

LONGITUDINAL ANALYSIS OF LABOUR FORCE SURVEY DATA

Geoff Rowe and Huan Nguyen¹

ABSTRACT

The Labour Force Survey (LFS) was not designed to be a longitudinal survey. However, given that respondent households typically remain in the sample for six consecutive months, it is possible to reconstruct 6-month fragments of longitudinal data from the monthly records of household members. Such longitudinal data – altogether consisting of millions of person-months of individual and family level data – is useful for analyses of monthly labour market dynamics over relatively long periods of time, 25 years and more.

We make use of these data to estimate hazard functions describing transitions among the labour market states: self-employed, paid employee and not employed. Data on job tenure, for the employed, and on the date last worked, for the not employed – together with the date of survey responses – permit the estimated models to include terms reflecting seasonality and macro-economic cycles as well as the duration dependence of each type of transition. In addition, the LFS data permits spouse labour market activity and family composition variables to be included in the hazard models as time-varying covariates. The estimated hazard equations have been incorporated in the LifePaths microsimulation model. In that setting, the equations have been used to simulate lifetime employment activity from past, present and future birth cohorts. Simulation results have been validated by comparisons with LFS age profiles of employment/population ratios from the period 1976 to 2001.

Keywords: microsimulation, counting process, censoring, truncation, employment

1. INTRODUCTION

In recent years, there has been an increasing understanding of the importance of studying labour market dynamics using individual level data. For this purpose, new panel surveys have been developed, for example, the Survey of Income and Labour Dynamics (SLID) (Statistics Canada, 1998). But, existing Labour Force Survey (LFS) (Statistics Canada, 2002) data provides a virtually untapped historical resource, in the form of many fragmentary event histories. From a conventional standpoint, the data currently comprises a time series of more than 300 cross-sectional surveys that were conducted monthly over more than 25 years. However, from a longitudinal perspective, those data consist of about 6.5 million fragmentary event histories covering overlapping time intervals within the past quarter century and totaling over 34 million person-months of observation.

Our analysis was specifically directed towards development of hazard models to be incorporated in LifePaths (Statistics Canada, 2001) – a micro-simulation model of the Canadian population. Further detail about the LifePaths model is available on the Statistics Canada website at <http://www.statcan.ca/english/spsd/index.htm>.

The paper is organized in the following way. In Section 2, we discuss the some features of LFS data when reorganized as longitudinal records and we present three examples comparing estimates derived from the resulting longitudinal file with corresponding estimates from external sources. In Section 3, we focus on the use of the data to model employment activity for LifePaths. There, we discuss use of LFS data for estimating hazard equations describing employment dynamics. Finally, we present some illustrations of estimation results and a validation of LifePaths simulations that make use of the hazard equations.

¹ Socio-Economic Analysis and Modeling Division; Analysis and Development Branch; Statistics Canada

2. LONGITUDINAL LFS DATA: DISTINGUISHING FEATURES AND VALIDATION

The longitudinal version of the LFS data was constructed by splicing together the monthly records of individual respondents. An LFS respondent normally remains in the LFS sample for six consecutive months. Thus, we obtain six-month histories that are not, by themselves, long enough for most longitudinal analyses. However, the LFS design with its overlapping rotation groups allows the combination of these six-month fragments, which may then be used in analysis of employment cohort experience over decades. Note that, analysis over such long periods may require some consideration of the impact that migration has on the cohort being studied.

Figure 1: Illustration of LFS fragmentary data on cohort starting jobs in January 1976

