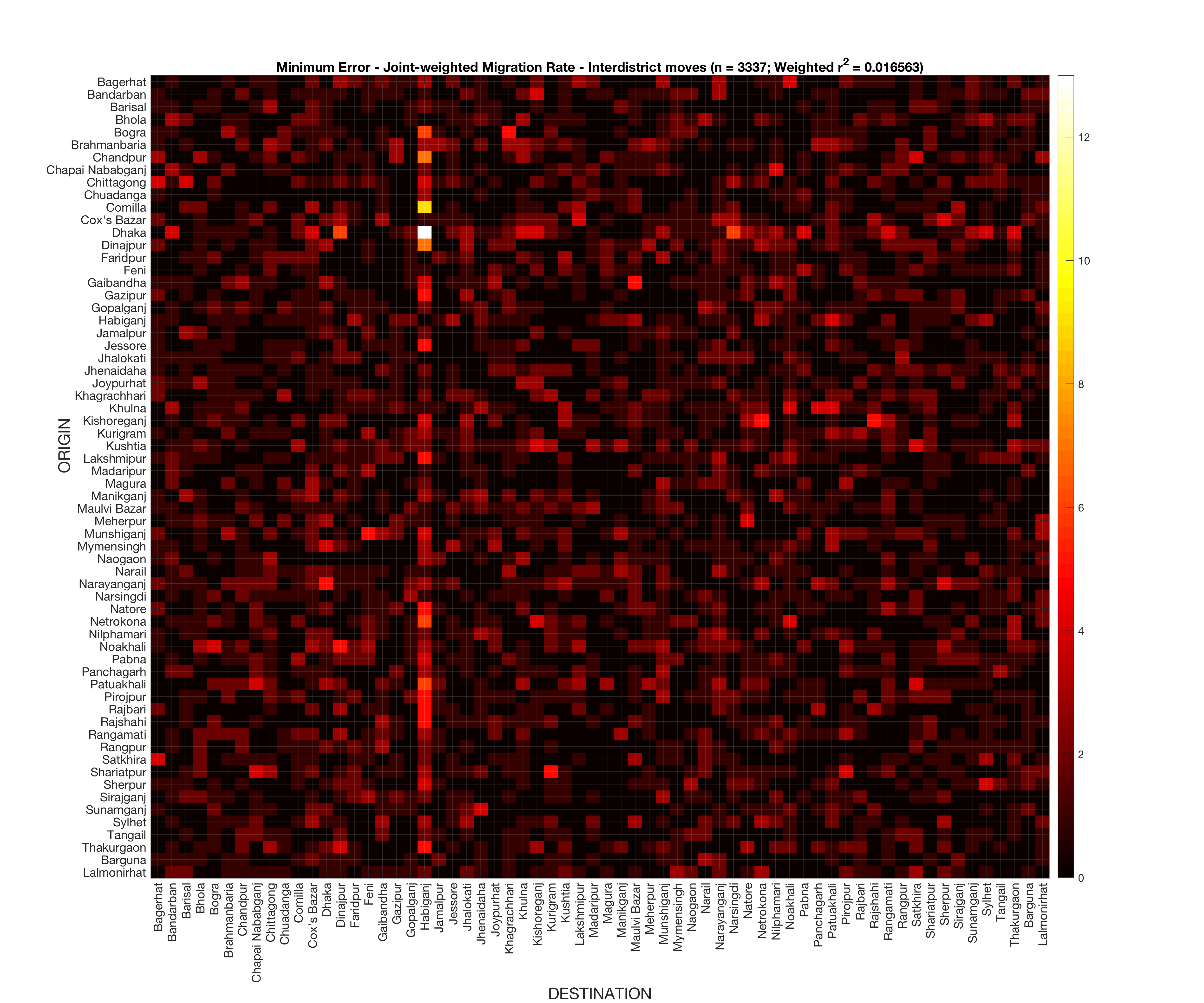
**Figure 2 – Simulated average annual interdistrict migration rates, from a calibration minimizing SSE of relative migration rates**

