INTEGRATING SEQUENCE ANALYSIS INTO EVENT HISTORY ANALYSIS: AN APPLICATION TO MIGRATION HISTORIES

Philippe Bocquier, Ashira Menashe-Oren, Carren Ginsburg

1 INTRODUCTION

Event history analysis is rooted in survival analysis. Its objective is to analyse the occurrence of an event for individuals that are observed for different lengths of time until censoring time t_c. A fundamental condition to validate the analysis is that censoring is non-informative, independent from the event of interest. Typically, censoring is the last time of observation, such as the day respondents were interviewed in a survey, which is independent from the events experienced by respondents. In some cases, for example in a cohort follow-up, respondents cease to be observed for various reasons (refusal, migration, death…) that are not independent from the event of interest. This is usually called attrition, or informative censoring. This is noticeably hard to handle and is the source of important biases.

Informative censoring is the reason why it is a bad idea to examine retrospectively residential histories for the entire recorded life of respondents, i.e. at the time of the event of interest, t_e, or at time of censoring, t_c, if the event was not experienced. The respondents who did not experience the event of interest until censoring time t_c will necessarily have, on average, longer and more complex residential histories than those who experienced the event at time $t_e < t_c$.

Therefore, the preferred strategy is to capture residential histories through the residential status at each analysis time t_a in the interval running from the first time the respondent was observed, t_0, to the last observation time, t_e or t_c. The variable capturing this residential status is called time-varying because its value can change at any analysis time t_a. In other words, respondents' residential status effect is compared without bias at the same time t_a, and not at last observation time, t_e or t_c. In contrast, the current migration or residential status only takes the last stage of sometimes complex migration itineraries. Hence accounting for long migration histories is meaningful.

The description of complex migration histories has often been achieved through sequence analysis. Sequential analysis has been used mainly to characterise migration histories of respondents surveyed at a fixed time t_c. The main advantage of sequential analysis is to account for the order, as well as the duration, of each sequence. Optimal matching is the preferred method to form clusters but cannot be applied on large samples. A random, representative sub-sample is generally drawn to perform sequence analysis.

How can event history analysis, with its rigorous handling of censoring, work with sequence analysis that reflects complex itineraries normally at a fixed period of time? The objective of the paper is to find a workable solution to reconcile the two. We will also compare results using sequence analysis and other characterisation of migration itineraries to evaluate the effectiveness of these different procedures. As a side product, the paper also proposes a methodology to extend sequence analysis to the whole population, thus lifting the constraint of sub-sampling.

2 MIGRATION HISTORIES: FROM MIGRATION STATUS TO SEQUENCES

As migration is a repeatable event, the resulting residential histories and migration statuses can become very complex over a lifetime. In response to this complexity, researchers use different strategies. We illustrate these with some examples taken from a non-exhaustive sample of the migration literature addressing health issues.

The last residence or the status acquired at last migration is the most obvious step in statistical analysis. Respondents' residential status effect is compared without bias at time t_a. This is capturing effect of order 1, or immediate effect, meaning effect of the status in the time unit just before t_a. It is sufficient to measure the effect of migration status on another phenomenon occurring at time t_e. For example, by distinguishing between remaining-in-place, voluntary

migration, and forced displacement, as well as between periods spent "on the move" versus periods spent in residence, it is possible to analyse both the probability of the first pregnancy and the subsequent spacing of higher order pregnancies (Verwimp et al., 2020).

However, the 1st order effect is generally not considered sufficient to reflect the impact of residential histories. Second order (or lagged) effects are often useful to better characterise migration status, as they account not only for the current residence or migration status (at destination), but also for the previous place of residence (at origin) or residence some years before (often five years). Typically, simple origin-destination comparison identifies migration status. For instance, rural-to-urban migration or intra-urban residential mobility which often lead to improved living conditions (Patel et al., 2020). The place of residence at birth is sometimes preferred to the previous place of residence (often due to data availability) but the effect remains 2nd order. Again, the resulting respondents' residential status effect is measured without bias at time t_e or t_c.

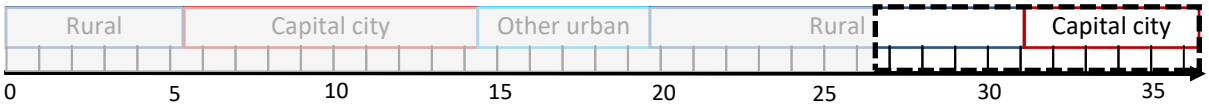
The duration since migration and the category of the place of origin (e.g. capital, other urban, rural, foreign), sometimes interacting with each other, can also add more details. For example, Ginsburg et al. (2016) used migration status and duration of exposure to identify differences in mortality risk between non-migrants (the reference category), new in-migrants, and return migrants, depending on exposure in the study area, as well as exposure outside the area in the case of return migrants. The assumption is then that the respondent retains in the current environment the memory of exposure to a previous environment, however defined by the researcher. In a meta-analysis on migration and cancer, cancer research was found to incorporate residential histories primarily focusing on incidence and estimating cumulative exposure (Namin et al., 2021). Another typical example is residence considered as exposure to a risk of infection disease, such as HIV. One study established that residence outside the rural study area is a strong predictor of HIV seroconversion in men, but not in women, and that residing in rural areas in a single or in multiple locations is a less significant risk factor for HIV acquisition compared to moving out of rural areas (Dobra et al., 2019). The long-term consequence of exposure is also sometimes of interest. For instance, urban residence during childhood and later in life may present cumulative risks for adult obesity (Kuuire et al., 2019).

With access to more detailed data on residential histories and the statistical tools to analyse them, researchers have begun to combine current residence with last residence, penultimate residence, etc., back to residence *n* years ago (or at birth) to form complete residential (or migration) histories over a fixed period of time, or over a lifetime. In other words, the residential histories are supposed to reflect *x*-order effects, or residences *t-n* years ago. The assumption is that, since current status is the product of a history, all individuals retain memories of the sequence of past residences and adapt their future behaviour accordingly. The number of combinations increases considerably with the number of past residences and the categories that compose them.

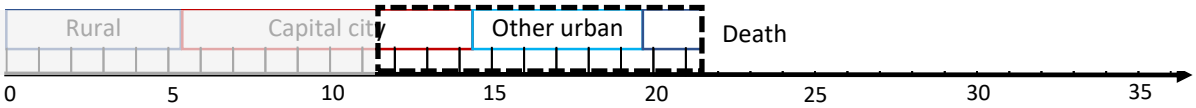
Whether of order one, two or higher, the resulting migration status variables are well suited to the analysis of a current situation, at a specific time *t*, or to the analysis of a situation over a fixed period (*t-n, t*). In either case, the analysis is cross-sectional even if the aim is to describe migration with a longitudinal perspective, in the past few years before *t*. Indeed, individuals must be observed over the same period of time in order to make meaningful comparisons. Generally, the sequences of residential histories will be standardised to cover the same period of *n* years before the end of observation or any other particular event of interest. For example, Chihaya et al. (2022) applied "*sequence analysis to decade-long residence and workplace histories of newly arrived migrants in Sweden to identify a typology of combined residence-work trajectories*", and Gosselin et al. (2018) applied sequence analysis to build a typology of migrants' residential pathways before and after HIV diagnosis.

Residential histories going back to birth lead to heterogeneity in duration, unless respondents are observed at the same age (cohort analysis). Therefore, a fixed duration of *n* years is preferable when comparing cohorts of different ages (generations). The normalisation of duration is related to the issue of censoring and selection bias. A fixed duration is a way to control (or reduce) the heterogeneity of the analysis time at the time of censoring t_c. If duration is not controlled, individuals with longer, and therefore on average more complex, itineraries will be over-represented, leading to selection bias, possibly survival bias (only those who survived the censoring time t_c are represented). However, imposing a fixed duration shifts the problem of temporal heterogeneity to time t_c. This temporal heterogeneity has to be controlled for, for example by age adjustment.

10-year life history of an individual from birth to time of survey at age 36 (sister 1):



10-year life history of an individual from birth to time of death at age 21, with the same life history as sister1 (sister 2):



10-year life history of an individual from birth to time of death at age 3, with the same life history as sister1 (sister 3):

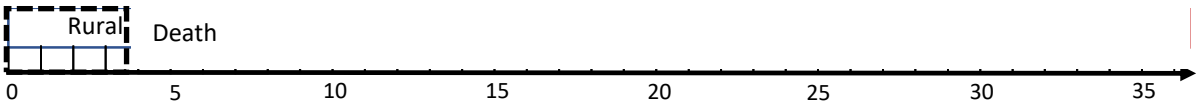


FIGURE 1: EXAMPLE OF A LIFE HISTORY WITH THREE 10-YEAR WINDOWS AT DIFFERENT TIME OF CENSORING (36-YEAR OLD, 21-YEAR OLD, AND 3-YEAR OLD)

To illustrate this, Figure 1 shows the example of triplet sisters who co-resided from birth. The first sister having survived until the survey time t_c (censoring), the second having died at 21, the third at 3. A fixed duration of 10 years would lead to very different migration histories (sister 1: rural-capital; sister 2: capital-other urban-rural; sister 3: rural), with different 1st order effect (capital, rural, rural) and 2nd order effect (rural, other urban, none). This is because migration histories are not identified at the same age for the three sisters, although they lived together in the same places from birth. Moreover, the migration history of the third sister will be discarded because it does not satisfy the minimum 10-year duration, thus leading to a survivor's bias in our analysis of the migration effect on mortality.