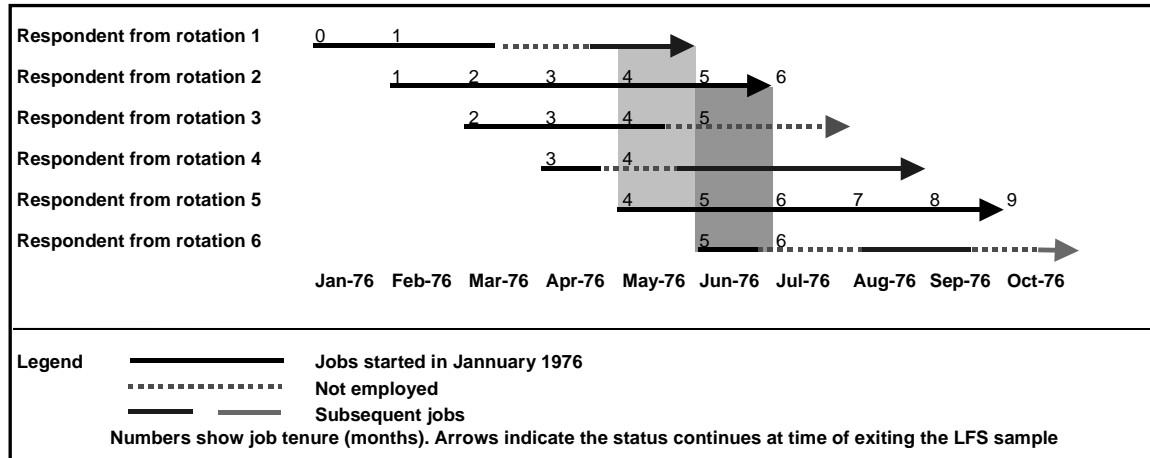


Figure 1 illustrates some characteristics of the LFS data after they are formed into longitudinal records, focusing on changes in the labour force status for the cohort who started a job in January 1976. Respondents who were members of this cohort and who entered the sample through rotation 1 contribute data on the first 6 months, from January 1976, when the job started, to June 1976, when they left the LFS sample. For respondents from rotation 2, the six-month longitudinal data windows shifts right one month (starting and ending one month later than those given by rotation 1). The data windows for respondents in subsequent rotations evolve similarly. These 6-month fragments of longitudinal data can be combined to provide complete information on an employment cohort and to identify new cohorts that may be defined in terms either of a new job or of a period without employment.

This combination of six-month fragments uses successive samples to represent a cohort's complete experience as it evolves over time. Over the long term, many different samples of individuals represent the same cohort at different points in time. However, month-to-month changes are observed largely for the same sample individuals. The two shaded areas in Figure 1 illustrate this. The respondents from each of the rotations 2-5 contribute data for both the May-June and the June-July intervals.

This is not the first attempt to use LFS data longitudinally. Lemaître (1988) studied errors in the estimation of 'gross flows' between labour force states (employed, unemployed and not in the labour force) over intervals of one month. He found that problems arose both because of response errors and because "Labour Force Survey concepts, designed for cross-sectional purposes, tend to "create" flows when consecutive months' responses are linked". (Examples include the treatment of on-call workers and of the self-employed without a business). Nevertheless, he concluded, "Administrative data have shown that not all sub-groups of status changers are seriously overestimated". Kinack (1991) examined the longitudinal consistency of responses to questions on job search activity that were employed in distinguishing between the categories unemployed and not in the labour force. He found substantial inconsistency, particularly when associated with proxy responses from different proxies. These studies have shown that focusing on transitions between the categories 'employed' and 'not employed' (i.e., without distinguishing between unemployed and not in the labour force) serves to reduce the impact of response error.

Cross-sectional data have previously been used to estimate frequencies of job hiring and job separation over monthly intervals (Lemaître, Picot and Murray, 1992). In the latter case, hiring was directly observed from the

