DÉPARTEMENT DE SCIENCE ÉCONOMIQUE
DEPARTMENT OF ECONOMICS

CAHIERS DE RECHERCHE / WORKING PAPERS

 # 0703E

by

Estimating Labour Market Transitions and Continuations using Repeated Cross Sectional Data

Pierre Brochu*

University of Ottawa

August 2007

ISSN: 0225-3860

uOttawa

Faculté des sciences sociales
Faculty of Social Sciences

CP 450 SUCC. A
OTTAWA (ONTARIO)
CANADA K1N 6N5

P.O. BOX 450 STN. A
OTTAWA, ONTARIO
CANADA K1N 6N5

* Department of Economics, University of Ottawa, 55 Laurier Street East, Desmarais Building, Ottawa (Ontario) K1N 6N5, Canada. Phone: (613) 562-5800 ext. 1424. Fax: (613) 562-5999. Email: pbrochu@uottawa.ca. I thank Aneta Bonikowska, David Gray, David Green, Thomas Lemieux, Kevin Milligan and Joris Pinkse for their helpful comments.

1 Introduction

There is a long tradition of exploring labour market transitions in economics. Although the unemployment-employment transition has been the most frequently explored, other transitions or continuations have also been examined, such as the transition out of the labour force (e.g. Jones and Riddell (1999)) and the continuation of a job (job stability, e.g. Swinnerton and Wial (1995); Neumark, Polsky, and Hansen (1999)). While using panel data to estimate these labour market outcomes is the preferred approach, there are circumstances where that approach is problematic. For example, limited historical coverage (Canadian panels) make it impossible to differentiate between cyclical and secular changes in job stability.[1] With the absence of this differentiation, one cannot address the real question of interest in the job stability literature: i.e. how and why has job stability changed? The unemployment-employment transition literature also faces panel data limitations. Panel data from the Spanish Labour Force Survey, for example, only became available after significant labour reforms, i.e. the introduction of fixed term contracts in Spain. As a result, Güell and Hu (2006) could not rely on panel data to examine how (important) changes in fixed costs faced by Spanish firms have affected the probability of leaving unemployment. In these instances, repeated cross sectional data sets offer a valid alternative.

In this paper, I propose a new non-parametric approach for estimating the continuation probability when using cross sectional data.[2] This approach takes advantage of the fact that repeated cross sectional data sets, like the Current Population Survey (U.S.) and the Labour Force Survey (Canada), are representative of the country's population. While the existing cross sectional literature applies panel tools to cross sectional data; those tools are designed to take advantage of the fact that in longitudinal data one can follow individuals over time—something not possible with

[1] See Brochu (2006) for a detailed discussion of panel data limitations when exploring for changes in job stability in North America.

[2] In a two state world, the continuation probability is simply one minus the transition probability.

repeated cross sections. As a result, existing cross sectional approaches have had to impose very strict identifying assumptions and make unverifiable claims that their approaches provide reasonable approximations.

The key insight of this paper is that the population continuation (or transition) probability can be written as a ratio of two unconditional means; a conceptual framework that is conducive to cross sectional analysis. Based on this result, I propose a cross sectional estimator for the continuation probability, and another for the standard errors. The required identifying assumptions are both relatively mild and easy to interpret.

Using the proposed framework, I also re-examine existing methods. Through this lens, I clearly identify the full set of underlying assumptions of the traditional non-parametric estimator and provide a consistent estimator for its standard errors—both of which fill gaps in the literature. I also illustrate the bias of the Güell and Hu (2006) method, and suggest a potential correction to their parametric method.

Finally, a researcher may prefer the use of repeated cross sections even when panel data is available. Repeated cross sections like the CPS and LFS have much larger sample sizes that allow for a more detailed analysis (e.g. Baker (1992)); one can focus on more narrowly defined groups without having to worry as much about small samples. Depending on the question (and group) of interest, panel attrition may become a significant problem. One can, for example, expect younger and more mobile workers to have higher attrition rates. For these groups, a self-selection bias may arise. The proposed estimator, however, does not face this difficulty—it is designed for cases where the same individual cannot be followed over time.