The above-mentioned duration standardisation strategy is suitable when it comes to a descriptive or exploratory analysis of residential histories or to measure the effect of migration status on another phenomenon observed at time t_c. However, this strategy is inappropriate in an event history analysis (EHA) framework where the phenomenon (event) can occur at any analysis time t_a, for example from birth or age 15 until the time of the survey t_c. In such a framework, the variable qualifying the residential status must be strictly comparable at each analysis time t_a along an observation period (t_0, t_c) that varies from one individual to another. In other words, neither the event nor the censoring occurs at the same time for all individuals.

Sequence analysis has become a very popular statistical tool for classifying complex residential histories and reducing them to a manageable number of categories. Studies of residential histories through sequential analysis do not adequately address the issue of censoring, while EHA often limit residential histories to 1st or 2nd order effects. The main objective of this paper is to try and reconcile the migration history approach (descriptive, inductive) with EHA (causal, hypothetico-deductive). This involves finding a method to include complex, high order residential histories as a time-varying covariate in EHA, whatever the size of the population at stake. Our secondary objectives are to provide a proof of concept that can be applied to any type of history (residential or not, e.g. fertility, health, employment, etc.), to provide codes adaptable to specific analyses, and to give an illustration of this approach with real data, in this case, by linking migration and death histories using data from health and demographic surveillance (HDSS) sites.

Following the above discussion, we identify four criteria to evaluate appropriateness of a method to identify migration effect on a time-dependent event. The method should:

1. Account for censoring at time t_c
2. Account for migration status independently from event or status of interest at time t_e
3. Account for the time-varying nature of migration status, from t_0 to the last observation time, t_e or t_c
4. Account for lagged migration effects to the highest order possible.

How does sequence analysis conform to these criteria? To account for censoring (not everyone is interviewed at the same age, or duration of exposure), it is not advised to account for the migration history over the life-time. Rather,

migration histories will be collated over a fixed period of time, say, the last 10 or 20 years before the last observation time, t_e or t_c. The past exposure, prior to these 10 or 20 years, is controlled through another variable (e.g. age group).

In this typical sequential analysis, the above criteria 1 and 4 are met. Exposure at censoring is neutralised by, say, controlling for age group, and maybe, place of birth (or at 15-year old), which represents the highest order migration effect prior the 10- or 20-year fixed period. However, the migration history is no longer independent from event or status of interest (criterion 2) since the longer the exposure (e.g. the older the respondent), the higher the selection. Respondents are selected by virtue of sheer survival to the time t_e at which the event or status of interested is recorded. It is therefore difficult to say whether the migration history reflects the conditions of the last n years before the survey or the respondent's exposure (e.g. age). In addition, sequential analysis applied to a fixed time, t_e or t_c, by design, does not account for time-varying migration status (criterion 3): migration history is fixed in time for each respondent.

3 PROPOSED METHOD: TIME-VARYING MIGRATION HISTORIES

Say that we observe individuals from birth to their death and that we believe that, as they age, these individuals' memory encompasses mainly the last 10 years. Each day of their life, individuals would remember only the last 10 years of their migration history or, more realistically, only the last 10 years of their migration history would matter to explain their current behaviour. The Figure 2 represents three different 10-year time observation windows along the life of an individual.

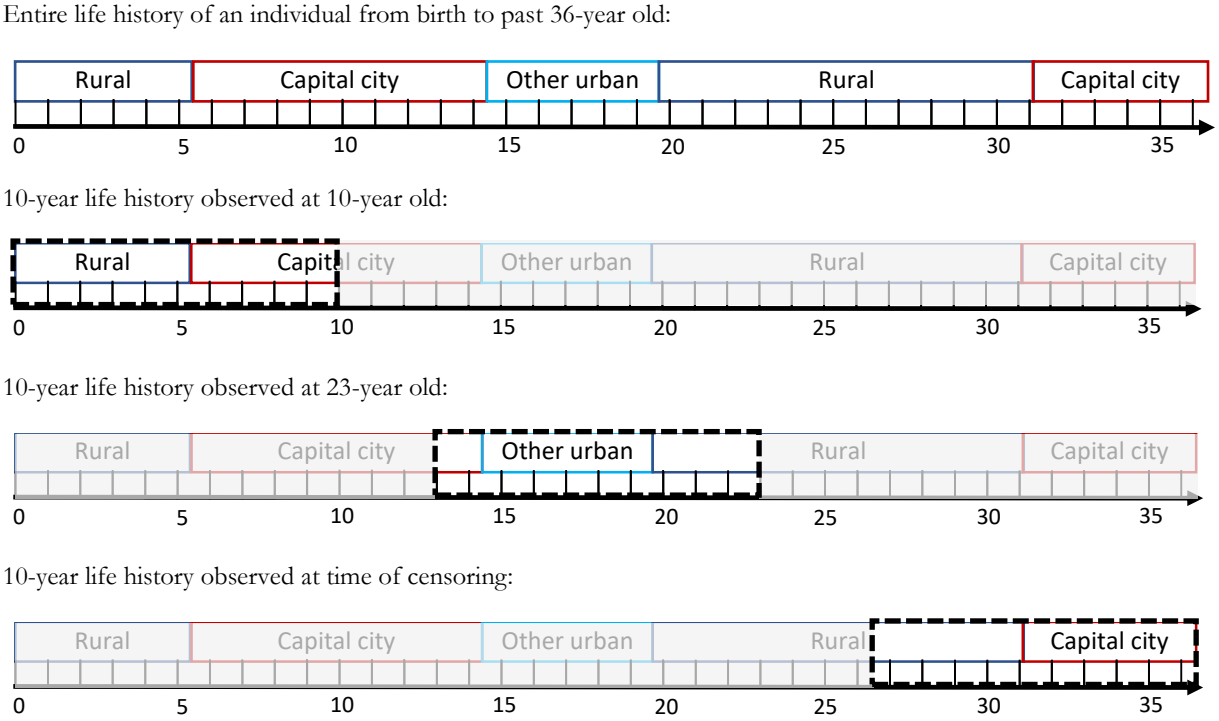


FIGURE 2: EXAMPLE OF A LIFE HISTORY AND THREE 10-YEAR WINDOWS AT DIFFERENT TIME (10-YEAR OLD, 23-YEAR OLD, AND AT TIME OF CENSORING)

The 10-year migration history starts to be observed from 10-year old, but between that age and the last observation date (censoring), there is an infinite number of 10-year migration histories that can be identified. In practice, the number of migration histories should be limited by the accuracy of the data. Typically, the threshold in the field is generally 3- or 4-months duration of residence. Only migration events that separate residence period of more than 3 or 4 months would be recorded. Considering this level of accuracy, a higher threshold of 6-month should be used for analysis. Moreover, if the final aim is to construct sequences out of series of places of residence, it might be good for computation purpose to limit the series to 10.

For the simplicity of the demonstration, we will therefore assume that we are interested in migration histories depicted by sequences of 10 places of residence: 1, 2, 3… 9, and 10 years ago. Of course, depending on data accuracy and focus of the study, the period could be extended beyond or reduced below 10 years, and the categories of sequences as well. But, whatever the choice of time period and categories of sequences, how will migration histories be included as a time-varying independent covariate in the analysis?

Let's start with the example of the 10-year life history observed at 23-year old in Figure 1. The sequence can be translated as a series of codes (K: capital city; U: other urban; R: rural) representing the different types of residence 1, 2, 3… 9, and 10 years ago (Figure 3). These codes can be indexed by the number of years before the observation time, e.g. U_7 for "other urban 7 years ago". Suppose that the migration from capital city to other urban occurred when the respondent was 14.4-year old and from other urban to rural at 19.7-year old. The 10-year sequence would then change at the anniversary of these migrations, i.e. when K_9 turns into U_9 at 23.4-year old, and also when U_4 turns into R_4 at 23.7-year old. In other words, the sequence is time-varying according to any anniversary occurring in the past 10 years.

