Learning about Internal Migration from Half a Billion Individual Records: Applying Localized Classification Trees to Large-Scale Census Data

Guy J. Abel[1,3,4]

Raya Muttarak[2,3,4]

Fabian Stephany[3]

1 Asian Demographic Research Institute, Shanghai University, China
2 School of International Development, University of East Anglia, UK
3 Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW and WU), Austria
4 International Institute for Applied Systems Analysis, Austria

Abstract

Understanding who migrates is crucial in explaining societal changes and forecasting future population composition and size. However, there is no empirical consensus on demographic and socioeconomic factors driving migration decision. Exploiting micro census data from the Integrated Public Use Microdata Series International (IPUMSI) database across 65 countries over the period 1960 to 2012 covering 477,296,432 individual records, this study aims to establish common demographic drivers of migration. Given an exceptionally large number of observations, a parametric approach would simply yield bias estimates of standard errors of the variables of interest. We apply a machine learning technique using decision tree models to establish common demographic patterns driving migration in our data. We find that globally, age, education, household size, and urbanisation are important drivers of internal migration. Age and education are particularly important predictors in Europe and Northern America whilst in South and Central America and Africa, urbanisation and household size are more relevant.

1. Introduction

Migration is a key demographic behaviour responsible for population distribution and redistribution in sending and receiving areas. Internal migration in China, for instance, is the key contributor of rapid urbanisation in the country where rural-urban migrants accounting for 78% of China's urban population since 1979 (Su et al. 2018; Zhang and Song 2003). Migration however is not a random process. Migrants are selected for a number of characteristics that make them differ from nonmigrant populations. Economic migrants, for example, are likely to be positively selected along characteristics underlying labour market abilities such as age, education and health (Grogger and Hanson 2011). Family-related migration, on the other hand, is more likely to involve females (Kudo 2015). This migrant selectivity, when occurring in a large scale, can contribute to demographic shifts in both the origin and destination (Findlay and Wahba 2013). Understanding who the migrants are thus is useful in explaining societal changes and projecting future population composition and size.

Whilst the evidence on the role of age on migration is fairly consistent (Rogers and Castro 1981), there is no empirical regularity regading other demographic factors such as gender and education. Quantiative studies of gender and migration are scarce compared to studies using ethnographic methods (Donato et al. 2006). As a consequence, there is no consensus whether gender is a key demographic characteristic driving migration and if so, in what direction. Recent studies using micro-level data pointed to a feminisation of migration, that is higher likelihood of migration for women as compared to men (Camlin et al. 2014; Reed et al. 2010). However, at the global level, total migration flows are still higher in men than in women over the period 1970 to 2010. There is also evidence that the increase of migration flows is faster for men than for women during the period 2000 to 2010 (Abel 2018). Given the unsettled evidence, it remains unclear to what extent gender matters in determining migration.

Similarly, there is no conclusive evidence on whether migrants are drawn upon a pool of less or more educated individuals. Empirical studies show that the direction of selectivity by educational level differs across countries (Cattaneo 2007; Gould 1982). Previous studies at the individual level provide inconsistent evidence on the relationship between education and the propensity to migrate. On the one hand, a series of studies reported a positive effect of educational attainment on the likelihood of migration (Donato 1993; Stark and Taylor 1991; Williams 2009; Yang and Guo 1999) . On the other hand, many studies found a negative relationship between education and migration (Massey et al. 1987; Massey and Espinosa 1997; Quinn and Rubb 2005) and some studies reported no significant association at all (Adams 1993; Curran and Rivero-Fuentes 2003). It is thus remains unsettled which direction education affects migration and to what extent.

The absence of empirical regularity regarding how demographic factors drive migration is partly due to variation in geographical areas considered as well as the scale of the study which may not always be nationally representative. Especially for internal migration, large-scale cross-national studies on drivers of migration is limited. Data scarcity and complexity in collecting migration data make it difficult to compare internal migration patterns across countries, not to mention understanding who the migrants are (Bell et al. 2002). The lack of understanding about the common characteristics driving migration consequently makes it difficult to predict how migration will look like in the future.

Exploiting micro census data from the Integrated Public Use Microdata Series International (IPUMSI) database, we are able to analyse common factors determining internal migration for 63 countries over the period 1960 to 2012. Indeed, the potential of using the IPUMS database for comparative migration research between places, over time and across population subgroups has been recognised (Sobek 2016). However, the large volume of data obtained from the harmonised micro census data with xxxx million individual records in our case making it inappropriate to estimate migration drivers using classical regression methods. This requires a new approach to handle and analyse big data like ours.