The remaining sections of this paper are divided as follows: Section 2 provides a discussion of standard cross sectional approaches and demonstrates their weaknesses; Section 3 proposes an alternative non-parametric method; Section 4 re-interprets the existing methods using the proposed framework; Section 5 provides some empirical examples; Section 6 illustrates the bias of a recently proposed parametric estimator; and finally, Section 7 provides some final remarks.

2 Existing Non-Parametric Approaches

This section provides an overview of the standard cross sectional methods. Given the well documented difficulties associated with estimating transition (or continuation) probabilities using only one cross section,[3] I focus on methods that require the use of two cross sections covering two consecutive time periods. Finally, to simplify the presentation, all estimators are applied in order to estimate the employment-continuation probability—the probability that the worker remains with the same employer. The framework can easily be generalized to explore any transition or continuation of interest.

The standard cross sectional approach is to estimate the continuation probability by subpopulation. Assume the researcher is interested in the employment-continuation probability of an "at-risk group", say, individuals with time-invariant characteristics c who have been employed for s periods at time 1. The continuation probability is simply the fraction of "at-risk" individuals in the *population* that remains with the same employer in the next period

$$\frac{N_2^{s+1,c}}{N_1^{s,c}} \tag{1}$$

where $N_1^{s,c}$ is the number of people in the population that have time-invariant characteristics c who have been unemployed for s periods at time 1.

A natural next step would be to use the sample analog of equation (1), i.e. the fraction of "at-risk" individuals in the *sample* who remain with the same employer in the next year. Yet this is not possible with cross sectional data—one cannot follow individuals over time. Instead, the cross sectional literature (e.g. Baker (1992); Neumark, Polsky, and Hansen (1999)) proposes a synthetic cohort approach that takes advantage of the fact that base weights of representative cross sections, like the Labour Force Survey (Canada) and the Current Population Survey (US), sum

[3] This approach requires constant inflows over time, an assumption not supported in the data. See Ureta (1992) for more details.

up to their respective populations.[4] The cross sectional estimator takes the form

$$\frac{\tilde{n}_2^{s+1,c}}{\tilde{n}_1^{s,c}} \tag{2}$$

where $\tilde{n}_1^{s,c}$ is the sum of the base weights of all individuals with characteristics c who have been employed s periods with the same employer as of period 1. By using weights as counts, the denominator (numerator) of equation (2) directly estimates the denominator (numerator) of equation (1). This result holds because the base weights sum up to the country's target population. Estimating population counts is not a common approach in empirical work; a more traditional approach would be to estimate population moments (e.g. means).

The approach is synthetic in the sense that one does not follow the same period 1 individuals over time, but instead uses a group with similar characteristics in period 2 to help estimate the continuation probability. The use of synthetic cohorts is not new to empirical work. What is unique to the present literature is its attempt to estimate a transition or continuation probability using synthetic cohorts.

Because it directly estimates population counts, the cross sectional estimator in equation (2) abstracts from the difficulty associated with sample sizes changing over time. An example will best illustrate this point. Consider the case where the population is unchanged from period 1 to period 2, but where the period 2 sample is relatively smaller. To ensure that the base weights sum up to the country's population, the data agency scales up the base weight of the period 2 cross section. As a result, one can still get an "unbiased" estimate of the population counts in equation (2), even if the sample size changes over time. This would not be the case if un-weighted counts were used. A smaller period 2 sample would reduce the estimate in the numerator.

The strength of the existing approach is also its weakness; directly estimating population counts may be intuitive, but it lacks statistical vigor. It is reasonable to

[4]For example, an observation with a base weight of 5 is meant to represent 5 people in the population.

think that the accuracy of the estimator will improve with larger samples, but this cannot be proven in any statistical sense.[5] Most importantly, one cannot lay bare all underlying identifying assumptions without such a proof. A clear understanding of the consistency requirements is critical considering that an important selling point of cross sectional data sets like the CPS and LFS have been their large sample sizes.