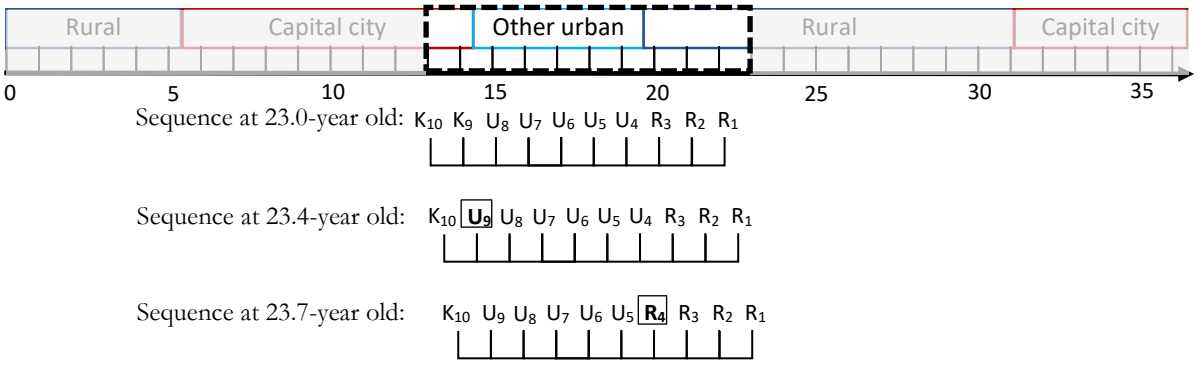


FIGURE 3: EXAMPLE OF CODING A 10-YEAR LIFE HISTORY AT DIFFERENT AGES (23-YEAR OLD, 23.4-YEAR OLD, AND 23.7-YEAR OLD)

4 IMPLEMENTATION OF TIME-VARYING MIGRATION HISTORIES

We test the value of time-varying migration histories by examining the risk of adult death according to migration history. We contrast and compare the use of 1) time-varying migration histories clustered using sequence analysis, 2) time-varying migration histories grouped manually, and 3) second-order classification. We detail the different steps taken for each of these methods below.

4.1 DATA

To test the solution of time-varying migration histories, we use longitudinal data from a Health and Demographic Surveillance System (HDSS) in South Africa, Agincourt. Agincourt is located in the rural north-east of South Africa, and the population in this area has been under surveillance since 1992. The HDSS data records events (births, deaths and migrations) with corresponding dates, for all individuals in the area. The data for Agincourt HDSS was obtained from INDEPTH i-share, and covers around 152,500 individuals. Life expectancy in the HDSS was around 73 for women and 68 for men in 1994, and this since declined significantly in the 2000s due to high HIV prevalence in the region, but recovered with rollout of antiretroviral treatment. Fertility in the HDSS has declined from 3.7 in 1994 to 2.3 in 2009 (Kahn et al., 2012). The HDSS is also characterised by a relatively high proportion of temporary migrants (ranging from around 15% of women aged 15-34 to 60% of men aged 35-54 in 2000) (Collinson et al., 2006).

Several variables are necessary for our analysis. We will use the same names and codes as in manuals initiated for HDSS, summarised in Table 1. These can be applied to any longitudinal setting. Each individual in the HDSS is assigned a unique identifier (IndividualID). An individual is considered a resident in the HDSS following enumeration, or birth, and after six months present on the site following an in-migration. Note, at the time of these events (enumeration, birth and in-migration), the individual is considered a non-resident, while at the time of death and out-migration or end of observation, she is considered a resident. The HDSS continuously observes individuals who enter observation by enumeration (at baseline census), by birth or by in-migration. They may exit observation by death or

out-migration. Since we consider whether death depends on the migration history of the last 10 years, we only use the last five years of observation in the available i-share data (1 January 2012- 1 January 2017), to allow for sufficient "history" for each individual. We focus on adults of both sexes aged 25-84: the lower boundary of this age range ensures we have a 10-year migration history since adolescence, and the upper boundary of the age range ensures we have enough deaths in our data.

TABLE 1: VARIABLES' NAME, LABEL AND VALUE LABELS

Variable name	Variable format	Variable label	Value labels
ID	string	Individual ID	None
EventDate	%tc	Date at the end of episode	None
EventCode	byte	Event at the end of episode	1 ENU "Enumeration" 2 BTH Birth" 3 IMG "In-migration" 4 OMG "Out-migration" 5 EXT "Exit HH" 6 ENT "Entry HH" 7 DTH "Death" 9 OBE "End of observation" 50 "migration history"
residence	byte	Residency status	0 "non-resident" 1 "resident"
datebeg_1	%tc	Earliest recorded date at the beginning of episode	None
date_OBE	%tc	Latest recorded date at the end of episode	None

4.2 METHODS

We use Stata MP 17 for all our analyses, and provide the programme in the appendix. However, nothing prevents readers from implementing the procedure in another software.

4.2.1 TIME-VARYING MIGRATION HISTORIES

We apply the migration history method simplifying things further compared to the example above by using only two codes for residence: 1 for resident, 0 for non-resident. The migration histories are therefore 10-digits codes made of series of 0 and 1. However, if one wants to account for different type of left-censoring, other codes will be needed to characterise periods before birth, immigration, or enumeration.

The procedure is detailed in the program in the appendix. The heart of the program is a loop (*n*: 1 to 10) to capture the status *n* years ago. Each loop creates from the original file in long format (each residency episodes ordered by time of occurrence) a new file corresponding to the situation *n* years ago (lag files). A new code, of "migration history", is created for EventCode indicating that only events before end of observation (OBE) should be lagged. Otherwise, the general case is that if the end of the episode falls after the last observation date, and the beginning of the episode falls before the last observation date, then, the indicator variable is coded "1": A new variable is then needed to capture the residency status of the lagged records. The dates of the lagged events (i.e. except OBE) are lagged by *n* years, and no records post-OBE should remain in the database and therefore are deleted. Only records with lagged events and OBE are kept in the database, and this new file is saved. This new file is then merged according to time with the file from the last loop. For that, the program "tmerge.ado" must first be installed.[1] We eventually create a new variable with the sequence of residency status over 10 years.

4.2.2 SEQUENCE ANALYSIS IN SUB-SAMPLES

Sequence analysis is based on the construction of a distance matrix between sequences. The matrix is used to group sequences in clusters. One computational issue with sequential analysis is that the distance matrix between individual sequences can be huge and intractable when exceeding a certain number of respondents. Sequential analysis is often

[1] `net install https://github.com/bocquier/mighealth/blob/master/tmerge.ado`
A version of "tmerge" is also available in R and SAS.

used on representative sub-samples, or conducted separately on different sub-category of the sample, to the expense of direct comparison of clusters across sub-categories. We propose a way around this constraint.

Figure 4 summarises the workflow from generating the time-varying histories to forming clusters for the full dataset. After generating lagged time-varying migration histories as explained in the previous section (step 1), to circumvent the constraint of drawing a single representative sample, we drew randomly 2% and 3% samples from the Agincourt data, giving us 40 sub-samples (step 2). In these sub-samples, we exclude the sequences of residence only (step 3): they will form a unique category of continuous residence over 10 years, coded "1111111111", akin to what we consider permanent residents. Sequence analysis is conducted on the remaining sequences from each sub-sample (step 4) and then the resulting clusters for the whole population are combined (step 5), as detailed below.

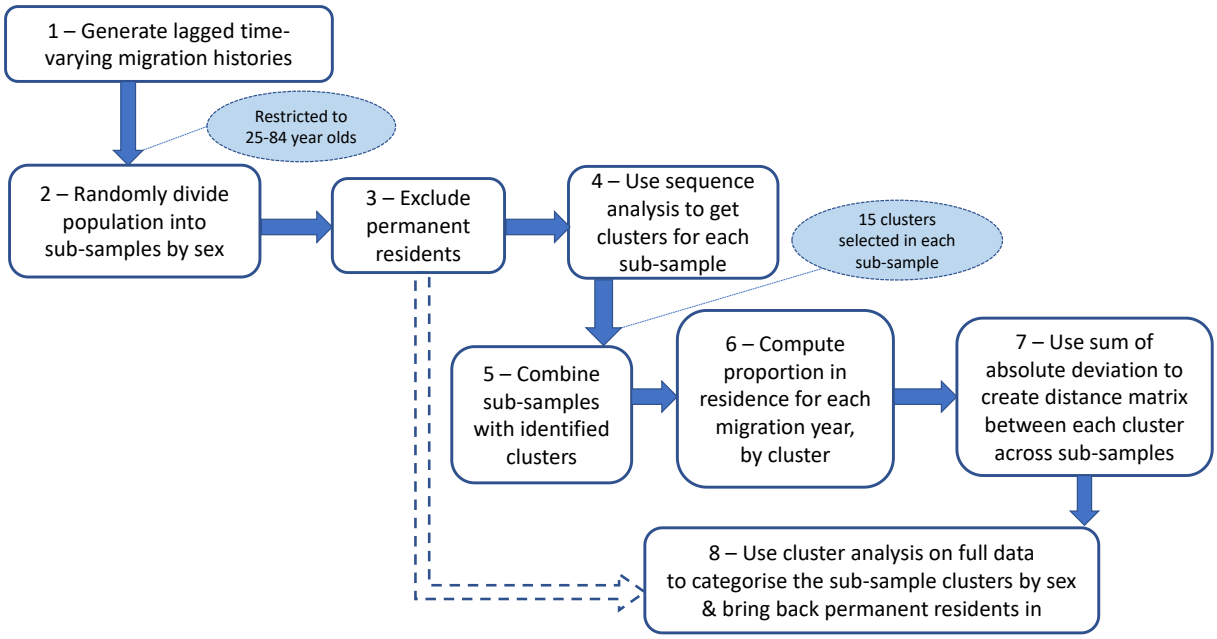


FIGURE 4: WORKFLOW DIAGRAM FOR MIGRATION SEQUENCE ANALYSIS

Using time-varying sequences, one can identify clusters of migration histories with sequence analysis tools (step 4). To run sequence analysis in Stata, it is necessary to install the package "sq". In order to group the migration histories, clustering is based on a distance matrix between all the different sequences representing migration histories. We use Wards linkage to identify 15 clusters for each sub-sample. This high number is the default value proposed by Stata (this can be changed to a higher value). This is far above the optimum number of clusters which one would obtain based on the highest Calinski/Harabasz pseudo-F value. However, the F-value is constrained by sample size. To make the selection of clusters less sample-dependent, we prefer to be less selective since we will eventually group the clusters in larger ones when merging all sub-samples together. It is indeed possible that a cluster is never big enough in any of the 40 sub-samples to be identified as a cluster using the F-value because of limited sub-sample size. This cluster would be systematically merged with another one. By retaining a maximum of 15 clusters per sub-sample, there is a higher chance of identifying small clusters represented across many sub-samples. In steps 5 to 7, we compare these clusters and group them when merging all sub-samples together.

