Interval-Censored Event History Analysis

Daniel Courgeau; Jamal Najim

Population: An English Selection, Vol. 8 (1996), 191-207.

Stable URL:
http://links.jstor.org/sici?sici=1169-1018%281996%292%3A8%3C191%3AIEHA%3E2.0.CO%3B2-5

Population: An English Selection is currently published by Institut National d'Études Démographiques.

INTERVAL-CENSORED EVENT HISTORY ANALYSIS

Event history analysis was no doubt already in existence when J. Hajnal used census information on marital status to investigate cohort nuptiality or when L. Henry studied cohort fertility from numbers of children ever born. It developed when other elements added to this information: age at first marriage, dates of birth of children; in this respect, the 1946 census in Britain was a turning-point. It has since branched out in two directions: on the one hand, surveys collecting retrospective information on respondent's life course, family and occupational histories, residential mobility...; on the other, individual biographies compiled by extracting information from administrative sources (vital registration data, census schedules, notifications of residence...). Each method has constraints – the second, in particular, because of depending on the availability of administrative data: job changes are never registered and in France, neither are residential moves. That limits observation to the information collected when a census is held or a vital event registered. Daniel COURGEAU* *and* Jamal NAJIM** *test here the validity of these incomplete data by comparing them to information supplied by an* ad hoc *survey.*

The use of retrospective survey and population register data has permitted numerous applications of event history analysis in all fields of demography. While these data are incomplete in that they observe the individual's life course only up to the time of the survey or when the register was consulted, reliable methods of analysis have been developed taking right-censoring into account (Courgeau and Lelièvre, 1989, 1993). But there are other kinds of data collection that produce more fragmented biographical data and call for specific forms of analysis. Let us take some concrete examples.

First, INSEE's Demographic Panel Survey (Échantillon Démographique Permanent, which we shall call the EDP survey)[1] (Sautory, 1987) is exceptionally rich, both in terms of size (slightly more than a one per cent sample of the population present in France, at every census from 1968 to 1982, but only half this sample for the 1990 census) and of wealth of information.

* INED.
** Institut de Statistique, Université Pierre et Marie Curie, Paris; INED.

[1] Similar panel surveys exist in other countries, for instance the OPCS Longitudinal Study in the UK (Social Science Research Unit, 1990).

The data file contains all the information in the individual census schedules since 1968 coupled with the corresponding registration data. The family history analysis of individuals born since 1953 is consequently straightforward, given the accuracy with which these events are dated. But the study of migration or occupational histories is another matter: residence and occupation are only known at time of census and of family events. We have no precise dates to mark residential or occupational mobility; we only know that a move has occurred between two censuses or family events. The usual methods of event history analysis cannot deal with such *interval-censored* data.

Second, the survey on social, geographical and wealth mobility in 19th and 20th century France, conducted by Jacques Dupâquier and Denis Kessler (Dupâquier, 1981), provides a most interesting file of historical data. Starting from 3,000 couples living at the time of the First Empire, whose family name began with the three letters TRA, and who were selected so as to be proportionally representative of the census population of 1806 by *départements* (in their present boundaries), the survey traces their male descent lines down to the present time. Again, we know current place of residence and current occupation at time of each vital event, but not the exact date of any move or job change. The case is therefore the same as in the preceding example.

In the present article, we shall propose and test a number of methods which allow for the fragmented nature of the data. We shall use them to estimate as accurately as possible the distribution of migrations or job changes over time, and show how various individual characteristics affect this distribution.

These methods will be tested on the complete retrospective data from the 'Triple Biography: Family, Occupational and Migration' (3B) survey conducted by INED in 1981. By fragmenting these data to make them comparable to the EDP and TRA data, we can test the loss of accuracy and degree of error that such methods entail. We suppose that we observe current residence and occupation of respondents at time of each census (1926, 1936, 1946, 1954, 1962, 1968 and 1975) and at time of their successive family events: in other words, we transform a complete data set into an interval-censored data set. We consider 1,995 female respondents born between 1911 and 1935.

We shall first discuss the hypotheses that are required for analysing interval-censored data. We shall then estimate non-parametric models to measure the effect of duration on the probabilities of occurrence of residence or job changes and semi-parametric models to take into account the effect of various individual characteristics. We shall also present some of the parametric models most commonly used for such analysis and discuss their drawbacks. Throughout the article, we shall check against the 3B survey data the validity of the different models and hypotheses.

In this study, we generalize some partial results we have presented elsewhere (Courgeau, 1993).

I. – Two fundamental hypotheses

Observing place of permanent residence or occupation only at specified times may leave some information in the shade, when several moves or job changes occur in the interval between two observations. It must therefore be assumed that no more than one of the events studied (say, migration) can occur between two observation times. Thus, to study first change of *département* occurring after marriage, we assume that no departure from and return to the *département* where the newly-weds set up home can occur between two observation times.

To minimize this drawback, the density of the events defining the individual's position must be sufficient for few migrations or job changes to slip through. We can therefore suppose that observing the individual at time of censuses *and* family events will be more satisfactory for estimating his or her mobility than using only one or the other. We shall use the 3B data to judge how far this improves the results, and also to show how much information does slip through in the case of interval-censored data.