To that end, we employ localised classification trees – a data mining method commonly used for establishing classification systems in large datasets. This non-parametric technique allows us to identify patterns in the data without imposing any statistical assumptions and being sensitive to misuse of significance testing like in conventional multiple regression models (Attewell et al. 2015). To the best of our knowledge, this is the first time classification trees method is applied to study drivers of migration.

2. Data and measurement

2.1 Data

Migration and socio demographic data are derived from harmonized census microdata samples from the Integrated Public Use Microdata Series International (IPUMSI) database (Minnesota Population Center 2015). Each set of census microdata contains a small random sample (0.4%-10%) of unidentified private households and associated persons based on a full census conducted by the national statistical agency in each country. The countries and years used in this study, shown in Figure 1, are based on censuses collected between 1960 and 2012. In total, data from 477,296,432 individual records in 65 countries and 190 censuses we used as the basis for our analysis.

Our eligibility criteria for including countries in our analysis was based on the availability of migration, age, gender and education measures in the IPUMSI, where all three measures were required in order to derive bilateral migration flows between regions by gender and educational attainment.

One advantage of using the IPUMSI database is that potential explanatory variables of migration are both available and standardized to allow for cross-country comparisons. However, the geographical detail available for each country is not uniform and depends on the density of the sample size, the distribution of the population and the administrative units in place.

2.2 Migration Measurement

An indicator measure to determine if individuals are migrants were based on a combination of IPUMSI migration variables depending on availability. These can be separated into two approaches. The first used a two questions on the length of stay in current location and whether the previous residence was in the same administrative unit. We coded an individual migrant as migrant from those who had changed their major administrative unit during the year previous to the census date. The second approach used a question asking respondents for their place of residence at a fixed time interval (such as place of residence five years ago). When the major administrative unit for the previous residence differed from the administrative unit at the time of the census we coded the individual as a migrant. Where responses to all sets of migration questions were available we gave preference to the former. In the latter, where multiple fixed interval measures were available we gave preference to the shortest of interval.

Other measures used in our study, on age, sex, education, employment, marital status, children ever born, household size and if the household was in an urban area were based directly on the IPUMSI measures.

3. Method

This work employs a classification technique often called classification and regression trees (CARTs). Generally speaking[1], classification trees apply a systematic truncation of a data sample with regard to the distribution an outcome variable (target feature). The target feature in our case is internal migration, which distinguishes each individual observation between migrants (labelled 1) and non-migrants (labelled 0). With regard to this target feature, the initial sample has a certain distribution. For this initial distribution, a measure of data purity, in our case the Gini impurity index, is calculated. The Gini impurity is a measure of how often a randomly chosen element from the sample would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the sample. This measure serves as a benchmark for the subsequent steps.

In the next step, external features[2] are considered for the truncation of the initial sample. The data purity of the two resulting subsets (nodes) is compared. The external feature that yields the highest level of data purity in the remaining subsets is chosen as a splitting criterion. After this split, the initial procedure is repeated for each of the resulting subsamples: a tree like, cascading structure emerges after repeating the procedure several times, as it is illustrated in figure one. For the sake of comparability, we decided to limit the growth of the tree to four subsequent splits (“pruning”).

[Figure 1 about here]

In comparison to parametric methods, like logistic regressions, the method of classification trees has the advantage of considering all possible features (covariates) simultaneously. At the same time, the problem of “greediness” arises. This means that the classification method selects only the one splitting criterion that results in higher data purity at each single point of the tree. When applying classification trees to a large and diverse sample, like given in our case with 65 countries over 52 years, local distributions of the target feature could remain undetected. For example, in our global sample, age could be the most relevant characteristic when it comes to migration. However, for some countries, at a certain point in time, education might be more influential.

[1] A more elaborate description of CARTS can be found in Friedman, Hastie & Tibshirani (2013).

[2] Namely, sex, age, education, employment status, number of children, marital status, household size, and a measure of urbanization.