4.2.3 COMBINING THE CLUSTERS FOR THE WHOLE POPULATION

We combine in one file the 15 clusters from sequence analysis for each of the 40 sub-samples (step 5). The clusters can be characterised by the proportion of time spent in the site (i.e. as resident) for each of the 10 years of the migration histories. Considering that there are only two codes, 0 or 1, for each year, a series of 10 proportions is sufficient (step 6). Using the proportions for each year, we can compare the series of 10 proportions in one cluster with the series of another, by computing the sum of absolute deviations over the 10 years (step 7). In the case of 3 or more codes (e.g. 1, 2, 3 for capital, other urban, rural), the series would be constituted of as many proportions as necessary to cover the alternative codes (e.g. with 3 codes we need 2 proportions for each year i.e. 20 proportions for 10 years; or 30 proportions for 4 codes, etc.).

From the previous stage, we obtain a symmetric matrix of the sum of absolute deviations from each cluster to all other clusters in all sub-samples. Considering that we selected 15 clusters in each 40 sub-samples, this distance matrix has a 600 X 600 dimension. We then form "clusters of clusters" using the distance matrix. We use Wards linkage to identify the clusters but here, instead of setting a maximum number of clusters, we use the highest Calinski/ Harabasz pseudo F-value out of a maximum of 15 clusters to identify the optimum number of clusters in the whole population (step 8). The sequences of 10-year residence only (permanent residence) form an extra cluster that we add. From hereon we refer to this method as the "sequence analysis" method.

4.2.4 MANUAL CLUSTERING

The clusters identified with the above adapted sequence analysis will be compared with the categories generated "by hand". These manually identified categories organise the migration history sequences into conceptually meaningful sequences, constrained by frequency (see Table 2 for the distribution of categories). The most frequent sequence is "1111111111" – permanent residence over 10 years. The basic concepts that were used to form the categories are number of years spent in and numbers of years spent out. The potential sequences are ranked according to these sequences of in and out of site and according to frequency. Some complex combinations are grouped (under code 56 in Table 2). Also, unique sequences that are the least frequent are grouped into a separate category ("other") and represent 6.8% of the total person-years.

TABLE 2: MANUAL CODING OF MIGRATION HISTORIES – AGINCOURT HDSS

Sequence	Code	Grouped code (17)	Grouped code (8)	Label	Person-years
1111111111	0	0	0	Permanent residence	70978.10
0000000000	1	1	1	6-12 month in-migrant	3276.58
0000000001	2	2	1	1yr in-migrant	2965.77
0000000011	3	3	2	2yr in-migrant	3241.25
0000000111	4	4	2	3yr in-migrant	2876.11
0000001111	5	5	2	4yr in-migrant	3194.70
0000011111	6	6	3	5yr in-migrant	3562.59
0000111111	7	7	3	6yr in-migrant	3059.22
0001111111	8	8	3	7yr in-migrant	3442.39
0011111111	9	9	3	8yr in-migrant	3574.61
0111111111	10	10	3	9yr in-migrant	4004.59
1111111110	11	11	4	6-12 month return, away for 1yr	502.92
1111111100	12	11	4	6-12 month return, away for 2yr	435.19
1111111000	13	11	4	6-12 month return, away for 3yr	317.33
1111110000	14	11	4	6-12 month return, away for 4yr	262.95
1111100000	15	11	4	6-12 month return, away for 5yr	249.45
1111000000	16	11	4	6-12 month return, away for 6yr	210.54
1110000000	17	11	4	6-12 month return, away for 7yr	207.38
1100000000	18	11	4	6-12 month return, away for 8yr	176.86
1000000000	19	11	4	6-12 month return, away for 9yr	200.26
1111111101	20	12	4	1yr return, away for 1yr	566.60
1111111001	21	12	4	1yr return, away for 2yr	394.25
1111110001	22	12	4	1yr return, away for 3yr	283.33
1111100001	23	12	4	1yr return, away for 4yr	227.71
1111000001	24	12	4	1yr return, away for 5yr	208.71
1110000001	25	12	4	1yr return, away for 6yr	183.89
1100000001	26	12	4	1yr return, away for 7yr	178.98
1000000001	27	12	4	1yr return, away for 8yr	171.06
1111111011	28	13	4	2yr return, away for 1yr	663.47
1111110011	29	13	4	2yr return, away for 2yr	468.30
1111100011	30	13	4	2yr return, away for 3yr	345.97
1111000011	31	13	4	2yr return, away for 4yr	296.16
1110000011	32	13	4	2yr return, away for 5yr	256.46
1100000011	33	13	4	2yr return, away for 6yr	230.19
1000000011	34	13	4	2yr return, away for 7yr	226.54
1111110111	35	14	5	3yr return, away for 1yr	608.64
1111100111	36	14	5	3yr return, away for 2yr	431.05
1111000111	37	14	5	3yr return, away for 3yr	339.25
1110000111	38	14	5	3yr return, away for 4yr	309.04
1100000111	39	14	5	3yr return, away for 5yr	253.05

1000000111	40	14	5	3yr return, away for 6yr	218.00
1111101111	41	15	5	4yr return, away for 1yr	690.20
1111001111	42	15	5	4yr return, away for 2yr	528.40
1110001111	43	15	5	4yr return, away for 3yr	416.59
1100001111	44	15	5	4yr return, away for 4yr	345.78
1000001111	45	15	5	4yr return, away for 5yr	299.57
1111011111	46	16	6	5yr return, away for 1yr	804.61
1110011111	47	16	6	5yr return, away for 2yr	625.87
1100011111	48	16	6	5yr return, away for 3yr	471.65
1000011111	49	16	6	5yr return, away for 4yr	432.51
1110111111	50	16	6	6yr return, away for 1yr	696.93
1100111111	51	16	6	6yr return, away for 2yr	587.34
1000111111	52	16	6	6yr return, away for 3yr	465.74
1101111111	53	16	6	7yr return, away for 1yr	811.17
1001111111	54	16	6	7yr return, away for 2yr	714.12
1011111111	55	16	6	8yr return, away for 1yr	865.47
0000000010	56	17	7	6-12 month return, away for 1yr, prior in-migrant	198.91
0000000110	56	17	7	6-12 month return, away for 1yr, prior in-migrant	
0000001110	56	17	7	6-12 month return, away for 1yr, prior in-migrant	
0000011110	56	17	7	6-12 month return, away for 1yr, prior in-migrant	
0000000100	57	17	7	6-12 month return, away for 2yr, prior in-migrant	59.91
0000001011	58	17	7	2yr return, away for 1yr, prior in-migrant	86.85
*	59	17	7	other	8900.52
Total					131602.00

Considering the large number of potential categories with many having few person-years, we collapse these into larger categories (with at least 1000 person-years each), reaching 17 categories (excluding permanent residents). We also combine the categories further to be in parallel with the categories identified with the second-order migration status. From hereon we refer to this method using migration histories as the "manual", and the "manual reduced" for the collapsed categories.

4.2.5 CLASSIC SECOND-ORDER MIGRATION CATEGORIES

Following previous work (Ginsburg et al. 2015), we classify period of exposure in the last 10 years between non-migrants (the reference category), new in-migrants, and return migrants. The in-migrants and return migrants are sub-categorised in 3 categories depending on the number of years of exposure (<2, 2-4, 5+) in the HDSS, thus forming 6 categories in addition to the non-migrants. In the case of return migrants, an extra covariate captures the number of years of exposure (<3, 3+) outside the HDSS, so the 3 categories of return migrants can be sub-divided by two. In total, we form 9 categories of migrants in addition to non-migrants. From hereon we refer to this method as the "classic" method.

5 RESULTS: MIGRATION AND MORTALITY

We applied the different methods to identify migration clusters (sequence analysis, manual, classic) in longitudinal data from Agincourt HDSS, South Africa, for adults aged 25-84. We then use Cox models to examine the risk of death according to each set of migrant categories, controlling for sex and period.