A related but not identical criterion is the maximized weighted r2 (different I think because the weights apply in multiple different parts of the calculation) for relative migration rates:



**Figure 3 – Simulated average annual interdistrict migration rates, from a calibration maximizing weighted Pearson r2 of relative migration rates**

Additionally, I have the following calibration to minimize SSE for absolute migration rates:

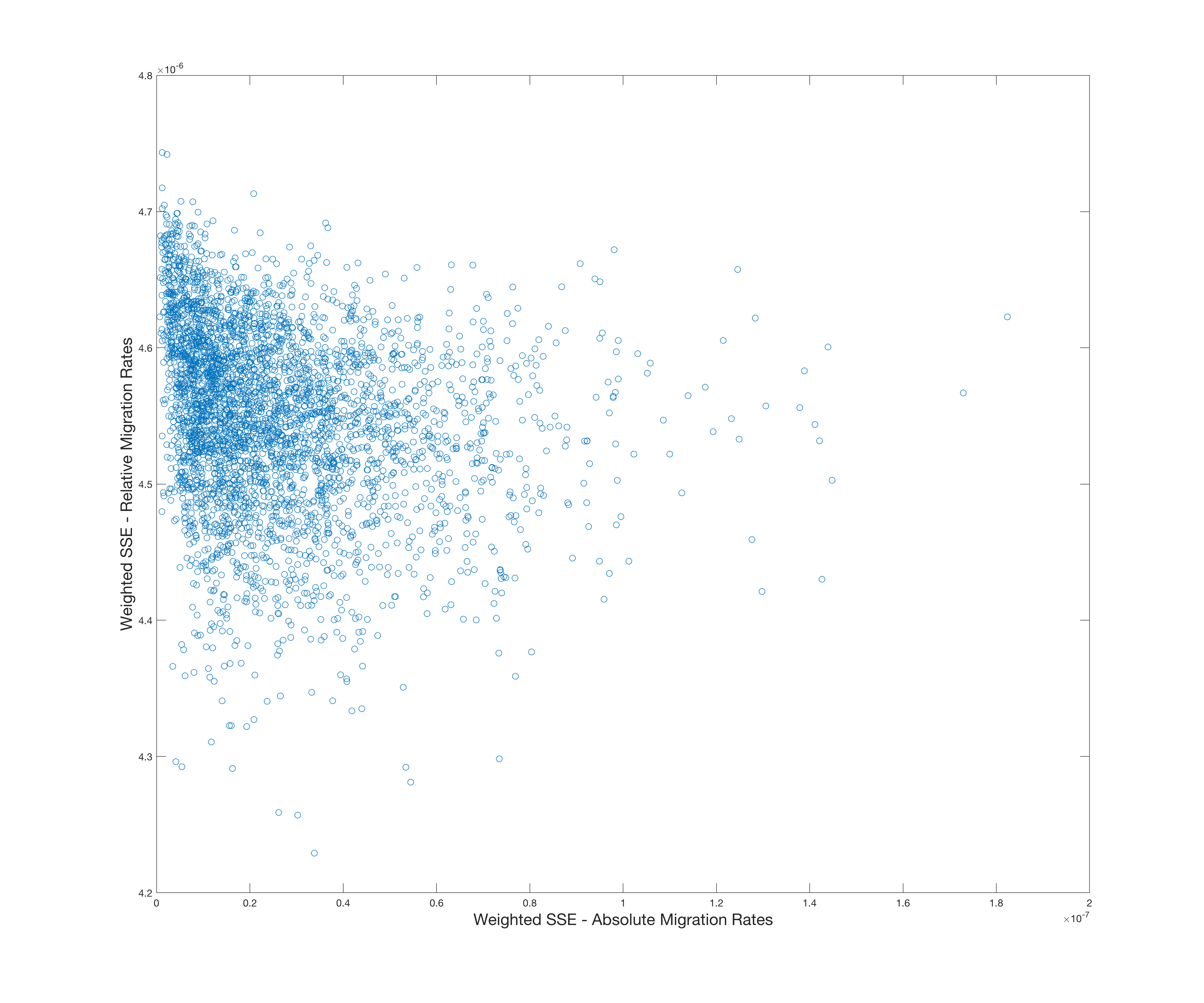


**Figure 4 – Simulated average annual interdistrict migration rates, from a calibration minimizing SSE of absolute migration rates**

There are some things to highlight here.