In order to overcome this limitation, we perform individual classification trees for each single country-year sample. At a second stage, we compare the structure of the individual trees. Here, the time of occurrence of features as a splitting criterion is relevant. The earlier a feature is chosen as a criterion for a split, the more it corresponds with the distribution of the target feature. This means that features that are chosen at an early stage of the tree are more closely related to internal migration than features that appear at a later splitting point. This descriptive analysis of localized trees helps to detect differences in the relevant migration determinants over time and space.

4. Results

For ease of interpretation the results of our analysis are shown in form of a heat map. Figure 2 lists all 190 country-year combinations vertically and groups them into 17 broad geographical regions. In the horizontal direction, the external features are listed. For each country-year and feature cells have different colours. If the feature appears as the first splitting criterion of the classification tree, the cell is marked in deep red. The later the feature appears in the splitting hierarchy of the tree, the lighter the colour becomes[3].

[Figure 2 about here]

The heat map illustration eases the comparison of results between countries from a broad perspective. In the following a set of general findings is presented. Overall, age is a prominent feature. However, age is not equally relevant in all regions. In contrast to most other regions, age is only of secondary importance in African countries. In Eastern and Western Africa, for example age never appears as a first splitting criterion. In Europe and Northern America, age is very often selected as the first feature in the tree. In Southern and Western Europe, besides age, gender and the level of education are most commonly associated with internal migration. In fact, only occasionally are other features, such as employment status or household size selected at all. In South and Central America, on the other hand, education is less dominant as a splitting feature. With the exception of Mexico, age, household size and the degree of urbanisation are the predominant features. It is only the case for three out of 60 country-years in South and Central America that none of these three features is selected at the first split.

[Figure 3 about here]

[3] The number of possible splitting features doubles after every split. This means that there can only be one feature for the first split, two for the second, four for the third, and eight for the fourth split. White cells indicate that a feature is not chosen up until the fourth split. However, since all trees are pruned after split four, it might be that some features could have occurred at a later split in the tree. Grey cells indicate that the feature has not been available in the survey of the respective country-year.

The prominent role of age is worth a more detailed investigation. Figure 3replicates the heat map structure. However, now the horizontal categories indicate when the feature age appears in the classification tree. The exact age which has been selected by the algorithm as the defining cut-off value are represented by the number inside the cell and the cell's colour. At first sight, one notices that the boxes for Southern and Western Europe, as well as Northern America and South-Eastern Asia are more populated with coloured cells than for regions of South and Central America or Africa. This can be explained by the fact that features, such as education, household size or urbanisation are more relevant in Africa and in South and Central American countries. In Europe, on the other hand, these characteristics are less important. Their place in the tree is taken instead by the age feature. A closer look at the precise age reveals that the cut-off value for the first and second split differs across regions. In African regions, the average cut-off age, for the first three nodes (split one and split two), is 28, while in Western Europe and North America, the average age for these splits is 37.

5. Conclusion

Globally, age, education, household size, and urbanisation are important drivers of internal migration. However, in Europe and Northern America age and education play a more important role than in South and Central America or Africa. Here, the migration pressure of urbanisation and household size are more relevant. Internal migrants are in their mid-twenties in Africa, about a decade younger than in the Global North. Non-parametric techniques, like classifications trees, are a helpful tool for the exploration of large-scale migration data. The presented case shows that with the help of localized classification trees, general patterns across time and space become visible. This could become a helpful tool for a pre-examination of multiple survey datasets in other disciplines.

References

Abel, G. J. (2018). Estimates of Global Bilateral Migration Flows by Gender between 1960 and 20151. *International Migration Review*, imre.12327. doi:10.1111/imre.12327

Adams, R. H. (1993). The economic and demographic determinants of international migration in Rural Egypt. *The Journal of Development Studies*, *30*(1), 146–167. doi:10.1080/00220389308422308

Attewell, P., Monaghan, D. B., & Kwong, D. (2015). *Data Mining for the Social Sciences: An Introduction* (1st ed.). University of California Press.

https://www.jstor.org/stable/10.1525/j.ctt13x1gcg. Accessed 12 September 2018

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., & Hugo, G. (2002). Cross-national comparison of internal migration: issues and measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *165*(3), 435–464. doi:10.1111/1467-985X.t01-1-00247

Camlin, C. S., Snow, R. C., & Hosegood, V. (2014). Gendered patterns of migration in rural South Africa. *Population, Space and Place*, *20*(6), 528–551. doi:10.1002/psp.1794

Cattaneo, C. (2007). *The Self-Selection in the Migration Process: What Can We Learn?* (No. 199). *LIUC Papers in Economics*.