Most importantly, the lack of precision carries over to the inference stage. Given the functional form of the estimator, there is no standard way to construct standard errors. The literature has used two approaches—both of which require very strict identifying assumptions. The first approach (e.g. Diebold, Neumark, and Polsky (1997); Swinnerton and Wial (1995)) applies an estimator designed for panel data to the cross section case. Yet, this estimator fails to account for the variability resulting from an inability to follow individuals over time. The second approach proposed by Neumark, Polsky, and Hansen (2000) treats the numerator of equation (2) as a random variable,[6] but the denominator as a constant. Different draws from the year 1 population would generate different values for the denominator, and therefore, it should also be treated as random. As such, this method also fails to account for the full variability of the cross sectional approach. In Section 5 I show that both approaches have a downward bias—a bias that may lead to a spurious identification of a change in continuation probability.

3 Proposed Method

In this section I propose a new non-parametric approach for estimating the continuation probability using cross sectional data. As with other non-parametric estimators, the proposed method is fully flexible—it does not impose any structure on how one

[5] Increasing the sample size will have two countervailing effects: The number of observations in both sums will increase, but the base weight associated with each observation will fall. Abstracting from the weight issue will not resolve the problem. In such a case, the numerator (and denominator) would not converge to a point, but instead go to infinity.

[6] The numerator is presented as the sum of a binary variable which equals one if individual i has characteristics c and has been unemployed for s periods in period 2, and zero otherwise. Their period 2 sample consists of all unemployed workers in period 2. They also abstract from the weight issue by using only un-weighted counts in the calculation of standard errors.

group's continuation probability relates to that of other groups. But contrary to other cross sectional methods in the literature, it does not approximate a panel approach. The proposed method is designed with cross sectional data in mind, and as a result does not face the same difficulties at the inference stage.

The approach to econometric modeling focusses on the population and its moments. Assume a population distribution $F(X_1, X_2)$ where X_j is a vector of characteristics for period j such as age, gender, and in-progress unemployment duration. This representation implies that for each individual in the population, there are two periods of information. As such, the object of interest - the continuation probability - is simply a "population cohort" parameter.

Let $Z_{ij}^{s,c}$ be a dummy variable which takes the value one if individual i in period j has characteristics c, and has been employed with the same employer for exactly s periods. Use the shorthand notation $D_{ij} = Z_{ij}^{s,c}$ and $N_{ij} = Z_{ij}^{s+1,c}$. The continuation probability for a randomly chosen individual with characteristics c who has been unemployed for s periods, $R_1^{s,c}$, is commonly expressed as

$$R_1 = R_1^{s,c} = Prob(N_{i2} = 1 | D_{i1} = 1) \tag{3}$$

Given its conditional structure, equation (3) is a good starting point for panel data. One can condition on an individual working in the first period of a panel, and therefore estimate the sample analog of equation (3). With repeated cross sections, however, this is not possible. In Proposition 1, I propose an alternative representation.

Proposition 1 *The continuation probability of a population cohort can be expressed as*

$$R_1 = \frac{E(N_{i2})}{E(D_{i1})} \tag{4}$$

proof: See Appendix A.1 ∎

Proposition 1 is a key insight of this paper. It shows that the population continuation probability can be written as a ratio of two unconditional moments. Most importantly, the numerator in equation (4) does not condition on period 1 events. As such, this representation is more conducive to cross sectional analysis than existing approaches which rely on equation (3) as their starting point.

The proposed estimator is the sample analog of equation (4)

$$\hat{R}_1 = \hat{R}_1^{s,c} = \frac{\sum_{i=1}^{n_2} N_{i2}/n_2}{\sum_{i=1}^{n_1} D_{i1}/n_1} \tag{5}$$

where n_j is the size of the year j representative sample.