First, if we feel it appropriate to do so, and are agreed on model structure and parameterization, we can make this procedure Bayesian by identifying the distribution of the top fraction of calibrations (the top 5 or 10%, for example), and using those distributions to develop the next pool of possible calibrations. This is Approximate Bayesian Calibration (ABC). We don’t have to do this if we don’t want to.

Second, we draw different kinds of simulations by using relative vs absolute migration rates as calibration metrics. In the former, we better capture the relative sizes of migration flows but overpredict migration rates in the population. In the latter, we better capture overall migration rates but at the expense of getting the sources and destinations right – we simply extinguish migration. In this particular case, the main migration hub is also wrong – Habiganj shows up in the income data as having high income rates, though it is actually much smaller than Dhaka (which does not have particularly high income rates but has many opportunities).

This is not a mystery. We are only including a narrow range of drivers (income sources) in our decision model, knowing that a wealth of other factors (assets, ties, place utility, etc.) for which we do not have data are simultaneously shaping the decision. My interpretation of the differences in calibration is that by calibrating to relative migration rates, we get migrant risk characteristics and preference right but lack the key factors that would keep many of them from migrating. By calibrating to absolute migration rate, we incorrectly calibrate preferences to capture the basic idea that most people don’t migrate, but in this case we don’t get the who or the where correct. To illustrate this more graphically, compare the SSE for relative vs. absolute migrations, across all members of the calibration pool:

**Figure 5 – Weighted SSE for relative migration rates plotted against weighted SSE for absolute migration rates**

The key message – SSE for absolute migration rates is minimized when error for relative migration rates is maximized – they are not capturing the same thing. In particular, the most important parameters shaping SSE for absolute migration are the mean r value (higher value means riskier behavior) and the likelihood of updating portfolio preferences in a timestep (Figure A2, identified using randomForest to identify relative variable importance in prediction). For both of these variables, SSE for absolute migration rates is minimized when these variables are minimized – i.e., extinguishing agent preferences to take risks or consider moves (Figures A3, A4). In contrast, the most important parameter for shaping SSE for relative migration rates is also the r value mean (Figure A1), but errors are minimized and remain lower when the mean r value is higher (Figure A5).

One line of work is to improve representation of the decision and options in our simulation – adding in assets, ties to place, etc. We should do this, but this is a longer-term endeavor.