Another hypothesis is also necessary to estimate durations of stay from interval-censored histories: the events defining the individual's spatial or social position must be independent of the geographical and occupational mobility we wish to measure. Otherwise interactions between the different phenomena would disturb the estimation and introduce errors.

To demonstrate more clearly what we mean, let us suppose we want to estimate duration of stay in the first dwelling occupied after marriage. The results can easily be generalized to other situations.

We do not observe directly the date of departure from this dwelling but locate it on a calendar where moves, births of children and censuses are mingled. We know when the stay began (date of marriage), but only that it ended somewhere between two birth or census dates, or had not ended at the time of the last census or when the individual died.

Some of these events, censuses for example, are obviously independent of the migration considered and will not bias the estimation. Others, such as births of children, will more frequently be dependent: a couple who set up home in a small dwelling are likely to move to a larger one when they start a family. In this case, the observation of successive dates of birth may not be independent of the fact that a move did or did not take place between them[2]. The bias that introduces will be all the greater when we consider short-distance mobility, which is closely linked to changes in family size; long-distance mobility (to another *département* or region) is affected less.

---

[2] This concept of unilateral or local independence was introduced by Schweder (1970) and developed by Aalen *et al.* (1980) and Courgeau and Lelièvre (1986, 1989, 1993).

A probabilistic formulation of these problems is given in the Appendix. It shows what conditions are required for durations of stay to be correctly estimated.

By applying these methods to the two data sets – the complete retrospective data from the 3B survey and those artificially interval-censored –, we have the opportunity, throughout this study, to test precisely the dependence between family events and geographical or occupational mobility.

II. – Estimation of durations of stay

We consider the general case where, starting from the kth geographical or occupational move that occurs at random time T_k, we observe the time elapsed between this move and the next one ($T = T_{k+1} - T_k$). We suppose here that the random variables T_k and T are discrete (say, annual) and independent. Thus, we measure the probabilities $m_h = P\ (T_k = t_h)$ that the kth migration will occur in year t_h and $v_i = P\ (T = t_i)$ that the following migration will occur after duration t_i ($1 \leq h \leq r$, $1 \leq i \leq s$ where r and s are the longest observed durations).

Given the method of observation, T_k and T_{k+1} can only be located in relation to the different family events or successive census dates. Each observation is thus of the form (t^j, t^{j+1}, $t^{j'}$, $t^{j'+1}$) such that $t^j \leq T_k \leq t^{j+1}$ and $t^{j'} \leq T_{k+1} \leq t^{j'+1}$. The data may be right-censored, when the (k+1)th migration does not occur before the last observed event.

The probabilities m_h and v_i can be estimated by maximizing the likelihood of the observations. To write this likelihood, we introduce parameters $\alpha^l_{h,i}$ which, for individual l, are equal to 1 when $t^j \leq t_h \leq t^{j+1}$ and $t^{j'} \leq t_h + t_i \leq t^{j'+1}$, otherwise to 0. Hence, the likelihood function can be written:

$$L = \prod_{l=1}^{N} \left(\sum_{h=1}^{r} \sum_{i=1}^{s} \alpha^l_{h,i}\, m_h\, v_i \right) \qquad [1]$$

where N is the number of individuals having previously moved k times. The values of m_h and v_i that maximize this likelihood are obtained by an iterative fitting procedure similar to the one proposed by De Gruttola and Lagakos (1989). We introduce a random variable $\alpha^l_{h,i}$ equal to 1 if the actual but unobserved value of (T_k, T) for the lth individual equals (t_h, t_i) and to 0 otherwise. Hence the conditional expectation of this variable given α^l is:

$$\mu^l_{h,i} = \frac{\alpha^l_{h,i}\, m_h\, v_i}{\sum_{h,i} \alpha^l_{h,i}\, m_h\, v_i} \qquad [2]$$

From this, we derive the following estimates of m_h and v_i:

$$\hat{m}_h = \frac{\sum_{l,i} \mu^l_{h,i}}{N} \quad \text{and} \quad \hat{v}_i = \frac{\sum_{l,h} \mu^l_{h,i}}{N} \qquad [3]$$

To obtain these estimators, we take arbitrary initial values of m_h and v_i and calculate the corresponding values of $\mu^l_{h,j}$ from [2], then recalculate new values for $\hat{m}_h$ and $\hat{v}_i$ from [3]. We repeat the process until the difference between the successively refined values of m and v becomes negligible. De Gruttola and Lagakos (1989) have demonstrated that when this convergence is achieved, we have a maximum likelihood estimate or a saddle point. By computing the opposite of the matrix of second derivatives of log L, we verify that we are not at a saddle point[3] and obtain, from its inverse, the variance and covariance matrix of $(\hat{m}, \hat{v})$[4].

III. – Application to the artificially interval-censored 3B survey data

We shall now apply these methods to the 3B survey data which we have fragmented for the occasion. By comparing the complete survey data to the interval-censored data, we can see to what extent the above hypotheses are verified.

Changes of residence after marriage When we know the home of a newly-wed couple, which is usually noted on the marriage certificate, we can study departure from this home, or first moves after marriage. This is the simplest situation, where the starting point is known (taken as time 0) and the finishing point can occur between two different kinds of events.