Proposition 2 shows that the continuation probability can be consistently estimated using two cross sections. Under clearly stated identifying assumptions, one can recover this probability without a panel data set.

Proposition 2 *Assuming iid samples for each year that are drawn from a population cohort, then* $\hat{R}_1 \rightarrow R_1$

proof: See Appendix A.2 ∎

The identifying assumption of drawing from a population cohort necessitates further attention. It requires that the two cross sections select vectors of observable characteristics from the same pool of individuals, but at different moments in time. It allows for the fact that the two cross sections do not select the same individual, and that the sample sizes may differ.

Sampling from a population cohort does require that the probability of an individual being drawn not vary across time. This rules out sample restrictions based on time varying characteristics. For example, restricting the sample to only include employed workers (as is the case in the job stability literature) will violate the identifying assumption of Proposition 2. The pool of employed workers in the population does not remain constant over time—some individuals will lose their jobs, while others will find employment. The proposed approach therefore requires that the

underlying population be broadly defined. Fortunately, repeated cross sections like the CPS and LFS are representative of their country's population.

A further threat to the validity of the proposed approach must be explored. Immigration, emigration and deaths will affect the country's population. However, the CPS and LFS are carried out at frequent intervals - on a monthly basis. As a result, slippage (changes in population) due to immigration, emigration or deaths will be minimal. In Section 4, I show how the proposed estimator can be adjusted when the cross sections are further apart.

In Proposition 3, I provide the asymptotic properties of the estimator. I allow for both N_{ij} and D_{ij} to have sampling distributions. No restrictions are imposed on the correlation within observations, i.e. between D_{ij} and N_{ij}, but independence across observations is assumed. For repeated cross sectional data, this implies no correlation across time.

Proposition 3 *Assuming iid samples for each year that are drawn from a population cohort, independence across years, and* $lim_{n_1,n_2\to\infty}\frac{n_1}{n_2} = 1$, *then* $\sqrt{n}_1(\hat{R}_1 - R_1) \xrightarrow{d} N(0, V)$ *where* V *is*

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) \tag{6}$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})} \tag{7}$$

proof: See Appendix A.3 ∎

As equation (6) illustrates, V is simply a weighted sum of the variance of D_{i1} and N_{i2}, with the weights reflecting the non-linearity of the estimator. Replacing the population moments in equation (6) with corresponding sample analogs generates a consistent estimator of the asymptotic variance.

Finally, applying this approach to survey samples where the probability of being

selected is not the same across observations is straightforward. For the continuation probability estimator, one replaces the population means in equation (4) with the weighted sample analogs

$$\tilde{R}_1 = \frac{\sum_{i=1}^{n_2} nw_{i2} N_{i2}/n_2}{\sum_{i=1}^{n_1} nw_{i1} D_{i1}/n_1} \tag{8}$$

where nw_{ij} is the normalized weight for individual i in year j, i.e. nw_{ij} sum to n_j. A similar procedure is used for the variance estimator. One replaces the population means in equation (6) with the weighted sample analog—again using normalized weights. The use of normalized weights reflects the more traditional use of weights; the weights are only used to reflect the varying probability of selection and not used as population counts.

4 Links to the Existing Method

I start this section by adapting Proposition 1 - the ability to present the continuation probability as a ratio of two means - in order to account for compositional changes in the population. I then show that one can re-interpret the existing continuation probability estimator as an extension of the one proposed. As a result, I clearly identify the underlying assumptions of the existing estimator and provide a consistent approach to estimating its standard errors. These areas are yet to be fully examined in the literature.

Proposition 4 shows that the continuation probability can be written as a function of two unconditional moments, when the composition of a country's population changes over time.