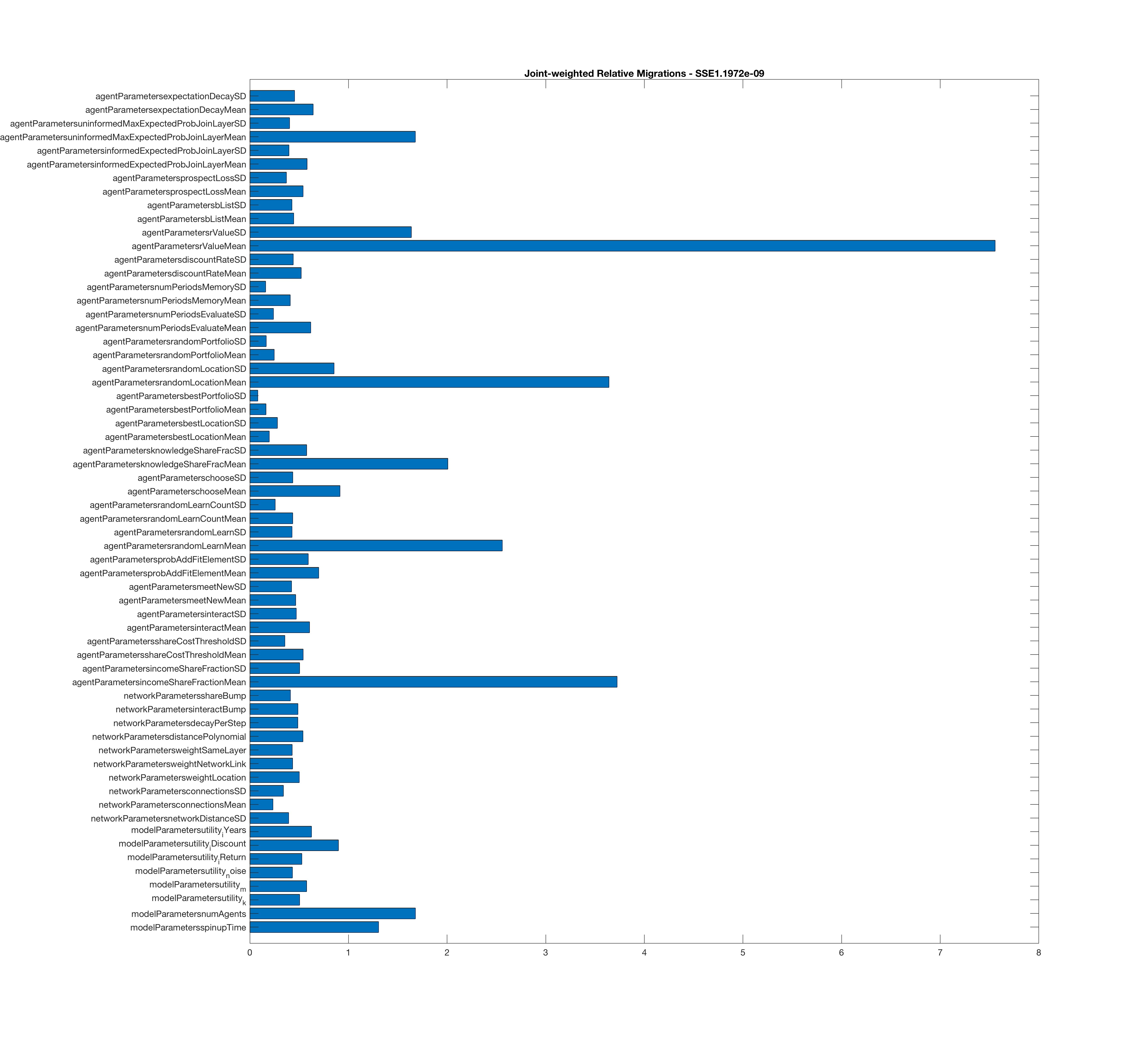
A separate line of work is to take what we have as a reasonable base model and choose appropriate calibrations. One possible approach, open to discussion is to choose a handful of the best calibrations of relative migration rate, interpret them as over-representing the migrants in the sample, and scale the predicted migration rates in any future projection by (Observed annual migration rate in data) / (Simulated annual migration rate in calibration exercise).

If we believe there is merit in also trying to capture the absolute rates correctly from the start, we can collect some of the best calibrations from this metric as well for inclusion in an ensemble. More generally, we can establish a range of values for a weight W that gives an overall SSE score as:

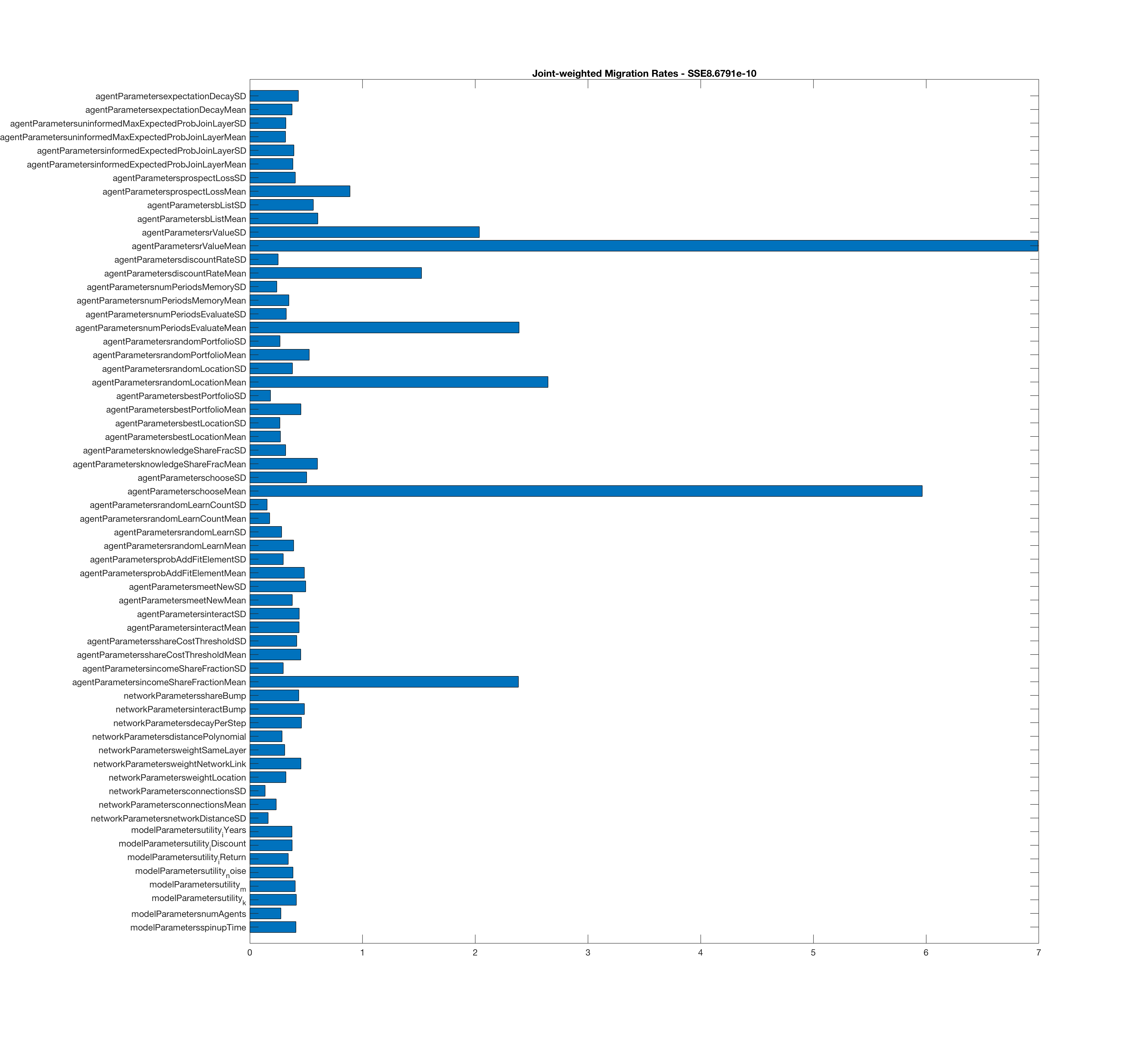
Open for discussion.

When we have reached paths forward for all of the questions raised in this document, we are ready to move on to applying this calibrated case to future projections of development and SLR.