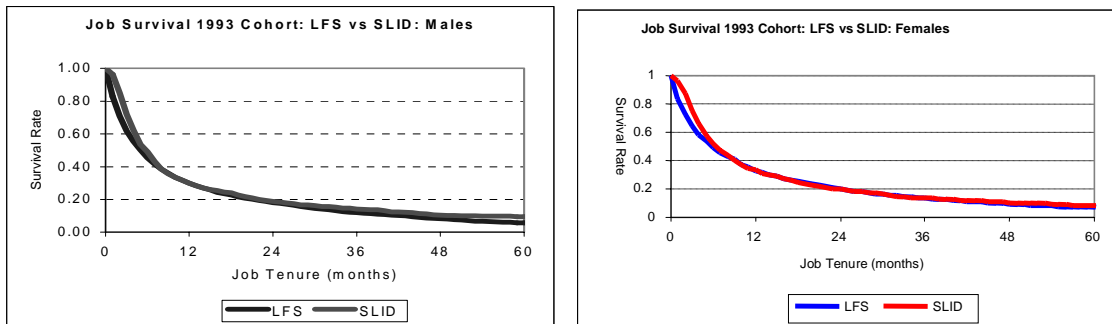
frequency of reported job-tenures of one month or less, while separation was determined residually using aggregate estimates of employment change and of hiring. Cross-sectional LFS data have also been used to calculate and compare duration statistics for synthetic-cohorts (Corak and Heisz (1995)). Those statistics were derived from retention rates obtained using, for example, the numbers of employed LFS respondents reporting job tenure 't' in month 'm' and of those reporting tenure 't+1' the next month. Such use of cross-sectional data has certain limitations. First, because the movement of individuals cannot be identified, destination states are unknown. (Although we may know the proportion that separated from a job, we do not know whether they became unemployed or began another job immediately). Second, the synthetic-cohort approach uses retention rates from a single time period to represent a complete cohort experience. Nevertheless, a time series of synthetic-cohort statistics – for example, the proportions of jobs that might last a given duration – may serve as an index reflecting changing labour market conditions.

Data Validation: Selected Examples

A second illustration of longitudinal use of LFS data involves month-to-month comparison of the number of children aged less than one year reported by female economic family heads or by the spouse of a male head. A new infant child that is reported by a woman aged between 15 and 50 likely indicates the birth of a child. In order to make direct comparisons between these LFS estimates and vital statistics, we made adjustments for births to other women living in economic families (i.e., teen lone parents living with their parents) and for births in the Yukon, NWT and Nunavut.

The two examples above indicate that – with careful attention to survey coverage, survey concepts and the possibility of response error – the LFS can provide useful longitudinal data. Figure 4 goes further in the validation of employment dynamics, comparing job ‘survival’ probabilities for males and females who started a job in 1993, as estimated from the LFS data and from SLID.

Figure 4: Comparing Job ‘Survival’ Probabilities: estimates from LFS and SLID



The survival probabilities were estimated from LFS data by the chained product of average retention rates derived from monthly main-job separation rates over the period 1993 to 1998. Survival probabilities from the SLID data were estimated in a similar manner using the reported job tenure and dates of job end. Both survival curves display the same characteristic shape; showing relatively high attrition for jobs of duration less than a year, but with much lower attrition at job tenures of one to five years. There are discrepancies between the estimates for durations of about six months or less, which may be related to the one-year recall period of SLID interviews and to the restriction of LFS job-tenure data to main-jobs. However, over periods as long as five years, the LFS and SLID provide quantitatively similar estimates. With the available LFS data, we can track some job cohorts for as long as 25 years after the job spell began.

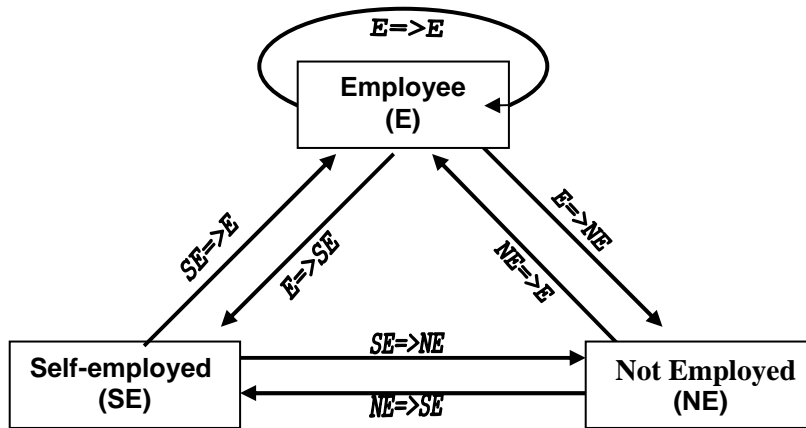
3. USING LONGITUDINAL LFS DATA FOR MODELING EMPLOYMENT ACTIVITY IN LIFE PATHS

This section focuses on the use of the LFS data to simulate employment activity in LifePaths. Currently, LifePaths uses a 3-category classification of employment status – employee (E), self-employed (SE), and not employed (NE). We have not analyzed transitions involving unemployment. Unemployment is a complex state requiring additional questions to ascertain and so, as noted above, unemployment transitions are particularly subject to response error.