Proposition 4 *Assume that the composition of a country's population changes from period 1 to period 2. Further assume that these compositional changes break (or interrupt) the spell of interest. The continuation probability for the period 1 population*

can be expressed as

$$R_1 = \frac{adj_1 E(N_{i2})}{E(D_{i1})} \tag{9}$$

where adj_1 is the population growth (or adjustment) factor.[7]

proof: See Appendix A.4 ∎

Contrasting this finding with that of Proposition 1 (where the composition was assumed unchanged) highlights the necessary adjustment; one must only account for net changes in population size, i.e. adj_1. More precisely, the numerator of equation (9) recovers the counterfactual of interest—the mean value of N_{i2} if the population composition had remained unchanged in the second period.[8] However, Proposition 4 does require that the compositional change break (or interrupt) the spell of interest. For the employment-continuation probability, this assumes that the emigrant or immigrant not remain with the same employer after migration, i.e. migration must interrupt the individual's tenure spell.[9]

Proposition 5 shows that the existing cross sectional estimator, i.e. equation (2), which I now define as $\tilde{Q}_1$, can be rewritten as

$$\tilde{Q}_1 = \hat{adj}_1 \tilde{R}_1 \tag{10}$$

where $\hat{adj}_1 = \frac{\sum_{i=1}^{n_2} bw_{i2}}{\sum_{i=1}^{n_1} bw_{i1}}$.[10] The proposed and existing estimators will differ by the estimated population growth factor. In the case of growing populations (e.g. Canada and the U.S.), for example, $\tilde{Q}_1$ will provide systematically larger estimates.

Proposition 5 *Given repeated cross sections where the base weights sum up to the target population, then $\tilde{Q}_1 = \hat{adj}_1 \tilde{R}_1$.*

[7] If the population size increased by 20%, the population growth factor would be 1.2.

[8] See proof of Proposition 4 for intuition on why this is the case.

[9] I will argue later in this section that for some continuation or transition probabilities, this identifying assumption may not be innocuous.

[10] $\hat{adj}_1$ can be interpreted as an estimate of the growth factor when the sum of the base weights add up to the target population, as is the case in the CPS and Canadian LFS.

proof: See Appendix A.5 ∎

By linking the two estimators (Proposition 5) and showing how to account for compositional change (Proposition 4), one can easily derive the asymptotic properties of $\tilde{Q}_1$.

Proposition 6 *Assume iid samples for each year that are drawn from a population where compositional changes break (or interrupt) the spell of interest. Further assume independence across years,* $lim_{n_1,n_2\to\infty}\frac{n_1}{n_2} = 1$*, and where* $\hat{adj}_1$ *is the true population growth factor. Then,* $\sqrt{n}_1(\tilde{Q}_1 - \tilde{Q}_1) \xrightarrow{d} N(0, V)$ *where* V *is*

$$V = adj_1^2 \left[\phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2})\right] \tag{11}$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})}$$

proof: See Appendix A.6 ∎

The variance estimator is simply the (weighted) sample analog of the population moments in equation (44).

The two main assumptions of Proposition 6 are that 1) the population growth factor is measured without error and that 2) compositional changes break the spell of interest. I argue below that the former is the less severe of the two identifying assumptions.

$\hat{adj}_1$ estimates a broadly defined parameter - the growth factor of the country's population. Given the large samples of repeated cross sections like the LFS and the CPS, one can therefore expect $\hat{adj}_1$ to provide accurate estimates. A second difficulty is the potential correlation between $\hat{adj}_1$ and $\tilde{R}_1$—which is ruled out by assumption. The fact that they each estimate very different objects may minimize any potential correlation; the former estimates the country's population growth, while the latter estimates the continuation probability for specific groups. Having

said that, one could rule out this possibility by simply using a different data set to estimate the population growth factor.

The more serious threat is the requirement that compositional change interrupt the spell of interest—the severity of which will depend both on the continuation (or transition) probability of interest and the target population. For the transition out of a job, for example, one must assume that the change in population resulting from immigration, emigration or death, leads to a break in tenure spell. In Canada and the U.S., immigration is the biggest source of population change, and assuming that immigrants find new employment upon arrival is a reasonable assumption.[11] For the exploration of the unemployment-continuation probability, on the other hand, assuming breaks in unemployment spells when the individual migrates is more severe of an identifying assumption; particularly for countries where the emigration decision is due to lack of employment prospects in the home country.