Let us suppose that we want to, or have to (this is the case for the TRA survey), use only dates of family events occurring after marriage. We shall first consider changes of dwelling, which verify the first hypothesis: the probability that an individual will return to a dwelling he or she has previously occupied is very low (Courgeau, 1973, 1979). Figure 1 shows the cumulative distribution function estimated from the whole set of observed data and from only the information provided at time of successive family events, estimated with a 95% confidence interval. The distribution function gives the probability that an individual will have moved by a specified duration, given in years elapsed since marriage. We can see

(3) If the eigenvalues of this matrix are all positive, the point considered is a maximum likelihood one.

(4) These probabilities and the variance and covariance matrix can be estimated by using computer programmes npara1.C and npara2.C written by J. Najim (1994).

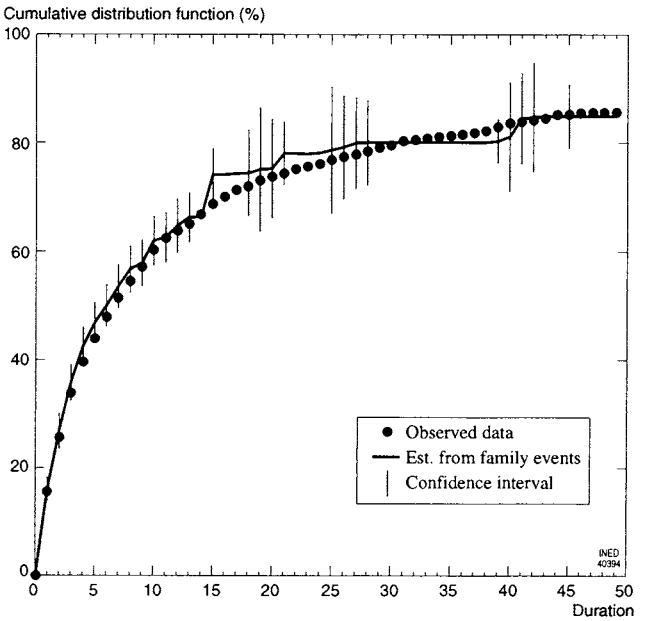

Figure 1. – Cumulative distribution (%) of first moves after marriage, estimated from observed data and from data interval-censored by family events (with a 95% confidence interval)

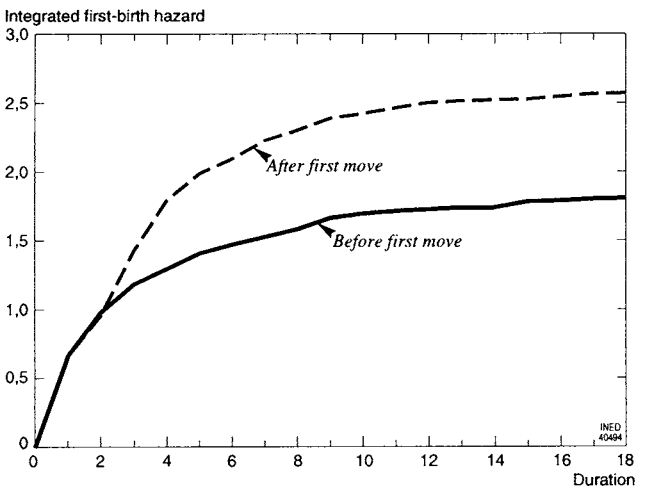

Figure 2. – Integrated first-birth hazards depending on whether or not the birth preceded the first move after marriage

that while the probability of remaining sedentary after a stay of 50 years is not affected by censoring (0.15), the distribution over time is very different for the complete and the interval-censored data: between 3 and 20 years, the curve derived from family data is higher than the actual curve. This would suggest that the second hypothesis is not verified: there is not independence between changes of dwelling and family events. Let us look more closely at the interactions between migration and fertility. We can, for instance, calculate the integrated hazard of a first birth, depending on whether or not there was a move before the birth occurred. If the hazards are identical, then it is verified that the observation of first births is independent of the phenomenon studied, in this case first moves after marriage. In fact, Figure 2 shows that, for durations higher than two years, the first-birth integrated hazards are larger for individuals who have moved from their dwelling. As indicated in the Appendix, this dependence results in higher estimates of cumulative distribution from the interval-censored than from the observed data, at least for durations of 3 to 20 years elapsed since marriage.

We now consider census dates, which are obviously independent of first moves after marriage. Figure 3 compares the cumulative distribution function estimated from census information and from all observations. The two are closely entwined across all durations, but the former is much less smooth. If we now use dates of censuses together with dates of family

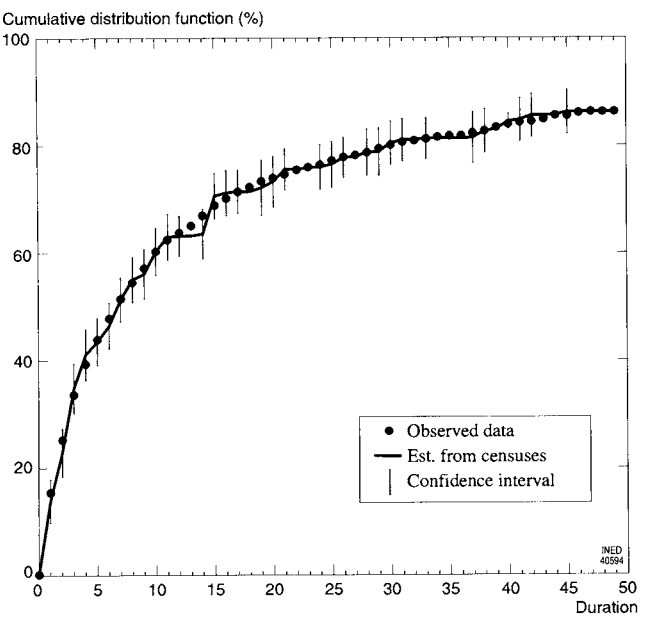

Figure 3. – Cumulative distribution (%) of first moves after marriage, estimated from observed data and from data interval-censored by censuses (with a 95% confidence interval)