There are six transitions that can result in a change of the employment status (as represented in Figure 6). LifePaths models all of these transitions. In addition, job changes that do not involve interruption of employment are also modeled by LifePaths (denoted here as E=>E). The LFS data was used to estimate hazard equations for each of these seven transitions. The estimated coefficients became parameters in the LifePaths ‘Career Work’ module. Below we discuss some technical issues that arise due to the limitations of the LFS data, followed by an illustration of the estimation results and then of a simulation outcome.

The fragmentary nature of these data poses a challenge for analysis. An important question is whether there are unavoidable biases that result from their fragmentary nature. In general, the answer is that the limitations of these data can be accounted for and potential sources of bias can be avoided with careful analysis.

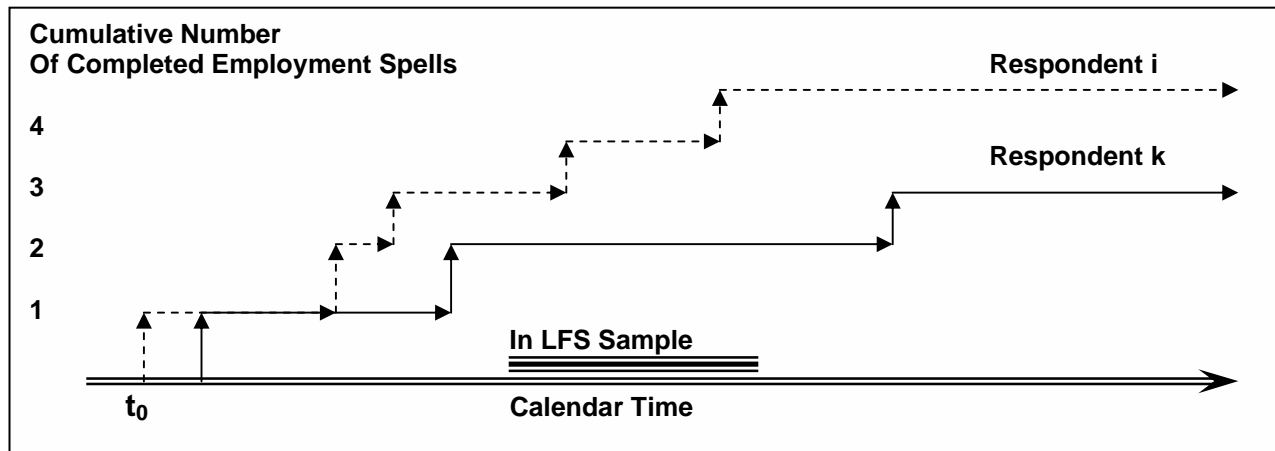
Figure 6: Employment Status and Transitions in LifePaths



Problems of Censoring and Truncation

An initial concern for the analyst of these data is the absence of retrospective employment information other than the length of the current employment spell. We might think of individual employment histories as consisting of a succession of contingent employment states (Figure 7) reflecting the process of career development. But, given only the transitions within the LFS window, estimable transition rates will inevitably involve pooling data from respondents having markedly different prior careers. In contrast, panel surveys like SLID, collect retrospective data at the first interview that, although limited, permits experience rating in terms of previous extended work interruptions or periods of part-time work.

Figure 7: Recurrent Events and Employment Spell Durations Observable within the LFS Sample Window



As also illustrated in Figure 7, LFS employment spell durations may be right-censored and/or left-truncated. Right-censoring refers to the circumstance in which a spell ceases to be observed or a respondent ceases to be at risk without an event occurring of the type being studied. This happens: either (1) because the respondent's household 'rotated out' of the LFS sample before any event occurred or (2) because another event occurred that was not of the type under active study. Similarly, these data are frequently left-truncated. This refers to the circumstance in which the beginning of a spell is unobserved, because it happened before the respondent 'rotated in' to the LFS sample (however we know the elapsed duration at the time of the first interview). Censoring usually occurs when a household is 'rotated out' of the LFS sample, on schedule, six months after being 'rotated in'. Since both truncation

and censoring are generally independent of employment event processes, neither should lead to bias in the estimation of transition probabilities, if properly accounted for.