Curran, S. R., & Rivero-Fuentes, E. (2003). Engendering migrant networks: The case of Mexican migration. *Demography*, *40*(2), 289–307. doi:10.1353/dem.2003.0011

Donato, K. M. (1993). Current trends and patterns of female migration: evidence from Mexico. *The International Migration Review*, *27*(4), 748–771.

Donato, K. M., Gabaccia, D., Holdaway, J., Manalansan, M., & Pessar, P. R. (2006). A Glass Half Full? Gender in Migration Studies. *International Migration Review*, *40*(1), 3–26. doi:10.1111/j.1747-7379.2006.00001.x

Findlay, A. M., & Wahba, J. (2013). Migration and Demographic Change. *Population, Space and Place*, *19*(6), 651–656. doi:10.1002/psp.1786

Gould, W. T. S. (1982). Education and internal migration: A review and report. *International Journal of Educational Development*, *1*(3), 103–111.

Grogger, J., & Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics*, *95*(1), 42–57. doi:10.1016/j.jdeveco.2010.06.003

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Kudo, Y. (2015). Female Migration for Marriage: Implications from the Land Reform in Rural Tanzania. *World Development, 65,* 41–61. doi:10.1016/j.worlddev.2014.05.029

Massey, D. S., Alarcón, R., Durand, J., & González, H. (1987). *Return to Aztlan: The Social Process of International Migration from Western Mexico*. Berkeley; Los Angeles; London: University of California Press. http://www.jstor.org/stable/10.1525/j.ctt1ppp3j. Accessed 17 April 2017

Massey, D. S., & Espinosa, K. E. (1997). What's Driving Mexico-U.S. Migration? A Theoretical, Empirical, and Policy Analysis. *American Journal of Sociology*, *102*(4), 939–999.

Quinn, M. A., & Rubb, S. (2005). The importance of education-occupation matching in migration decisions. *Demography*, *42*(1), 153–167.

Reed, H. E., Andrzejewski, C. S., & White, M. J. (2010). Men's and women's migration in coastal Ghana: An event history analysis. *Demographic Research*, *22*(25), 771–812. doi:10.4054/DemRes.2010.22.25

Rogers, A., & Castro, L. J. (1981). Model migration schedules. *IIASA Research Report*, *81*(RR-81-30), 1–160.

Sobek, M. (2016). Data prospects: IPUMS-International. In M. J. White (Ed.), *International Handbook of Migration and Population Distribution* (pp. 157–174). New York: Springer Netherlands. //www.springer.com/gb/book/9789401772815. Accessed 12 September 2018

Stark, O., & Taylor, J. E. (1991). Migration Incentives, Migration Types: The Role of Relative Deprivation. *The Economic Journal*, *101*(408), 1163–1178. doi:10.2307/2234433

Su, Y., Tesfazion, P., & Zhao, Z. (2018). Where are the migrants from? Inter- vs. intra-provincial rural-urban migration in China. *China Economic Review*, *47*, 142–155. doi:10.1016/j.chieco.2017.09.004

Williams, N. (2009). Education, gender and migration in the context of social change. *Social science research*, *38*(4), 883–896.

Yang, X., & Guo, F. (1999). Gender Differences in Determinants of Temporary Labor Migration in China: A Multilevel Analysis. *The International Migration Review*, *33*(4), 929–953. doi:10.2307/2547358

Zhang, K. H., & Song, S. (2003). Rural–urban migration and urbanization in China: Evidence from time-series and cross-section analyses. *China Economic Review*, *14*(4), 386–400. doi:10.1016/j.chieco.2003.09.018