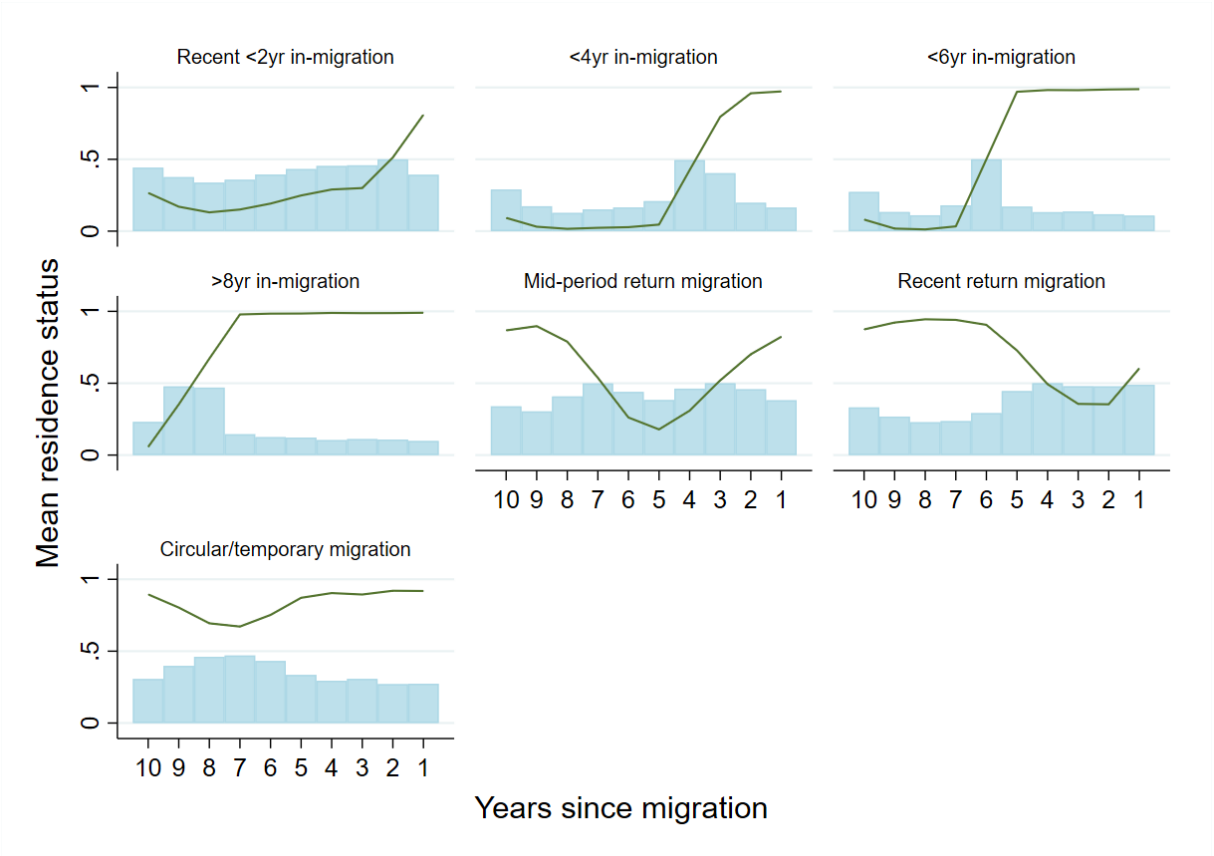
5.1 COMPARISON OF THE THREE METHODS IN CHARACTERISING MIGRATION

5.1.1 SEQUENCE ANALYSIS

We clustered the ten-year migration histories using Wards hierarchical clustering for each of the 40 sub-samples, and then combined these sub-samples, obtaining "clusters of clusters", through sequence analysis of the migration histories (step 8 of Figure 4). The optimal number of clusters according to the pseudo-F value was seven (excluding permanent residents). We characterised the clusters relying on the proportion of time spent in the HDSS using the mean residence status at each year over the last 10 years (Figure 4). To these clusters, we add the cluster of permanent residents.

For example, the cluster ">8yr in-migration" in Figure 4 illustrates how for the last seven years the majority of individuals in the cluster were classified as resident (with very small standard deviations for these years). At the time of moving, in years eight to ten, the standard deviation is greater, and in year ten the mean residence status is close to

zero. This cluster can be characterised as capturing in-migration a relatively long time ago. The "recent return migration" cluster is another example: we see that up to around five years ago individuals were resident, then tended to spend time out of the HDSS, and then in the last year there is an upturn in the mean residence status, indicating return not so long ago. Since the timing of both leaving and returning is not exactly the same for all individuals in the cluster, the standard deviation is higher. This cluster also only consists of 9.5% of observations. The last cluster, where the mean residence status lies closer to one over the entire period with a return-like pattern, we classify as circular or temporary migration. This cluster is quite large, 20.5% of observations, and is typical of known circular migration patterns in Agincourt (Collinson et al., 2006).

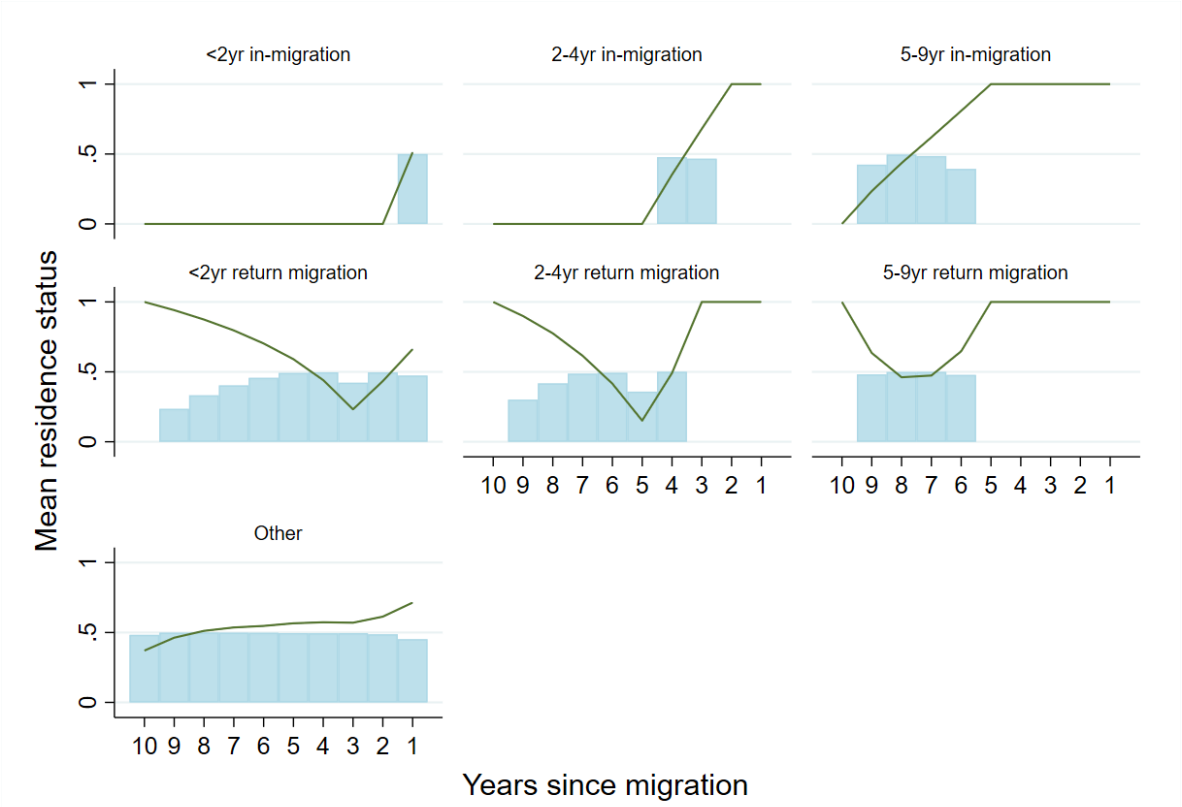


Notes: Graphs by cluster based on sequence analysis; Bars indicate standard deviation; Excludes permanent resident cluster

FIGURE 5: CLUSTERS OF 10-YEAR MIGRATION HISTORIES BASED ON SEQUENCE ANALYSIS

5.1.2 MANUAL METHOD

In Figure 5 we present a similar perspective of looking at the manually grouped migration histories as in Figure 4, using the mean residence status at each year over the last 10 years. Unlike the sequence analysis method, this method is more clear-cut in categorization. The first group clearly represents in-migration less than two years (all individuals were non-resident beforehand, and therefore the standard deviation around these years is zero). Similarly, in the "5-9yr return migration" group all individuals are resident over the last five years, but returned at different times before then, and so the standard deviation around these earlier years is high. The category we classified as "other" with small numbers of various sequences seems to capture circular or temporary migration patterns, similar to the cluster identified through the sequence analysis method.