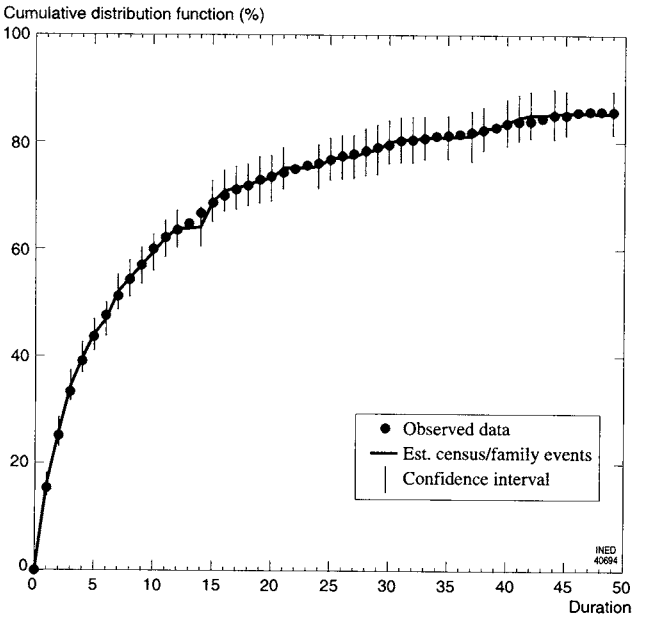

Figure 4. – Cumulative distribution (%) of first moves after marriage, estimated from observed data and from data interval-censored by censuses and family events (with a 95% confidence interval)

events, the estimate is improved further (Figure 4), in spite of the dependence illustrated above. Apparently combining independent and dependent events, thus increasing the density of events used to locate the individual, makes up for the error due to dependence. It is thus preferable to use all the information from the EDP rather than only census information.

Changes of département after marriage

Let us now consider moves from one *département* to another. They should *a priori* verify the second hypothesis better: we have shown elsewhere (Courgeau, 1985) that mobility related to family events is essentially short-distance, while long-distance mobility is motivated by economic reasons. Family events might then be expected not to be much affected by moves between *départements*. In contrast, the first hypothesis will no longer be verified: we have shown (Courgeau, 1973, 1979) that roughly 16% of residential moves bring people back to the *département* where they used to live.

This time, whether we use family events alone (Figure 5), censuses alone (Figure 6) or both (Figure 7), the cumulative distribution function

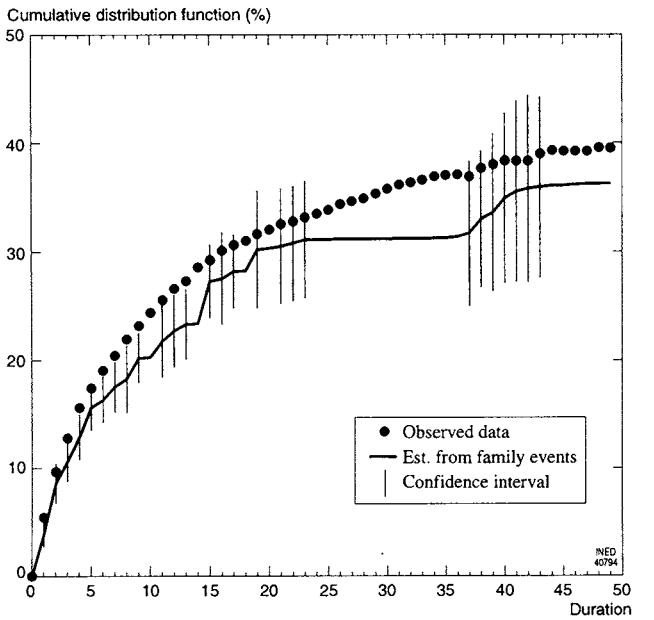

Figure 5. – Cumulative distribution (%) of first changes of *département* after marriage, estimated from observed data and from data interval-censored by family events (with a 95% confidence interval)

is always below the observed one. Some moves across *départements* are omitted because of return migration: the probability of not changing *départements* within a period of 50 years is 0.605 when all observations are used, it rises to 0.636 when estimated from family events alone, and falls to 0.620 in the other two cases. The timing of moves across *départements* is also modified: the difference between the observed and estimated distributions is greatest for the shorter durations, then decreases and the estimation even retrieves belated migrations occurring after 40 years in the same *département*. Once again, the estimation based on *département* of residence at time of census and of family events gives the best results.

Departure from parental home and the following move

We shall end this series of examples by studying the occurrence of two events for which both the time origin and the failure event can be interval-censored: the departure from the parental home and the first move after this departure. Figure 8 gives the cumulative distribution functions for departure from parental home estimated from observed and from interval-censored data (census plus family events), with a 95% confidence interval.