The combination of full and partial information provided by left-truncated and right-censored data can be represented in a conditional likelihood (L_{ijt}) (Wang, 1991). For respondent i and event type j , this likelihood may be expressed in terms of the spell duration at calendar time t (i.e., denoted $m_{i,j,t}$). Terms in the likelihood involve: the probability density $f(m_{i,j,t})$ (or the corresponding cumulative probability $F(m_{i,j,t})$); a binary variable ($C_{i,j,t}$) indicating whether or not censoring has occurred; and a further binary variable (LT_i) indicating whether or not the data from this respondent was left-truncated. To account for left truncation, the likelihood is expressed in terms of conditional probabilities given the spell duration first observed (m_t): conditional probabilities evaluated at the duration of an observed event ($f(m_{t+k}|m_t)$) and conditional probabilities of surviving to an observed duration ($1-F(m_{t+k}|m_t)$).

$$\begin{aligned} L_{i,j,t+k} &= f(m_{j,t+k} | m_{j,t})^{1-C_{i,t+k}} (1-F(m_{j,t+k} | m_{j,t}))^{C_{i,t+k}} \\ &= \frac{f(m_{j,t+k})^{1-C_{i,t+k}} (1-F(m_{j,t+k}))^{C_{i,t+k}}}{(1-F(m_{j,t}))^{LT_i}} \end{aligned} \quad (1)$$

The conditional likelihood (1) can be approximated by a Poisson likelihood (Holford, 1980; Laird and Olivier, 1981), thereby also acknowledging the discreteness of the data (i.e., events are generally observed in the one month interval between successive interviews). Equation (1) can be re-expressed in terms of a binary variable $Y_{i,j,t+k}$ that represents occurrence or non-occurrence of an event in a particular time interval (i.e., $Y_{i,j,t+k} = 1 - C_{i,t+k}$). $Y_{i,j,t+k}$ is treated as a Poisson random variable having an expected value equal to a piecewise constant hazard $h_{i,j,t+k}$. Under this model, the log-likelihood contribution from respondent i over n periods (using the identity $h(m) = f(m) / (1-F(m)) = -\partial \ln(1-F(m)) / \partial m$) is:

$$\ln(L_{i,j}) \approx \sum_{k=1}^n \left[Y_{j,t+k} \ln(\hat{h}_{j,t+k}) - \hat{h}_{j,t+k} \right] \quad (2)$$

The full-sample, conditional log-likelihood for event j may be expressed as a weighted deviance, with W derived from survey weight (since events typically occur between interviews, we use averages of consecutive cross-sectional survey weights to obtain W):

$$D_j \approx -2 \left(\sum_i \left[\sum_{k=1}^{n_i} W_{j,t+k} Y_{j,t+k} \ln(\hat{h}_{j,t+k}) \right] + \sum_i \left[\sum_{k=1}^{n_i} W_{j,t+k} [Y_{j,t+k} - \hat{h}_{j,t+k}] \right] \right) \quad (3)$$

In the analysis of event type j , we treat other events as censoring (i.e., non- j events occurring to the same population-at-risk), and therefore the deviance for sets of events will be the sum of component deviances (i.e., if the overall hazard is the sum of competing hazards, then the competing risks may be treated as independent (Prentice et al., 1978)).