Note: Graphs by manually defined groups (reduced); Bars indicate standard deviation; Excluding permanent residents

FIGURE 6: MIGRATION HISTORIES ACCORDING TO MANUALLY DEFINED CLUSTERS

Note: Graphs by 2nd order migrant status; Bars indicate standard deviation; Excludes permanent resident cluster

FIGURE 6: MIGRATION STATUS USING SECOND ORDER "CLASSIC" METHOD

5.1.3 CLASSIC METHOD

The classic migration categories (second-order status) are depicted in Figure 6, parallel to Figures 4 and 5. What stands out the most from this method is the lack of a category of temporary or circular migration which is captured in the sequence analysis and manual methods. The groups in Figure 6 are also more clear-cut than the clusters obtained with the sequence analysis, though they overall appear comparable. For instance, there are recent in-migration and recent return migration patterns.

5.1.4 COMPARING THE THREE MIGRATION "STATUS"

Figures 4 to 6 suggest considerable overlap in the migration groups according to the three methods. We examine this further by examining the proportion of person-years common in the groups across the methods (Tables 3 and 4). Apart from a small number of person-years considered as "<4yr in-migration", the "recent <2yr in-migration" cluster is very consistent with the classic "<2 year in-migration" group, with 98% of person-years in common (Table 3). Since the cut-offs of the categories on in-migration are not exactly consistent, we find that some classic categories are represented in two clusters. For instance, migrants in the classic "5yr+ in-migrant" category fall both in the "<6yr in-migration" and ">8yr in-migration" categories – naturally. These two clusters capture 98% of all distant in-migration. In-migration in general is more commonly classified by all methods, while return migration appears to differ more across methods. What we consider as circular and temporary migration in the sequence analysis method appears to be classified frequently in the return migration categories of the second-order status (Table 3).

TABLE 3: COMPARISON OF CLASSIC AND SEQUENCE ANALYSIS METHODS

		Clusters based on sequence analysis								
		Permanent resident	**Recent <2yrs in-migrant**	**4yrs in-migrant**	**6yrs in-migrant**	**>8yrs in-migrant**	**Mid-period return migrant**	**Recent return migrant**	**Circular migrants**	**Total**
Classic 2nd order migration status	**Permanent resident**	69,276.67	2.73	0.26	2.00	0.00	0.00	0.00	2.00	69,283.67
		99.99	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*100*
	<2y in-migrant	0.00	2639.99	66.07	0.00	0.00	0.00	0.00	0.00	2,706.06
		0.00	*97.56*	*2.44*	*0.00*	*0.00*	*0.00*	*0.00*	*0.00*	*100*
	<5y in-migrant	0.00	1625.66	5892.48	250.66	3.09	1.91	0.00	1.01	7,774.81
		0.00	*20.91*	*75.79*	*3.22*	*0.04*	*0.02*	*0.00*	*0.01*	*100*
	5y+ in-migrant	223.31	0.00	0.00	4834.50	6674.51	0.00	1.00	4.98	11,738.30
		1.90	*0.00*	*0.00*	*41.19*	*56.86*	*0.00*	*0.01*	*0.04*	*100*
	<2y return migrant	209.10	2832.99	466.30	176.83	211.66	2046.75	3285.30	1485.25	10,714.19
		1.95	*26.44*	*4.35*	*1.65*	*1.98*	*19.10*	*30.66*	*13.86*	*100*
	<5y return migrant	393.84	1698.89	2246.12	499.01	400.62	3330.53	1182.70	2868.76	12,620.48
		3.12	*13.46*	*17.80*	*3.95*	*3.17*	*26.39*	*9.37*	*22.73*	*100*
	5y+ return migrant	875.22	366.95	2.88	1932.04	4862.30	225.16	67.79	5154.38	13,486.73
		6.49	*2.72*	*0.02*	*14.33*	*36.05*	*1.67*	*0.50*	*38.22*	*100*
	Total	70978.15	9167.22	8674.11	7695.05	12152.18	5604.35	4536.80	9516.39	128,324.24
		55.31	*7.14*	*6.76*	*6.00*	*9.47*	*4.37*	*3.54*	*7.42*	*100*

The manual categories compare remarkably well with the classic categories of migration status when it comes to in-migration with over 97% of person-years in common across in-migration categories (Table 4). However, return migration is not as systematically distributed. 43% of the person-years considered as "<2yr return migration" in the classic method are classified as "other" in the manual categories. There also appears to be some inconsistencies not to be disregarded in what is considered more than five-year return migration in the class method. Notably, 41.6% of person-years are considered distant in-migrants rather than return migrants when relying on the manual grouping. This is likely because the migration histories only consider the last ten years of the individual, while the classic method considers a longer time period.

TABLE 4: COMPARISON OF CLASSIC AND MANUAL REDUCED METHODS

Classic 2nd order migration status	Manual clusters (reduced to 7 categories): Permanent resident	<2yrs in-migrant	2-4yrs in-migrant	5-9yrs in-migrant	<2yrs return migrant	2-4yrs return migrant	5-9yrs return migrant	Other	Total
Permanent resident	69,276.67	0	0	1.998683	2.00	0	0	2.99	69,283.67
	99.99	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*100*
<2y in-migrant	0	5,315.08	148.97	0	0	0	0	16.05	5,480.10
	0	*96.99*	*2.72*	*0*	*0*	*0*	*0*	*0.29*	*100*
<5y in-migrant	0	0	7,553.74	193.41	0.91	3.40	0.01	23.33	7,774.81
	0	*0*	*97.16*	*2.49*	*0.01*	*0.04*	*0*	*0.3*	*100*
5y+ in-migrant	223.31	0	0	11,504.54	0	0	4.98	5.46	11,738.30
	1.9	*0*	*0*	*98.01*	*0*	*0*	*0.04*	*0.05*	*100*
<2y return migrant	209.10	927.28	101.43	86.76	4,963.84	64.30	53.43	4,810.59	11,216.73
	1.86	*8.27*	*0.9*	*0.77*	*44.25*	*0.57*	*0.48*	*42.89*	*100*
<5y return migrant	393.84	0	1,507.91	250.11	2,297.71	4,371.84	278.25	3,520.81	12,620.48
	3.12	*0*	*11.95*	*1.98*	*18.21*	*34.64*	*2.2*	*27.9*	*100*
5y+ return migrant	875.22	0	0	5,606.58	0	0	6,138.72	866.21	13,486.73
	6.49	*0*	*0*	*41.57*	*0*	*0*	*45.52*	*6.42*	*100*
Total	70,978.15	6,242.36	9,312.06	17,643.40	7,264.47	4,439.55	6,475.40	9,245.45	131,600.80
	53.93	*4.74*	*7.08*	*13.41*	*5.52*	*3.37*	*4.92*	*7.03*	*100*

5.2 COX MODEL RESULTS

When the clusters or categories from each of the three methods are then included in Cox models (with sex as a covariate) to estimate the risk of death among 25-84 year olds in Agincourt, we note that there is some relationship between migration status and adult death, consistent across models, and some that differ (Figure 7). Regardless of the method used to capture migration, there seem to be similar results in terms of direction and strength of association. Return migrants tend to face an increased risk of death (between 19-44% higher depending on the model, with confidence intervals mostly above one). In all models except for the one using the classic method, recent in-migrants face a higher risk of death. In-migrants over five years ago have a lower risk of death in general – seemingly having time to adapt. However, in the manual method with multiple (18) categories, the hazard ratios for in-migration are not consistent in direction of effect.

The sequence analysis method and the manual method are advantageous in classifying a unique category of what we characterise as temporary or circular migration (named "other" in the manual method). The manual method is advantageous in having multiple categories, potentially allowing to see whether categories (or clusters) are masking some effects. However, the confidence intervals for these categories are wide and the direction of effects is not clear.