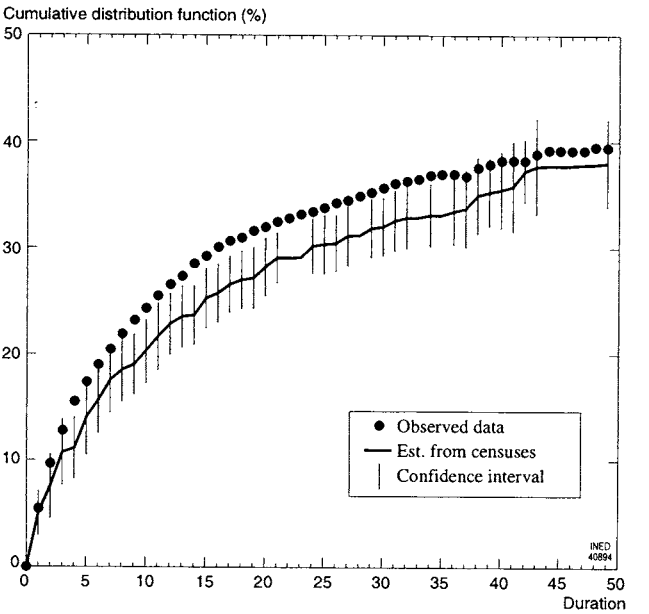

Figure 6. – Cumulative distribution (%) of first changes of *département* after marriage, estimated from observed data and from data interval-censored by censuses (CI 95%)

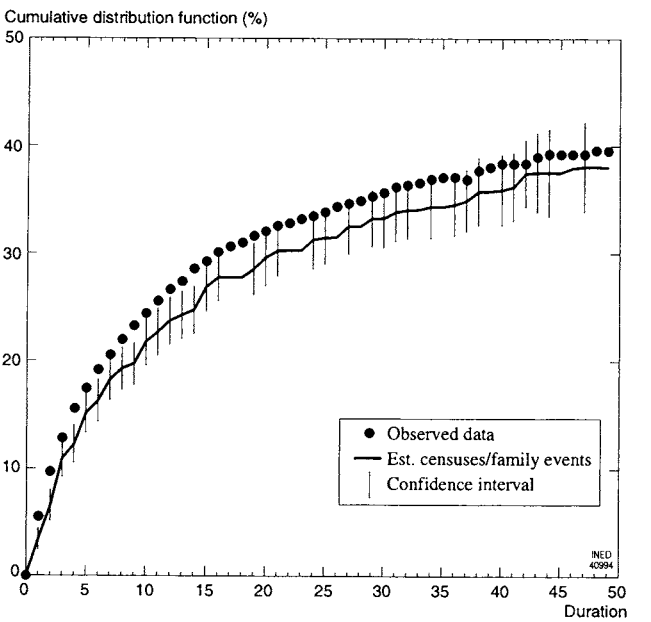

Figure 7. – Cumulative distribution (%) of first changes of *département* after marriage, estimated from observed data and from data interval-censored by censuses and family events (CI 95%)

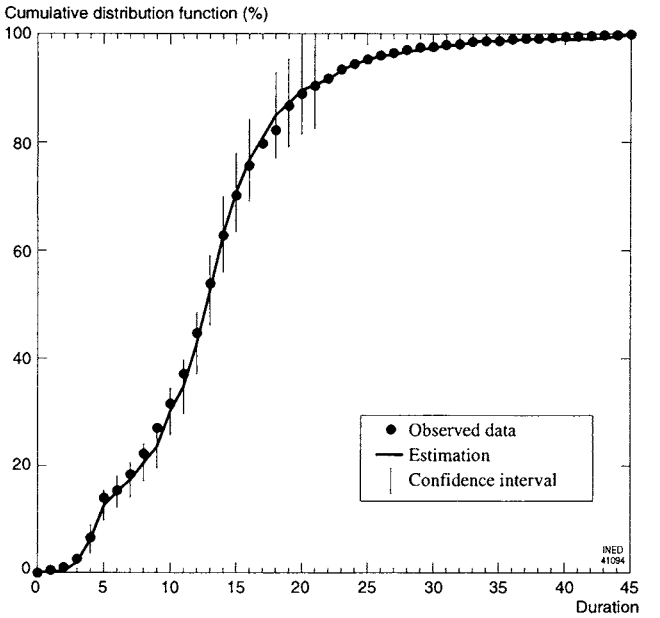

Figure 8. – Cumulative distribution (%) of departures from parental home, estimated from observed data and from interval-censored data (with a 95% confidence interval)

We recall that our sample consists of 1,995 women born between 1911 and 1935. All the women in the sample have left their parents' home, which gives this event an ultimate frequency of 1, and the observed data are always within the confidence interval for the interval-censored data. Figure 9 illustrates the corresponding timing and intensity of first moves after this departure from the parental home (with the same confidence interval). The intensity is estimated perfectly (0.67) and the observed data are again always within the confidence interval for the interval-censored data. We note, however, that this confidence interval is much greater than that observed for departure from the parental home. This is an effect of the imprecision concerning the beginning as well as the end of the stay.

IV. – The effect of some individual characteristics

Now let us look at the effect of various individual characteristics on these durations of stay. For this, we utilize a semi-parametric proportional hazards model (Courgeau and Lelièvre, 1989, 1992).