A more direct motivation of the same deviance has Poisson processes as the starting point (Borgan, 1984; Andersen, 1985; Andersen and Borgan, 1985; Lawless, 1987), rather than starting with postulated duration densities (durations which must be latent in a multivariate situation (Prentice et al., 1978)). Here, we would model sample multivariate counting processes that represent the number of occurrences of each specific event in time intervals $[t_0, t]$. Sample counting processes, represented by the step functions in Figure 7, are observable counterparts of cumulative hazard functions. The assumption that underlying hazard functions are approximately piecewise constant leads to the Poisson deviance as an approximation (Lindsey, 1995). To avoid bias, the principal concerns are that the full population-at-risk is identified, that censoring or truncation mechanisms are conditionally independent of the underlying employment processes and that the intervals over which hazards are assumed constant are not too large.

It is possible to obtain non-parametric estimates of employment hazard functions by implicitly splicing together all available information on members of a defined cohort from the longitudinal LFS samples. That would be a relatively simple problem compared with the complex observation schemes considered by Alioum and Commenges (1996). The implicit splicing of information takes place in the deviance (3) which has two components: the first component is non-zero only at observed events, while the second component represents cumulative hazards experienced over all times prior to those events or to censoring times. To the extent that the LFS cross-sections are representative samples for each reference week, then – taken together – they will provide an accurate estimate of the numbers of events occurring over the ‘life’ of an employment cohort (i.e., identified by a common date of job gain or job loss). Similarly, within employment cohorts, we can expect to find appropriate numbers of sampled left-truncated and right-censored respondent spells that correspond in time to the missing histories of all those left-truncated spells that terminate with an event. As such, the first component of the deviance will accurately reflect whether hazard estimates tend to be large over periods where observed events are frequent. And the second component, summed over all respondent-months, will produce a sample estimate of the cumulative hazard similar to that which we might have obtained had there been no left-truncation. So, for these data, the conditional deviance may be close to an unconditional deviance.

Hazard Equation Estimation

Patterns of employment transition differ significantly among different demographic groups. For example, full-time students are most active in the labour market during their summer break, while the maternity leave that an employed pregnant woman takes may be largely determined by Employment Insurance regulations. Accordingly, LifePaths distinguishes among the following groups and models their employment activities separately:

- Those who are full-time students;
- Those who have just graduated or left school and are in a transition to a first after-school job;
- Pregnant women for whom a maternity-leave may apply;
- Those who are in prime ages of employment; and
- Older workers in transition to retirement.

We discuss here only the estimation for the fourth group, which includes individuals in their “career employment” phase, the most important phase in terms of impact on the economy. Particulars for the other groups are available from the Statistics Canada website noted above.

Our parametric hazard estimation uses a log-linear form of regression equation, one equation for each of 7 transitions and for each sex separately, giving a total of 14 equations grouped in three sets of competing risks:

$$E(Y_{j,t+k}) \approx \hat{h}_{j,t+k} = \exp(\hat{g}(m_{j,t+k}) + X_{j,t+k} \hat{\beta}) \quad (4)$$

where $E(Y)$ is an expectation, $g(m)$ is a log-linear spell duration spline, X is a vector of time-varying covariates and β is a vector of regression coefficients. The term $g(m)$ corresponds to a piecewise Weibull baseline hazard, which distinguishes between employment transition risks at durations of less than a year from those at durations of more than a year. The covariates (X) include variables representing age, education, province of residence, presence of children by age group, spouse’s employment status, calendar month and calendar year, as well as interactions between some of these factors. Estimates of β are found which minimize the deviance (3).

The only example of detailed results that we present here involves the mutual influence of husband’s and wife’s employment status on each other’s respective transition hazards. Figure 8 compares coefficient estimates across the seven equations that represent the seven transitions we specified. The two panels correspond to separate sets of equations for males and females. The category ‘no spouse present’ was treated as the reference category and the spouse’s employment status was classified into ‘with paid employment’, ‘self-employed’, and ‘not employed’. The estimated coefficients are presented here in terms of risk relative to the reference group. Thus, with other covariates controlled, the hazard of becoming self-employed for female employees whose husbands are self-employed is about 2.5 times (250%) higher than the hazard of their counterparts who do not have a spouse (see tallest bar in the top panel).