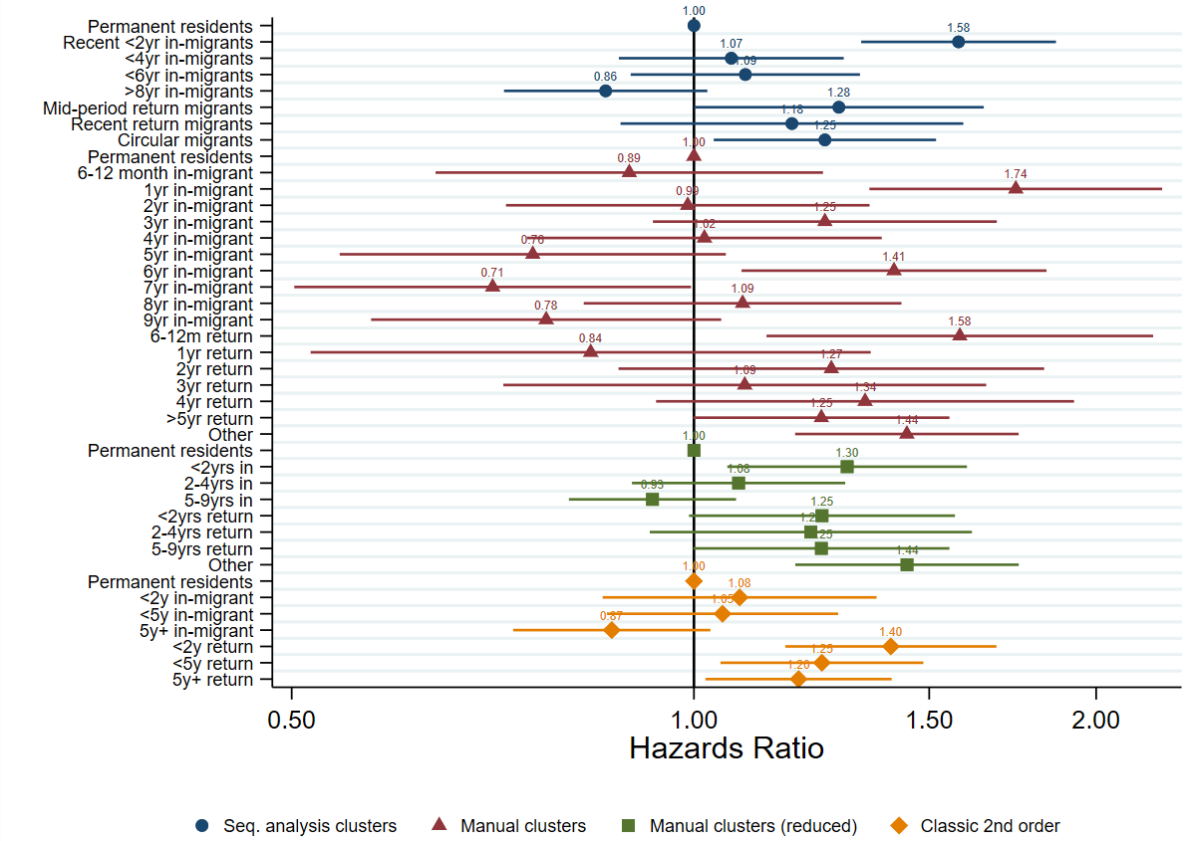


FIGURE 7: COX MODELS OF DEATH ACCORDING TO MIGRATION "STATUS"

6 CONCLUSION

Sequence analysis seemed very promising, but proved complex when tested in an event history framework, even after lifting the sample size constraint. Our general advice for researchers who would think of using migration histories in an event history analysis framework would be to use sequence analysis if one has no prior hypotheses to test and know little on the relationship, provided that the analysis is performed on the whole population at risk (using our method of "clustering the clusters"), and not just on one sample as often done since the number of person-years would not be enough to identify any relevant effect. This will help identify the relevant clusters. In our analysis on Agincourt HDSS, this method identified the recent migrant effect (a reflection of selection prior to migration), an adaptation effect (with the reduction of the migrant- permanent resident gap over time), and higher mortality of return migrants (including circular migrants). From a methodological point of view, this paper gives a proof of concept: we combined two previously-separated life history methods, sequence analysis and event history analysis, by introducing time-varying life histories as a covariate in time-to-failure analysis.

In situations with prior knowledge and informed hypotheses on the relationship, the manual categorisation of migration histories gives the optimum results, provided the categories are reduced to meaningful clusters. As shown above, identifying migration histories using simple criteria like duration in each state (by order: last, last but one, last but two…) allows the researcher to account for the actual frequency of each category. Duration (e.g. in years) being a continuous dimension, the categories can be ordered in a meaningful way. Most of all, the resulting categories can be handled whatever the size of the sample or population and the number of these categories, which will be well below the software's limits. Therefore, this higher-order time-varying migration variable satisfies all the four evaluation criteria: the variable accounts for censoring, for migration status independently from event of interest, for the time-varying nature of migration status, and for lagged migration effects to the highest order possible. With this method, the recent migrant effect, the adaptation effect and those of the return and circular migrants are all well identified.

The 2nd order classic method is sufficient for most analyses, whatever the sample size, since it satisfies the three first criteria set above (censoring, independence from event of interest, time-varying) and partially the fourth (2nd order

effect only). However, this method underestimated the recent migrant effect and adaptation effect, and overestimated the return migration effect by failing to identify the circular migrants.

Although our results based on HDSS data may not be generalisable to any analysis of the effect of migration on an outcome, it is striking how, in the migration-mortality relationship, the interpretation of our results based on cluster from sequence analysis or manual categorization of sequences (both of which identified a residual category interpreted as circular migrants) differ very little from the classic use of 2nd order time-varying variables, even without the circular migration category. A classic 2nd order time-varying variable "augmented" by a circular migrant category (identified by grouping migration histories of order 3 or more), would be quite efficient and easy to implement. However, with the benefit of hindsight and for this particular analysis on Agincourt HDSS data, the manual grouping seems to be the most efficient: it is more tedious than the classic grouping, and does involve some guided decision-making as regard to forming the circular migrant category, but it is still much less burdensome and computer-intensive than sequence analysis. It remains to be seen how the three different approaches can be implemented in other settings, with more complex migration histories or different outcomes, and which will prove most efficient.

7 BIBLIOGRAPHY

Chihaya, G. K., Marcińczak, S., Strömgren, M., Lindgren, U., & Tammaru, T. (2022). Trajectories of spatial assimilation or place stratification? A typology of residence and workplace histories of newly arrived migrants in Sweden. *International Migration Review*, *56*(2), 433–462.

Collinson, M., Tollman, S. M., Kahn, K., Clark, S., & Garenne, M. (2006). Highly prevalent circular migration : Households, mobility and economic status in rural South Africa. *Africa on the move: African migration and urbanisation in comparative perspective*, *35*(69), 194–216.

Dobra, A., Bärnighausen, T., Vandormael, A., & Tanser, F. (2019). A method for statistical analysis of repeated residential movements to link human mobility and HIV acquisition. *PLoS One*, *14*(6), e0217284.

Ginsburg, C., Bocquier, P., Béguy, D., Afolabi, S., Augusto, O., Derra, K., Herbst, K., Lankoande, B., Odhiambo, F., Otiende, M., Soura, A., Wamukoya, M., Zabré, P., White, J. M., & Collinson, M. (2016). Healthy or unhealthy migrants ? Identifying internal migration effects on mortality in Africa using health and demographic surveillance systems of the INDEPTH network. *Social Science & Medicine*, *164*, 59-73.

Gosselin, A., du Loû, A. D., & Lelièvre, E. (2018). How to use sequence analysis for life course epidemiology? An example on HIV-positive Sub-Saharan migrants in France. *J Epidemiol Community Health*, *72*(6), 507–512.

Kahn, K., Collinson, M. A., Gómez-Olivé, F. X., Mokoena, O., Twine, R., Mee, P., Afolabi, S. A., Clark, B. D., Kabudula, C. W., & Khosa, A. (2012). Profile : Agincourt health and socio-demographic surveillance system. *International journal of epidemiology*, *41*(4), 988–1001.

Kuuire, V. Z., Bisung, E., & Were, J. M. (2019). Examining the connection between residential histories and obesity among Ghanaians : Evidence from a national survey. *Journal of Public Health*, *27*, 569–579.

Namin, S., Zhou, Y., Neuner, J., & Beyer, K. (2021). The role of residential history in cancer research : A scoping review. *Social Science & Medicine*, *270*, 113657.

Patel, A., Joseph, G., Killemsetty, N., & Eng, S. (2020). Effects of residential mobility and migration on standards of living in Dar es Salaam, Tanzania : A life-course approach. *Plos one*, *15*(9), e0239735.

Verwimp, P., Osti, D., & Østby, G. (2020). Forced displacement, migration, and fertility in Burundi. *Population and Development Review*, *46*(2), 287–319.

8 APPENDICES

8.1 STATA PROGRAMME: STEP 1

Before running the following code, the program "tmerge.ado" must be installed (available at: https://github.com/bocquier/mighealth/)

The heart of the program is a loop (*n*: 1 to 10) to capture the status *n* years ago. Each loop creates from the original file in long format (each residency episodes ordered by time of occurrence) a new file corresponding to the situation *n* years ago (lag files).

```
forvalues n = 1/10 {
        capture erase lag`n'.dta
        use residency.dta
        keep ID datebeg EventDate EventCode residence datebeg_1 date_OBE
        sort ID EventDate EventCode
        rename residence residence_lag`n'
        * special cases:
```

This is to code episode that runs from the start of observation (here, 1st January 2000) to birth:

```
        replace residence_lag`n'=8 if EventCode==2 & residence==0 // 8 for "not born yet"
```

This is to code episode that runs from the start of observation (here, 1st January 2000) to enumeration:

```
        replace residence_lag`n'=7 if EventCode==1 & residence==0 // 7 for "not enumerated yet"
```

A new code is created for EventCode indicating that only events before end of observation should be lagged:

```
        bysort ID (EventDate EventCode): replace EventCode=50 if EventCode!=9
```

A special case is when the birth date *n* years forward is greater than the last observation date ("date_OBE"):

```
        tempvar postOBE
        bysort ID (EventDate EventCode): gen `postOBE'= ///
```

If the birth *n* years forward is greater than end of observation…

```
        cond(residence_lag`n'==8 & Cmdyhms(month(dofC(EventDate)),day(dofC(EventDate)), ///
        year(dofC(EventDate))+`n',hhC(EventDate),mmC(EventDate),ssC(EventDate))>=date_OBE, ///
```