We still have, for each individual, observations of the form (t^j, t^{j+1}, $t^{j'}$, $t^{j'+1}$) and a vector of m characteristics that can be quantitative variables (age at end of schooling, child's birth order, etc.) but are generally quali-

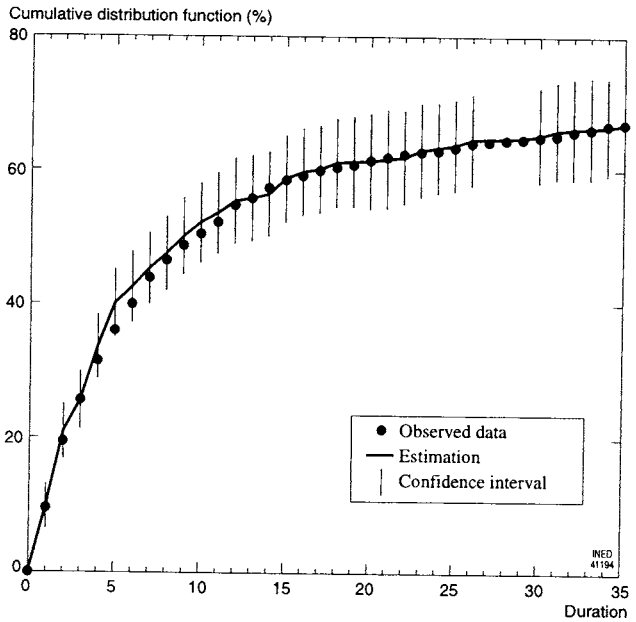

Figure 9. – Cumulative distribution (%) of first moves after departure from parental home, estimated from observed data and from interval-censored data (with a 95% confidence interval)

tative variables, represented in binary form (0 if the person is single, 1 if married, etc.).

Again, we attempt to measure the probabilities $m_h = P(T_k = t_h)$ which have the same significance as before, and the probabilities of occurrence of the second event at duration t_i since the first, conditional on individual characteristics $x_l = (x_l^1, \ldots x_l^n)$, that is $v_i(x_l) = P(T = t_i \mid x_l)$. The model being a discrete time proportional hazards model, we can write (Kalbfleisch and Prentice, 1980):

$$P(T > t_i \mid x_l) = [P(T > t_i \mid x = 0)]^{\exp(x_l\beta)} = (p_1 \times p_2 \times \ldots \times p_i)^{\exp(x_l\beta)} \quad [4]$$

where $p_i = P(T > t_i \mid T > t_i - 1,\ x = 0)$ is estimated for an individual whose characteristics are all null. Hence:

$$v_i(x_l) = (p_1 \times p_2 \times \ldots \times p_{i-1})^{\exp(x_l\beta)} (1 - p_i^{\exp(x_l\beta)}) \quad [5]$$

Under these conditions, the likelihood function is written:

$$L = \prod_{l=1}^{N} \left(\sum_{h=1}^{r} \sum_{i=1}^{s} \alpha_{h,i}^{l} m_h v_i(x_l) \right) \quad [6]$$

To estimate m_h, p_i and β, we can use the above algorithm in combination with that of Newton-Raphson (Kim *et al.*, 1993; Najim, 1994). We

estimate simultaneously the variance and covariance matrix of these different parameters[5].

Again, we use the 3B survey data file to test the quality of the estimation. We now turn our attention to the validity of the β parameters estimated either from all the observed data or from the interval-censored data.

The 3B survey collected information on tenure status at entry into and exit from each dwelling. It was thus possible to capture changes in tenure status not only at the time of a move, but also within periods of residence in a same dwelling. We study here the first dwelling occupied after parents' home and the following move.

We distinguish the behaviour of individuals whose tenure status changed in the course of a same period of residence (initially they were living at their parents') from those who left their parents' home either to be housed free of charge by their employer or to become owner-occupiers. It is the effect of these three characteristics on the following residential move that we analyse here. In Table 1, we show this effect as estimated from the whole data file and from the fragmented data obtained at time of censuses and family events.

TABLE 1. – EFFECT OF INDIVIDUAL CHARACTERISTICS ESTIMATED FROM COMPLETE AND INTERVAL-CENSORED DATA, SEMI-PARAMETRIC OR PARAMETRIC (MOVER-STAYER) PROPORTIONAL HAZARDS ESTIMATION

Characteristic	Observed data		Interval-censored data			
			Semi-parametric estimation		Mover-stayer model estimation	
	β	Standard deviation of β	β	Standard deviation of β	β	Standard deviation of β
Living at parents' home	– 2.298	0.222	– 2.341	0.227	– 2.225	0.222
Housed by employer	0.575	0.058	0.662	0.059	0.675	0.061
Owner-occupier	– 1.619	0.171	– 1.641	0.173	– 1.637	0.179

The results are very similar whether complete data or interval-censored data are used. Individuals whose tenure status changed while they were in the same dwelling, generally farmers, were the least mobile, followed by owner-occupiers, while those housed by their employers (mostly domestics: see Villeneuve-Gokalp, 1994) were the most mobile; tenants formed the control group ($\beta = 0$). The standard deviation of the β parameters increases slightly from the complete to the interval-censored data, while all the effects remain significant.