5 Empirical Examples

In this section, I empirically compare existing and proposed approaches using two large scale data sets: the Canadian Labour Force Survey (LFS) and the U.S. Current Population Survey (CPS). This section is divided into two parts. In the first, I focus on the employment-continuation probability estimators, and rely on repeated cross sections from the LFS files to illustrate the differences. The second part focusses on the standard errors. Using the CPS, I show that the bias of existing approaches to estimating standard errors can lead the researcher to falsely reject the null hypothesis of no change (or difference) in the continuation probability.

[11] One can limit the importance of deaths as a source of population change by restricting the focus to prime-aged individuals.

5.1 Employment-Continuation Probability

Table 1 compares the 1-year employment-continuation probability estimators using the master LFS files.[12] The LFS is a large monthly household survey of approximately 54,000 households per month, with a focus on gathering information about labour market activities of Canadians. The LFS is a rich source of tenure data. As part of their regular questionnaire, respondents are asked when they started working for their present employer. In Table 1, I present estimates of the 1-year continuation probability for select groups in the year 2000.[13]

Two important conclusions can be drawn from Table 1. One, the two methods generate very similar 1-year continuation-probability estimates. As equation (9) indicates, the two estimates will only differ by a common scaling factor, i.e. the population growth factor. Table 1 shows the 2000 growth factor only slightly exceeded 1. As a robustness check, I estimated the population growth factor for each year over the 1977-2003 period. The growth factor averaged 1.0033 over this period, and was close to 1 in all years—despite the fact that the Canadian population has increased over time due to immigration. This is because one is only looking at whether an individual remains with the same employer in the next year, and not, say, 10 years from now. The second conclusion is that the standard errors are relatively small. The loss in efficiency due to the fact that one cannot follow individuals over time in a cross sectional approach is compensated by the large samples of the LFS. As a result, the the probability estimates are very precise. It should be noted that both sets of standard errors were estimated using proposed methods; Equation (6) was used to construct the standard errors of $\tilde{R}_1$, and equation (11) for the standard errors of $\tilde{Q}_1$.

[12] These files were accessed on site at the British Columbia Interuniversity Research Data Centre (BCIRDC). The BCIRDC is run and sponsored by the University of British Columbia, University of Victoria and Simon Fraser University, in collaboration with Statistics Canada.

[13] Following Brochu (2006), all continuation probabilities condition on being 20 to 54 years of age. This imposes the following sample restrictions: the 2000 sample only included those 20 to 54 years of age, while the 2001 sample was restricted to individuals 21 to 55 years of age. The continuation probabilities also conditions on not being self employed, nor a full-time student. See Brochu (2006) for more details.

5.2 Hypothesis Testing

Within a continuation probability approach, testing for differences in job stability across time or groups is straightforward—only a single restriction needs to be tested. For ease of exposition, I focus on time differences; the arguments are similar when testing across groups. The null and alternative hypotheses are

$$H_0 : R_j - R_1 = 0 \tag{12}$$

$$H_a : R_j - R_1 \neq 0 \tag{13}$$

respectively, where $R_j - R_1$ is the difference in retention rate over a $j-1$ period. I use a t-test approach. The t-statistic, t_n, is [14]

$$t_n = \frac{\hat{R}_j - \hat{R}_1}{\sqrt{\hat{V}_{R_j - R_1}/n}} \tag{14}$$

where $\hat{V}_{R_j - R_1}$ is the estimator of $Avar(\hat{R}_j - \hat{R}_1)$.