…then the indicator variable is coded "1":

```
        1, ///
```

Otherwise, the general case is that if the end of the episode falls after the last observation date…

```
        cond(Cmdyhms(month(dofC(EventDate)),day(dofC(EventDate)), ///
        year(dofC(EventDate))+`n',hhC(EventDate),mmC(EventDate),ssC(EventDate))>date_OBE ///
```

… and the beginning of the episode falls before the last observation date…

```
        & Cmdyhms(month(dofC(datebeg)),day(dofC(datebeg)), ///
        year(dofC(datebeg))+`n', hhC(datebeg), mmC(datebeg), ssC(datebeg))<date_OBE, /// then,
```

…then the indicator variable is coded "1":

```
        1, /// else:
```

Otherwise, the indicator variable is coded missing ".":

```
        .)) ///
```

The variable "postOBE" only applies to events before the end of observation (coded "50" above):

```
                if EventCode==50
```

A new variable captures the residency status of the lagged records:

```
        tempvar status_lag
```

By dividing "residence_lag`n'" by "postOBE", the new "status_lag" variable will be coded with the lagged residential status or with a missing value "." if no lagged record:

```
        egen `status_lag'=max(residence_lag`n'/`postOBE'), by(ID)
```

Attribute the lagged status to the OBE record, otherwise the record retains the lagged residential status:

```
        bysort ID (EventDate EventCode): replace residence_lag`n'= ///
                cond(`status_lag'!=. & EventCode==9,`status_lag',residence_lag`n')
```

The records post-OBE are deleted from the database:

```
drop if `postOBE'==1
```

The dates of the lagged events (i.e. except OBE) are lagged by *n* years:

```
replace EventDate=Cmdyhms(month(dofC(EventDate)),day(dofC(EventDate)), ///
        year(dofC(EventDate))+`n',hhC(EventDate),mmC(EventDate),ssC(EventDate)) ///
        if EventCode==50
```

No records post-OBE should remain in the database and therefore are deleted:

```
drop if EventDate>date_OBE & EventCode==50
```

Only records with lagged events and OBE are kept in the database:

```
keep if EventCode==50 | EventCode==9
rename EventDate EventDate_lag
rename EventCode EventCode_lag
```

File remaining from the last but one loop is deleted:

```
local VeryOld = `n'-2
capture erase `hdss'_lag`VeryOld'.dta
```

The new file is saved:

```
save `hdss'_lag, replace
local Old = `n'-1
clear
```

The new file is merged according to time with the file from the last loop ("tmerge.ado" must be installed first):

```
tmerge ID ///
        `hdss'_lag`Old'(EventDate_lag`Old') `hdss'_lag(EventDate_lag) ///
        `hdss'_lag`n'(EventDate_lag`n')
replace EventCode=EventCode_lag if _File==2
label define eventlab 50 "Mig hist", modify
save, replace
} // end of loop
```

Delete unnecessary files after the loop:

```
capture erase `hdss'_lag.dta
forvalues n = 0/9 {
        capture erase `hdss'_lag`n'.dta
}
```

Delete unnecessary records and variables after the loop:

```
drop if EventCode==9 & EventDate<date_OBE
drop EventDate_lag* _File date_OBE
```

Rename and label variables:

```
rename EventDate_lag EventDate
replace EventDate=EventDate_lag10
lab def residence 0 "non-resident" 1 "resident" ///
        6 "not observed yet" 7 "not enumerated yet" 8 "not-born yet", modify
lab val residence_lag* residence
```

Create a new variable with the sequence of residency status over 10 years. First a string version…

```
capture drop mig_history
gen str mig_history=""
forvalues n = 1/10 {
        replace mig_history= ///
                cond(residence_lag`n'==.," ",string(,residence_lag`n')) + mig_history
}
duplicates drop
```

… then an alphanumeric version:

```
capture drop mig_history_NUM
gen mig_history_NUM=mig_history
destring mig_history_NUM, replace
format mig_history_NUM %10.0f
```

Note that migration histories starting with non-residency episodes do not appear as 10-digit numbers:

```
codebook mig_history_NUM

sort CentreId ID EventDate EventCode
```

```
order CentreId ID DoB datebeg EventCode EventDate
compress
save, replace
```

Check the results:

```
browse Id EventCode residence EventDate residence_lag* mig_history mig_history_NUM
```

8.2 STATA PROGRAMME: STEP 4

For the sequence analysis, first ensure that the sq package is installed

```
* ssc install sq
```

sq works with long format, so first reshape the file

```
keep IndividualId EventCode EventDate residence_lag*
reshape long residence_lag , i(IndividualId EventDate EventCode) j(pos)
```

The pos variable needs to be flipped so that the figures are read from left to right (10yrs ago to most recent)

```
recode pos (1=10) (2=9) (3=8) (4=7) (5=6) (6=5) (7=4) (8=3) (9=2) (10=1), gen(pos_rev)
```

Create a unique ID variable of individual + eventdate+code

```
gen double  date = (EventDate)
format date %20.0g
egen int uniqID =concat(IndividualId date EventCode), punct(,) format(%20.0g)
drop date
```

Set the data for sequence analysis (trim drops missing values at beginning or end of sequence- but we shouldn't have such cases)

```
sqset residence_lag uniqID  pos_rev , trim
```

Use the optimal matching with full option to cluster the sequences

```
sqom, full
sqclusterdat ,  keep(IndividualId EventDate EventCode pos residence_lag pos_rev uniqID )
clustermat wardslinkage SQdist, name(wards) add
```

For visualization of the potential clusters, a tree diagram (dendrogram), with n (here 30) branches

```
cluster tree wards ,cutnumber(30) countin horiz
```

To identify the optimum number of clusters (produces a table with F-value, for which the highest value indicates best number of clusters:

```
cluster stop, var(_SQid) matrix(Fclusters)
```

According to the highest value, generate the clusters (in this case 7)

```
cluster gen grps= groups(9) , name(wards)
```

Save file with the unique IDs and the clusters, to be merged back into the main dataset

```
keep uniqID grps
save mig_hist_clusters.dta, replace
```

Use main file, and create again the unique ID

```
gen double  date = (EventDate)
format date %20.0g
egen int uniqID =concat(IndividualId date EventCode), punct(,) format(%20.0g)
drop date
```

Then merge the file with the saved clusters, according to the unique ID:

```
duplicates drop numId EventCode EventDate , force
merge 1:1 uniqID using mig_hist_clusters
```

For “carpet” plots of the sequences by cluster:

```
sqindexplot if grps!=0, by(grps)
```

8.3 STATA PROGRAMME: STEP 6

Proportion in residence for each migration year by cluster

```
forval i=1/40 {
	use ZA031_mighist_sub_samp_`i'.dta, clear

	*create id variable of individual + eventdate+code
	gen double  date = (EventDate)
	format date %20.0g
	egen int uniqID =concat(IndividualId date EventCode), punct(,) format(%20.0g)
	drop date

	* combining cluster info with full data
	duplicates drop numId EventCode EventDate , force
	merge 1:1 uniqID using mig_hist_clusters_`i'.dta
	keep if _merge==3
	collapse (mean) residence_lag* , by(grps)
	reshape long residence_lag, i(grps) j(yr_mig)
	rename residence_lag mean_subs_`i'
	rename grps clusters
	save mean_of_clusters_`i'.dta, replace
}
use mean_of_clusters_1.dta, clear
drop _merge
forval i=2/40 {
	merge 1:1 clusters yr_mig using mean_of_clusters_`i'.dta
	drop _merge
}
save mean_of_clusters_all.dta, replace
```

8.4 STATA PROGRAMME: STEP 7

Sum of absolute deviation and distance matrix

```
*new matrix for final SAD values
matrix sadval = J(600, 600,0)

*filling in the SAD values
use mean_of_clusters_all.dta, clear
forval i=1/40 {
	forval cl_orig=1/15 {
		forval cl_dest=1/15 {
			mkmat mean_subs_`i' if clusters==`cl_orig', matrix(v_`i'_`cl_orig')

			local n=1
			while `n' <=40 {
				mkmat mean_subs_`n' if clusters==`cl_dest', matrix(v_`n'_`cl_dest')

				*di "Origin:" "v_`i'_`cl_orig'" " Destination:" "v_`n'_`cl_dest'"
				matrix diff = v_`i'_`cl_orig' - v_`n'_`cl_dest'
				mata: st_matrix("absval",abs(st_matrix("diff")))
				 * Compute Sum of Absolute Deviation b/w data & intrinsic distributions
				  mata: X = st_matrix("absval")
				mata: st_matrix("SAD", colsum(X[.,1]))
				matrix sadval[(15*(`i'-1)) + `cl_orig', (15*(`n'-1))+ `cl_dest'] = SAD[1,1]
				*matrix list sadval
				local n = `n' + 1
			}  // close while
		} // clusters dest
	} // clusters orig
}
mata: st_replacematrix("sadval", makesymmetric(st_matrix("sadval")))
```