[5] The computations required are considerable; they can be handled by the Semipar.e programme written by Najim (1994).

We can thus conclude that estimation of the effects of these characteristics is not sensitive to the fact that the data are interval-censored.

V. – Parametric estimation

The most common method for the parametric modelling of durations of stay, whether residential or occupational, is to use either a Gompertz model or a mover-stayer model (Blumen *et al.*, 1955; Courgeau, 1973, 1979; Ginsberg, 1979; Myers *et al.*, 1967). These are continuous time models and the annual probabilities used above are replaced by probability densities $m(\theta)$ for the first event and $v(t)$ for the second in relation to the first. In this case, the likelihood can be written:

$$L = \prod_{l=1}^{n} \int_{t_l^i}^{t_l^{i+1}} m(\theta) \int_{t_l^r}^{t_l^{r+1}} v(t) dt \, d\theta \qquad [7]$$

In fact, $m(\theta)$ is generally estimated non-parametrically, assuming T^k to be a discrete random variable. It is $v(t)$ that is assumed to follow a Gompertz or mover-stayer model. Given a mover-stayer model, we write:

$$v(t) = \rho k \exp(-\rho t) \exp(x\beta) [1 - k(1 - \exp(-\rho t))]^{\exp(x\beta)-1} \qquad [8]$$

where k, ρ and β are the parameters to be estimated. Given a Gompertz model, we have:

$$v(t) = \lambda\mu \exp(x\beta) \exp\{\mu t + \lambda \exp(x\beta) [1 - \exp \mu t]\} \qquad [9]$$

where λ, μ and β are the parameters to be estimated.

Estimating the parameters from likelihood [7] is straightforward using the Newton-Raphson method, and is naturally much less time-consuming than non-parametric or semi-parametric estimation[6]. We note, however, that these models are much more sensitive than the semi-parametric ones presented above to the effects of unobserved heterogeneity (Courgeau and Lelièvre, 1989, 1992). This restricts their utilization for event history analysis.

To illustrate the method, we show in Table 1 the mover-stayer model parameter estimates for individuals living in their parents' home, housed free of charge by employer, and owner-occupiers. There is nothing to distinguish them from the other parameter estimates and their standard deviation is of the same order. The advantage of such models lies in their economy of computer time. But they impose a parametric distribution which must be checked to see how it fits the data.

[6]This parametric estimation can be performed by computer programmes migsed2.c, migsed4.2 and gomp2.c for models in which individual characteristics are not entered and by migsed6.c for the mover-stayer model in which they are; they were all written by J. Najim (1994).

Conclusion

We have solved here the theoretical problems set by the utilization of interval-censored data, under certain conditions which it is important to verify.

The first condition concerns the density in time of the events permitting the location of the individual in geographical or occupational space. The higher this density, the more chance that events that are close to others in time will not be omitted. We saw when studying the case of migration to another *département* that such omission may be substantial. The cumulative distribution curve was also deflected: at higher durations, it included events that should not have been there (moves of higher orders following a return to the *département* of first residence).

The second condition concerns the dependence that may exist between the events defining the individual's spatial location and his or her geographical or occupational mobility. This dependence modifies the timing of the phenomenon studied, but has little effect on its intensity. We saw that it affected mainly the historical data (Dupâquier's 'TRA' survey) where individuals were located only at time of family events. The impact is particularly clear for short-distance mobility, which is very sensitive to family events, and becomes negligible for long-distance mobility (moves to another *département* or region). More in-depth investigation of the interactions between the different demographic phenomena is needed to estimate more precisely the timing of short-distance mobility.

In the framework of this study, we have developed a number of computer programmes to handle all the non-parametric, semi-parametric or parametric estimations based on interval-censored event history data. (These programmes, written in C language, are available from the authors.) As we stated in our introduction, many files contain data of this kind, and it is essential to allow for such censoring to obtain correct estimates of durations of stay and of the effects of individual characteristics on these durations.

Before using these programmes, it is also important to check the quality of the data contained in the files. The EDP survey data, for instance, need to be controlled for misreporting or omissions in the census schedules; also, some individuals and whole households escape observation (Coeffic, 1993).

The pursuit of this work lies not only in the exploitation of existing interval-censored event history sources, but also in the development of models which allow for the fact that the fundamental hypotheses are not verified. The utilization of large data sets from population registers or event history surveys should make it possible to correct these errors. There is still a lot of work in hand.

Daniel COURGEAU, Jamal NAJIM

REFERENCES

Aalen O., Borgan O., Keiding N., Thorman J., 1980.– «Interaction between life history events: Non parametric analysis for prospective and retrospective data in presence of censoring», *Scandinavian Journal of Statistics*, 7, 161-171.

Blumen I., Marvin K., Mc Carthy P.J., 1955.– *The industrial mobility of labour as a probability process*, Cornell Studies of Industrial and Labour Relations, Vol. VI, New York.

Coeffic N., 1993.– «L'enquête post-censitaire de 1990. Une mesure de l'exhaustivité du recensement», *Population*, 6, 1655-1682.

Courgeau D., 1973.– «Migrants et migrations», *Population*, 1, 95-129.

Courgeau D., 1979.– «Migrants and migrations», *Population, Selected Papers*, 3.

Courgeau D., 1985.– «Changements de logement, changements de départements et cycle de vie», *L'Espace géographique*, 4, 289-306.