The standard errors estimator - the denominator in equation (14) - must be able to account for the possible correlation between $\hat{R}_j$ and $\hat{R}_1$. More precisely, $\hat{R}_2$ and $\hat{R}_1$ may be correlated since both the denominator of $\hat{R}_2$ and the numerator of $\hat{R}_1$ are functions of the same (year 2) observations.[15] By allowing both N_{ij} and D_{ij} to have sampling distributions, the proposed approach can easily generate the necessary covariance term. This is not the case for existing methods. The Neumark-Polsky-Hansen (NPH) method, for example, rules out the possibility of any correlation between $\hat{R}_2$ and $\hat{R}_1$ by assuming that D_{ij} is a constant.

I use CPS data to illustrate how the choice of standard errors estimator can matter at the inference stage. As with the LFS, the American CPS is a large monthly household which asks respondents about their labour market activities. But contrary

[14] To simplify the presentation I assume that the cross sections are all of size n. The asymptotic properties of the employment-continuation probability differential are left to Appendix A.7.

[15] A similar difficulty occurs when testing across groups, say, groups A and B. The numerators (and denominators) of R_j^A and R_j^B may also be correlated.

to the LFS, a tenure question is not part of the regular CPS questionnaire—it is only included in select supplements. I therefore rely on 4-year employment-continuation probabilities—as was previously done in the American job stability literature. Finally, I use the $\tilde{Q}_1$ estimator (instead of $\tilde{R}_1$) to estimate the continuation probabilities; assuming no change in the underlying population would be too restrictive an assumption.

Tables 2 and 3 examine changes in 4-year employment-continuation probabilities from 1996 to 2000 for males and females, respectively.[16] Standard errors are calculated for the NPH method, the DNP method,[17] the proposed method, and the proposed method with no covariance term. In all cases, weights were used to make a clearer comparison of the various methods. Robustness checks for different years and sub-populations indicate that weights do not significantly affect the results.

Tables 2 and 3 indicate that accounting for the covariance term can increase or decrease the standard errors. One can easily show that the covariance term will be positive if and only if [18]

$$Pr(N_{i2} = 1, D_{i2} = 1) > Pr(N_{i2} = 1)Pr(D_{i2} = 1) \tag{15}$$

For the 4-year employment-continuation probability of males in the 0-2 tenure group, for example, the probability that both D_{i2} and N_{i2} equal 1 is zero, and as such, the covariance term is negative.

Even when the covariance term is negative, a consistent pattern emerges with the proposed method generating standard errors consistently larger than either DNP or NPH methods. This pattern was found to be robust for other time periods and other sub-populations. The proposed method generates standard errors that are up

[16] All U.S. continuation probabilities condition on being at least 16 years of age, and not being self-employed.

[17] The DNP method, refers to the application of the longitudinal standard error estimator to cross sectional data. Diebold, Neumark, and Polsky (1997) may not have been the first to use the method with cross sectional data, but they were one of the first to provide a detailed explanation of the approach.

[18] From equation (50) one can see that the covariance term will be positive if and only if the covariance between D_{i2} and N_{i2} is also positive.

to 172.9% larger than the DNP estimates and up to 55.6% larger than the NPH estimates. From Tables 2 and 3 one can observe a systematic differential in the standard errors; the gap is larger for longer tenured groups - groups with higher job stability. In general, the extent to which the NPH method underestimates the correct standard errors will be correlated with the size of the employment continuation probability. Focussing on equation (6) illustrates this point. Conditioning on a sample distribution for D_{i1}, a larger $E(N_{i2})$ is associated with a larger first term; a variance term not accounted for by the NPH method.

As a result, the DNP and NPH approaches to estimating standard errors may lead the researcher to falsely reject the null hypothesis of no change in job stability. Calculating t-statistics for males with 12+ years of tenure illustrates this point. Using either the DNP or NPH methods, one strongly rejects the null hypothesis at the 5% significance level. In fact, my method suggests that the null hypothesis should not be rejected, not even at the 10% level.

6 Parametric Approach

In this section, I illustrate the bias of the parametric cross-sectional estimator proposed by Güell and Hu (2006). Using the framework proposed in this paper, I also suggest an adjustment to their method—one that may lead to consistent estimates of the transition (or continuation) probabilities.