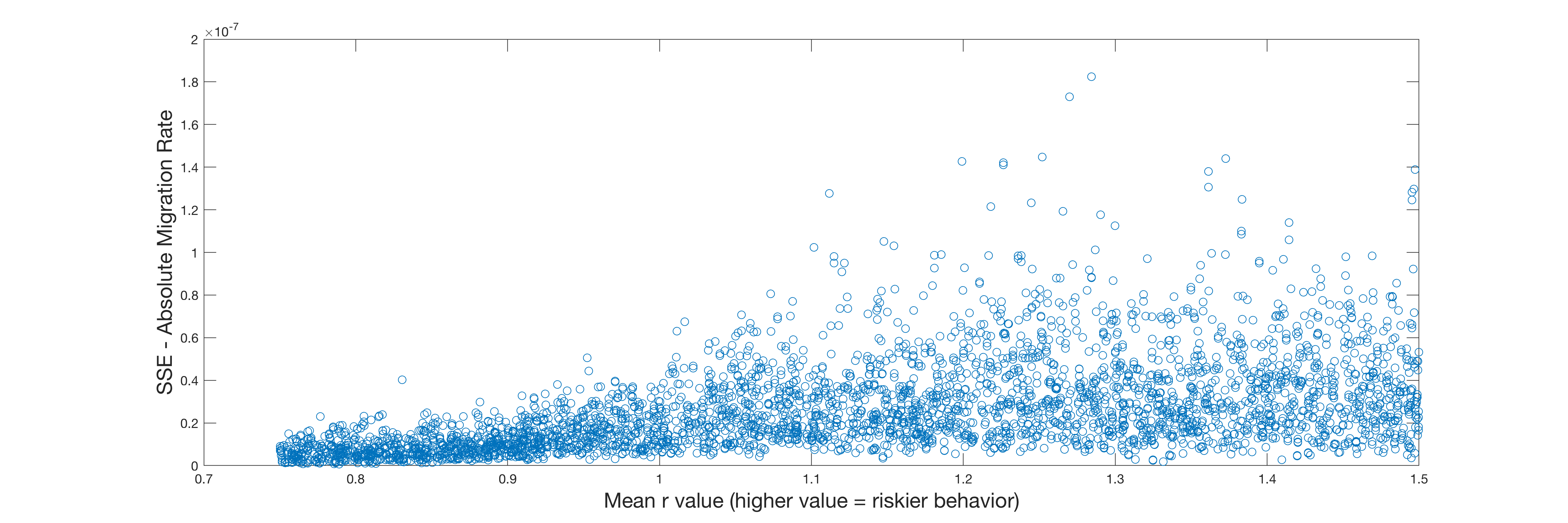
**Appendix – additional figures**



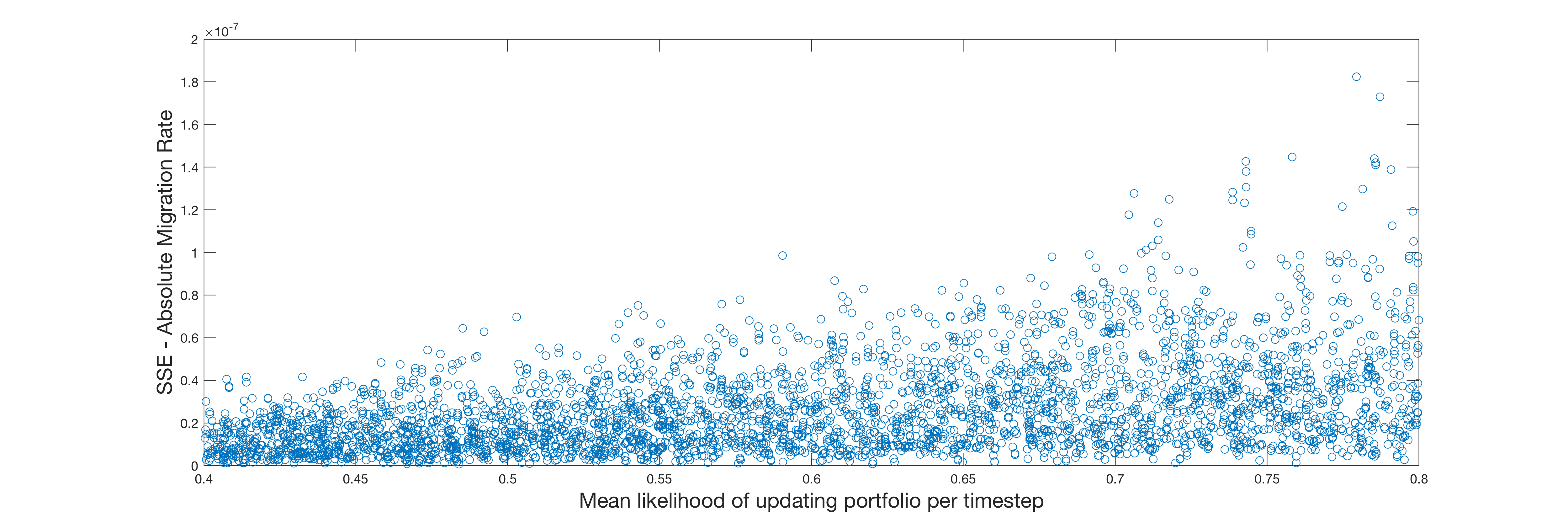
**Figure A1 – Relative variable importance in minimizing error in relative migration rates (as relative change in out-of-bag prediction error when variable is excluded) from a random forest of 100 regression trees**