Figure 8: Impact of Spouse's Employment Status on Employment Transition Risks



Figure 8 shows that the very presence of a spouse can work in opposite directions for males and females. The most frequent transitions for both sexes are E=>E, NE=>E and E=>NE. For females, the first two of those transitions are less likely to occur to married women than to single women, while the transition to 'not employed' is more likely. (The presence of children is not the reason for this, as children have been controlled for.) For males, the pattern is reversed. The results for these transitions are consistent with conventional gender roles. However, given the magnitudes of these relative risks, we are not given the impression that gender roles have a strong influence (after other variables are controlled for).

Figure 8 also reveals other conspicuous patterns. First, the relative risks of a transition into self-employment, for spouses with husbands/wives in self-employment, stand out as the highest among all other transitions. In addition, spouses with husbands/wives in self-employment have the lowest relative risks of a transition out of self-employment. Thus, self-employment status is, in a sense, mutually reinforcing within families. These observations are consistent with forms of self-employment that involve a family business and also with endogamy among professionals (e.g., lawyers marrying other lawyers).

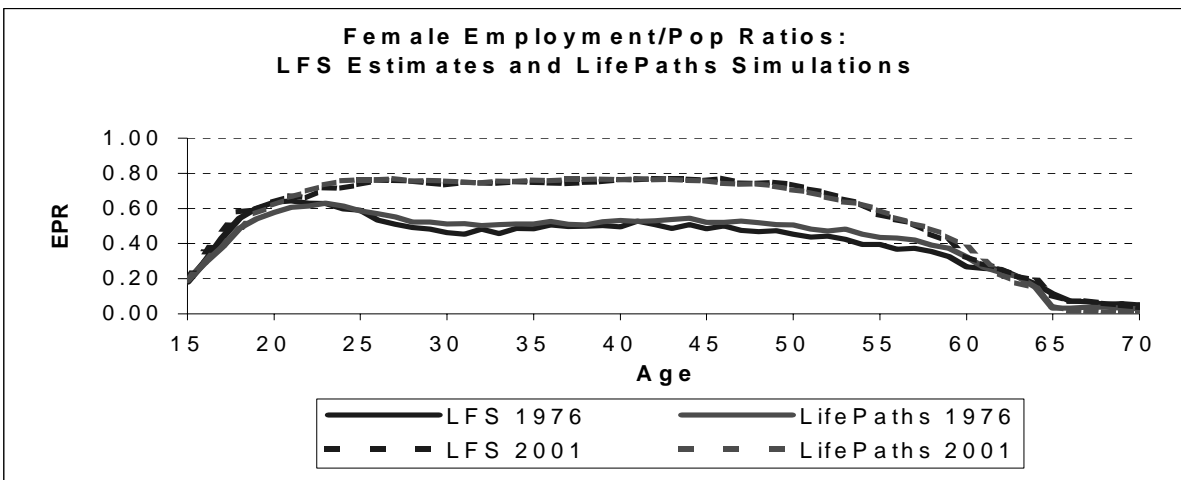
Other results (not reported here) confirm the influence of a range of explanatory variables on individual's employment transitions. These covariates include age, job tenure (or duration not employed), educational attainment, presence of young children (especially for women), province of residence, seasonality, and business cycles.

From Estimated Parameters to the Simulation Results: One Illustration

Our example of the role of spouse's employment status points to the need for family context in the simulation of employment activities. It is a challenge for LifePaths to integrate these relationships into the simulation process. For example, if individual education progression or the effects of education on employment transitions are not modeled appropriately and accurately, then the consequences will cascade from direct education-employment relationships to a chain of indirect impacts, including relationships between education and marriage, fertility, interprovincial migration, etc. The impact will then spill over to the simulated spouse, as indicated above. It is not difficult to see that, unless these relationships are specified appropriately and the parameters are estimated with reasonable accuracy, bias would be spread over a wide range of simulated outcomes.

An overall validation of the LifePaths employment hazard equations was obtained by comparing simulated annual average employment/population ratios with direct cross-sectional estimates from the LFS. The simulated employment/population ratios (EPR) were from a synthetic population whose members were subject to several of the seven types of employment hazards over the course of each simulated year. Moreover, the simulations also involved generating appropriate distributions of covariates that in turn determine the distributions of employment transition hazards. As may be seen in Figure 9, LifePaths accurately reflects the age patterns of female employment in both 1976 and 2001 and correspondingly accounts for the dramatic change observed in those age patterns over the past quarter century.