Figures

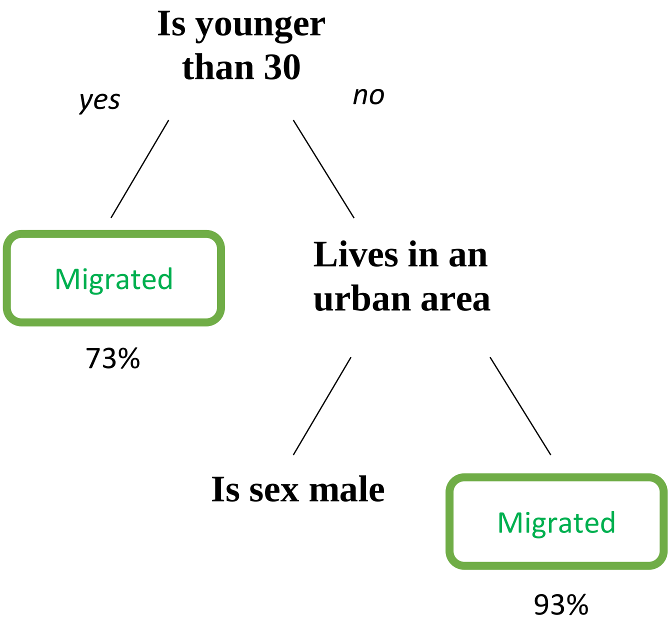

Figure 1: The algorithm truncates the sample at each node with regard to the feature criterion migraiton

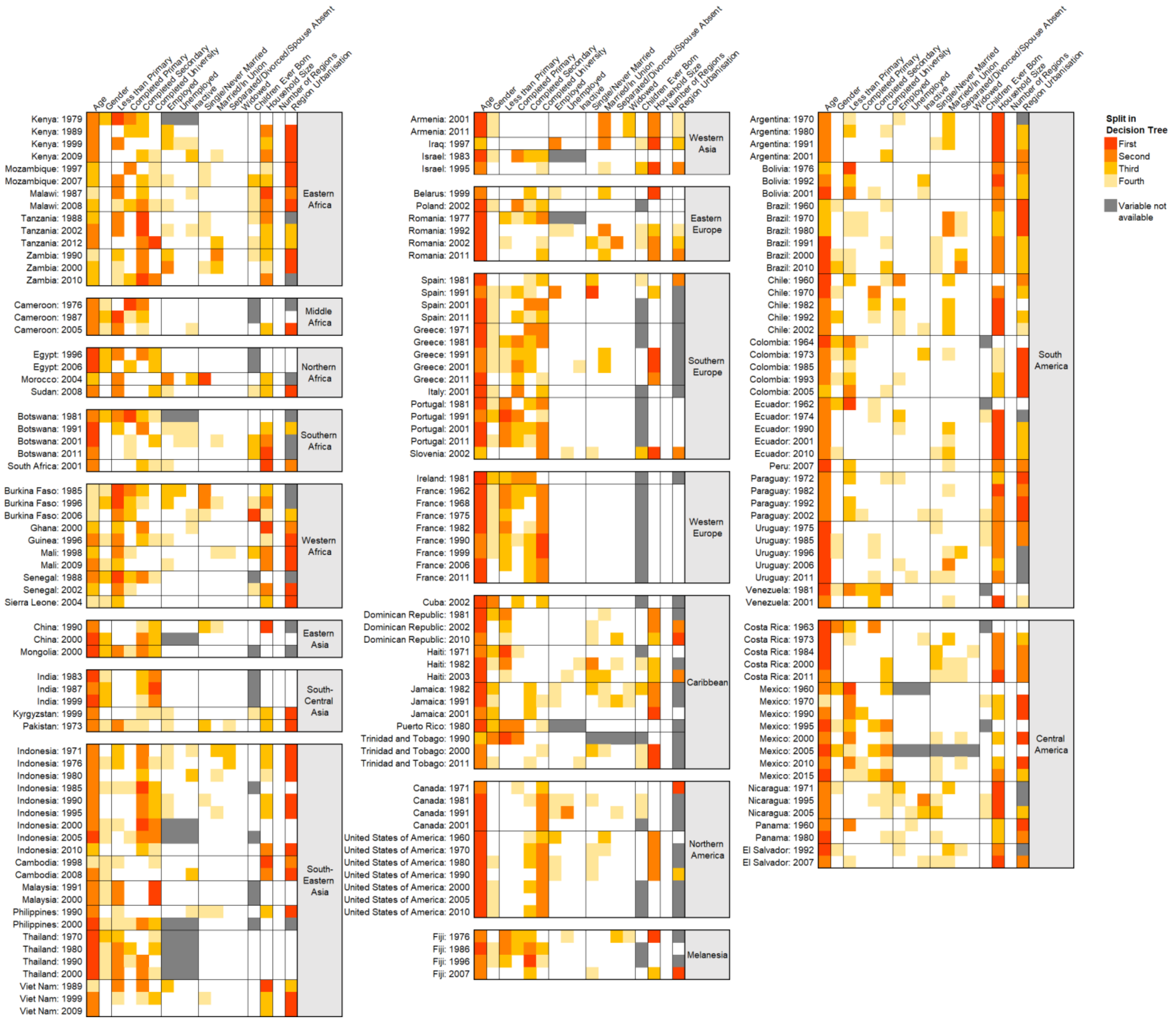

Figure 2: Heat map representing the order in which the variables determining migration are being split in decision tree classification