Courgeau D., 1993.– «An attempt to analyse individual migration histories from data on place of usual residence at the time of certain vital events. France during the nineteenth century», in *Methods in Historical Demography*, D. Reher and R. Schofield (eds.), Clarendon Press, Oxford, 206-222.

Courgeau D., Lelièvre E., 1986.– «Nuptialité et agriculture», *Population*, 2, 303-326.

Courgeau D., Lelièvre E., 1989.– *Analyse démographique des biographies*, Editions de l'INED, Paris.

Courgeau D., Lelièvre E., 1992.– *Event history analysis in demography*, Clarendon Press, Oxford.

Dupâquier J., 1981.– «Une grande enquête sur la mobilité géographique et sociale du XIXè et XXè siècles», *Population*, 6, 1164-1167.

Ginsberg R., 1979.– «Timing and duration effects in residence histories and other longitudinal data. II Studies of duration effects in Norway, 1965-1971», *Regional Sciences and Urban Economics*, 9, 369-392.

De Gruttola V., Lagakos S.W., 1989.– «Analysis of doubly-censored survival data, with application to AIDS», *Biometrics*, 45, 1-11.

Kalbfleisch J., Prentice R., 1980.– *The statistical analysis of failure time data*, Wiley, New York.

Kim M.Y., De Gruttola V.G., Lagakos S.W., 1993.– «Analysing doubly censored data with covariates, with application to AIDS», *Biometrics*, 49, 13-22.

Myers G.C., McGinnis R., Masnick G., 1967.– «The duration of residence approach to a dynamic stochastic model of internal migration: a test of the axiom of cumulative inertia», *Eugenics Quarterly*, 14(2), 21-126.

Najim J., 1994.– *Les méthodes d'estimation de la probabilité d'arrivée d'un événement et leurs utilisations en logiciels*, Mémoire de l'Institut de Statistique de l'Université Pierre et Marie Curie, Paris.

Sautory O., 1987.– «L'échantillon démographique permanent de l'INSEE», *Courrier des Statistiques*, 41, 1-4.

Schweder T., 1970.– «Composable Markov processes», *Journal of Applied Probabilities*, 7, 400-410.

S.S.R.U. 1990.– *OPCS Longitudinal Study. User manual*, London.

Villeneuve-Gokalp C., 1994.– «Les gens de maison», *Population*, 4-5.

APPENDIX

Probabilistic formulation of interactions between dates of moves, of censuses and of family events

We formalize here more fully how the probabilities we estimate compare to those we observe.

For simplicity's sake, let us work on first moves after marriage. The results obtained can easily be generalized to more complex situations. We denote T_1 the random variable corresponding to the duration between marriage and the first following move. Let T^1, T^2,... T^m be the durations between marriage and the different

events which locate the individual (births and censuses). These random variables are positive and ordered.

While our data are not sufficient to estimate the probabilities of occurrence of first moves after marriage, we can estimate those of more complex events which involve migrations, births and censuses all together. We know the exact date of beginning of duration (the date of marriage), but only that it ended some time between two dates of birth or census, or that it had not ended by the time of the last census or the individual's death. We estimate therefore conditional probabilities. If the first move occurred between the jth and the $(j+1)$th events observed at durations t and t', the following probability can be estimated:

$$P(t \leq T_1 \leq t' \mid T^j = t \cap T^{j+1} = t') = \frac{P(t \leq T_1 \leq t' \cap T^j = t \cap T^{j+1} = t')}{P(T^j = t \cap T^{j+1} = t')} =$$

$$P(t \leq T_1 \leq t') \frac{P(T^j = t \cap T^{j+1} = t' \mid t \leq T_1 \leq t')}{P(T^j = t \cap T^{j+1} = t')} \qquad [1]$$

where the last relation is obtained by applying the compound probabilities theorem twice. We have thus obtained the factor by which the probability that we wish to estimate, $P(t \leq T_1 \leq t')$, needs to be multiplied to obtain a probability that we can actually estimate. If the probability of occurrence of the two events that locate the individual at two different times is the same irrespective of whether or not the individual had moved between the two events, our estimated probability is equal to that which we wish to measure.

If the first move occurs between two censuses, this hypothesis is verified perfectly. If it occurs between two family events, we have to verify whether there is independence between the births and the first move.

Similarly, when no first move has occurred by time of last census, at duration t, we estimate the probability:

$$P(T_1 \geq t \mid T^m = t) = P(T_1 \geq t) \frac{p(T^m = t \mid t \leq T_1)}{P(T^m = t)} \qquad [2]$$

Again, the survivor function $P(T_1 \geq t)$ is correctly estimated if the probability of occurrence of the last event does not depend on whether the individual has previously moved or not.

Going back to our test on the interactions between migration and fertility (section III), we can re-write formula [1] where $t = 0$ given that we are working on the interval between marriage and first birth:

$$P(T_1 \leq t' \mid T^1 = t') = P(T_1 \leq t') \frac{P(T^1 = t' \mid T_1 \leq t')}{P(T_1 = t')} \qquad [3]$$

In this case, the cumulative distribution function we estimate will be higher than the one we observe, since $P(T^1 = t' \mid T_1 \leq t') > P(T^1 = t')$ for the durations when first births after marriage occur. This result can be completed by studying births of higher orders.