Figure 9: Validating the hazard equations using LifePaths



4. CONCLUSIONS

We have demonstrated that the LFS data – when organized into the fragmentary event histories collected over the six-month periods that most respondents spend in the sample – represents a significant longitudinal data asset. There is sufficient sample and breadth of content to provide for important analysis of labour market dynamics and, conceivably, of demographic processes such as fertility. Moreover, the data is monthly and spans more than a quarter century, so that analysis based on it has uninterrupted time depth that is unique in Canada.

Much remains to be done. Future work will involve extending and refining our models. In addition, we need to make further progress in variance estimation. So far, we have only preliminary estimates of the design effects on the variances of coefficients in our hazard equation (i.e., in the range of 1.1 to 1.2). To date, with this line of research still in its initial stages, our approach to inference has been informal.

REFERENCES

- Alioum, Ahmadou and Daniel Commenges (1996). "A Proportional Hazards Model for Arbitrarily Censored and Truncated Data". *Biometrics*, 52, 512-524.
- Andersen, Per Kragh (1985). "Statistical models for longitudinal labor market data based on counting processes" in *Longitudinal Analysis of Labor Market Data*. James J. Heckman and Burton Singer (eds). Cambridge University Press, Cambridge.
- Andersen, Per Kragh and Ørnulf Borgan (1985). "Counting Process Models for Life History Data: A Review". *Scand.J.Statist.*, 12, 97-158.
- Borgan, Ørnulf (1984). "Maximum Likelihood Estimation in Parametric Counting Process Models, with Applications to Censored Failure Time Data". *Scand.J.Statist.*, 11, 1-16.
- Corak, Miles and Andrew Heisz (1995). "The Duration of Unemployment: A User Guide". Research Paper Series No. 84, Analytical Studies Branch, Statistics Canada.
- Holford, T.R. (1980). "The analysis of rates and survivorship using log-linear models". *Biometrics*, 36, 299-305.
- Kinack, Mark (1991). "Measuring Data Quality with Longitudinal Data". 1991 Proceedings of the American Statistical Association Section on Survey Research Methods, 514-519.
- Laird, Nan and David Olivier (1981). "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques". *JASA*, 76, 231-240.
- Lawless, J.F. (1987). "Regression Methods for Poisson Process Data". *JASA*, 82, 808-815.
- Lemaître, Georges (1988). "The Measurement and Analysis of Gross Flows". Staff Reports, Labour and Household Surveys Analysis Division, Statistics Canada.
- Lemaître, Georges, Garnett Picot and Scott Murray (1992). "Workers on the move: An overview of labour turnover". *Perspectives on Labour and Income*, 4(2), Statistics Canada.
- Lindsey, J.K. (1995). "Fitting Parametric Counting Processes by using Log-linear Models". *Appl.Statist.*, 44, 201-212.
- Prentice, R.L., J.D. Kalbfleisch, A.V. Peterson Jr., N. Flournoy, V.T. Farewell and N.E. Breslow (1978). "The Analysis of Failure Times in the Presence of Competing Risks". *Biometrics*, 34, 541-554.
- Statistics Canada (1998). "Permanent Layoffs, Quits and Hirings in the Canadian Economy: 1978-1995". Catalogue # 71-539-XIB.
- Statistics Canada (1998). "Overview of the Survey of Income and Labour Dynamics". Catalogue # 75F0011XPB, <http://www.statcan.ca/english/freepub/75F0011XIE/free.htm>.
- Statistics Canada (2001). "The LifePaths Microsimulation Model: An Overview". http://statcan.ca/english/spsd/LifePathsOverview_E.pdf
- Statistics Canada (2002). "Guide to the Labour Force Survey". Catalogue # 71-543-GIE, <http://statcan.ca/english/freepub/71-543-GIE/0000071-543-GIE.pdf>.
- Wang, Mei-Cheng (1991). "Nonparametric Estimation from Cross-sectional Survival data". *JASA* 86, pp.130-143.