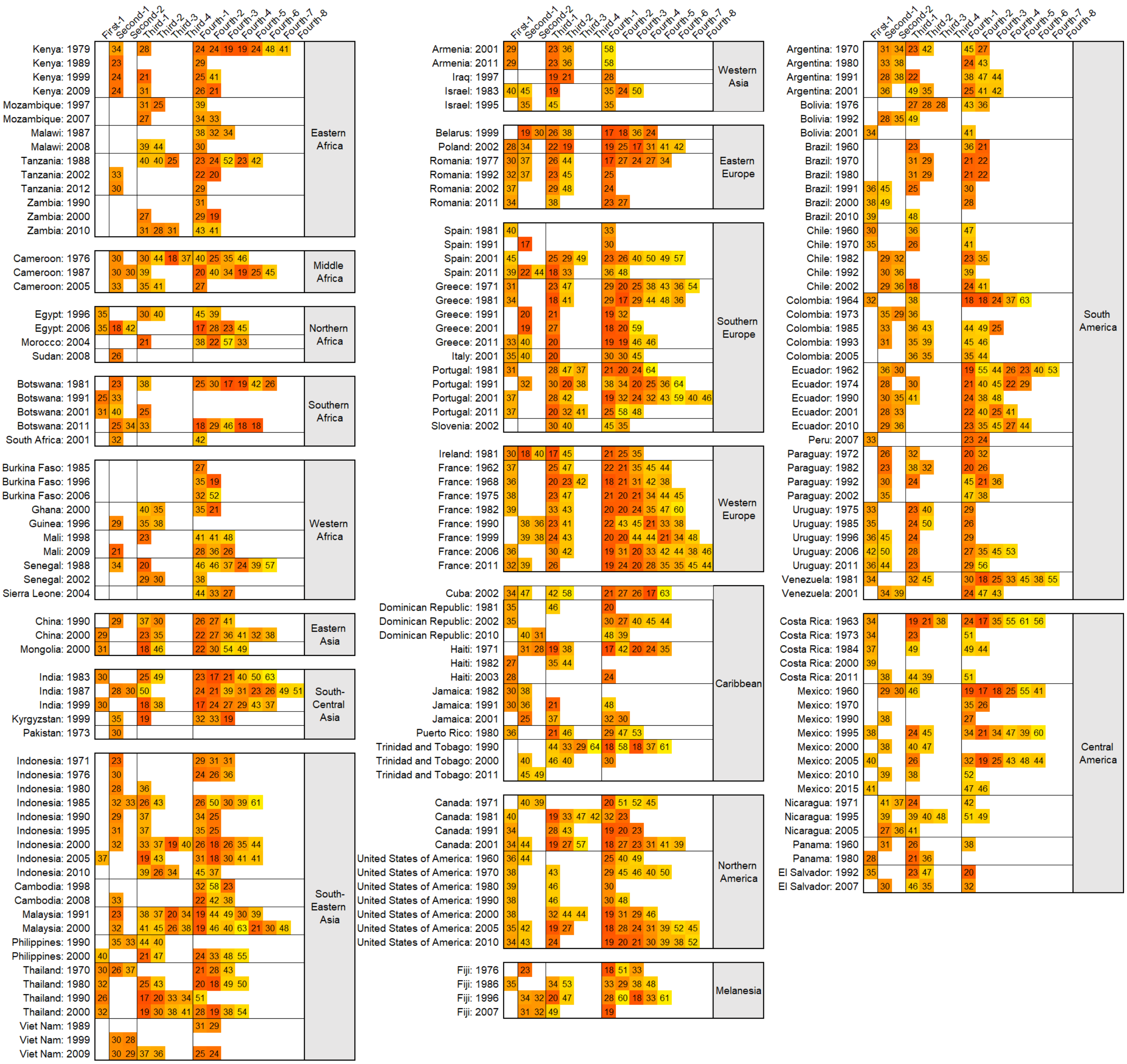

Figure 3: Heat map representing the order in which age is being split in decision tree classification