****

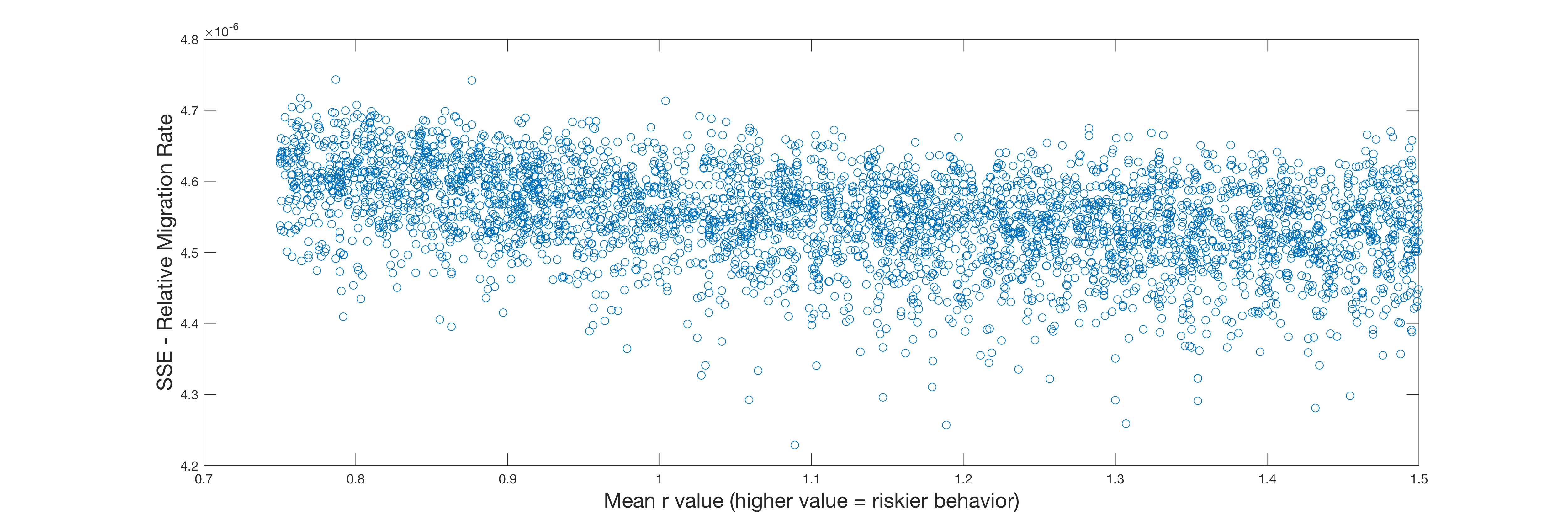
**Figure A2 – Relative variable importance in minimizing error in absolute migration rates (as relative change in out-of-bag prediction error when variable is excluded) from a random forest of 100 regression trees**

****

**Figure A3 – SSE in absolute migration rate as function of mean r Value**

****

**Figure A4 – SSE in absolute migration rate as function of likelihood of re-evaluating portfolio in a timestep**

****

**Figure A5 – SSE in relative migration rate as function of mean r Value**