The standard approach in the cross sectional literature has been to estimate the transition probability using a non-parametric approach, but Güell and Hu (2006) are the exception. Their parametric approach is based on the Logit estimator. Assume a two year panel where X_{i1} represents the vector of characteristics of individual i in period 1. For individuals that are unemployed in period 1, let Y_{i1} be a dummy variable equal to one if individual i remains unemployed into the next period, and zero otherwise. Finally, define U_{i1} as the length of the on-going unemployment spell of individual i as of period 1. As shown in Güell and Hu (2006), the Logit estimator

for the unemployment-employment transition is simply the sample analog of

$$E(X_{i1}Y_{i1}|U_{i1}=s) = E(X_{i1}\Lambda(X_{i1}\beta)|U_{i1}=s) \tag{16}$$

where $\Lambda(X_{i1}\beta)$, the conditional probability of individual i staying unemployed in period 2, is $\Lambda(.) = exp(.)/(1+exp(.))$. Güell and Hu (2006) argue that one can mimic equation (16) using two representative cross sections of unemployed workers. In particular, they state that the mean characteristics of workers who have been unemployed for $s+1$ periods in period 2 can be used to estimate the left-hand side of equation (16)—as long as the period 2 sample is representative of the unemployment population in period 2. Yet, by the Law of Iterative Expectations

$$E(X_{i1}Y_{i1}|U_{i1}=s) = Prob(Y_{i1}=1|U_{i1}=s)E(X_{i1}Y_{i1}|U_{i1}=s,Y_{i1}=1) \tag{17}$$

$$= Prob(Y_{i1}=1|U_{i1}=s)E(X_{i1}|U_{i1}=s,Y_{i1}=1) \tag{18}$$

$$= Prob(Y_{i1}=1|U_{i1}=s)E(X_{i2}|U_{i2}=s+1) \tag{19}$$

As a result, the Güell and Hu (2006) claim will hold if and only if all unemployed workers remain unemployed into the next period, i.e. if the first right-hand side term in equation (19) equals 1; one must therefore assume a priori a zero transition probability.

Güell and Hu (2006) face the same difficulty as the rest of the cross sectional literature. Approximating panel methods makes the researcher focus exclusively on one group of individuals; for Güell and Hu (2006) which examines the unemployment-continuation, its unemployed workers. Yet, their cross sectional samples of unemployed workers are drawn from a population that changes over time - some workers will lose their jobs while others will find employment. As such, the researcher is faced with the difficult task of of estimating *one* transition probability using samples drawn from *two* different populations.

A possible solution would be to add a first step to the Güell and Hu (2006)

method. In the first step, one estimates the first RHS term of equation (19) - a transition probability - using the framework proposed in this paper, i.e. the results of Proposition 1 and 2. As a second step, one can estimate the remaining moments - as stated in Güell and Hu (2006) - using sample analogs.

7 Conclusion

In this paper, I propose a non-parametric method for estimating labour market transitions. The proposed method, with its clearly stated identifying assumption, is necessary in cases where good panel data are not available. Because it is designed for cross sectional data, it does not face the same difficulties at the inference stage as other cross sectional approaches which approximate panel methods. The key identifying assumption is that one selects from the same population cohort.

Using the conceptual framework proposed in this paper, I also re-examine existing cross sectional methods. I identify the underlying assumptions of the non-parametric approach used in the literature, and propose a consistent estimator for its standard errors. Finally, I show the bias of the recently proposed parametric approach and suggest a possible correction.

A Appendix

A.1

Proposition 1 *The transition probability of a population cohort can be expressed as* $R_1 = \frac{E(N_{i2})}{E(D_{i1})}$.

proof:

$$R_1 = Prob(N_{i2} = 1 | D_{i1} = 1) \tag{20}$$

$$= \frac{Prob(N_{i2} = 1, D_{i1} = 1)}{Prob(D_{i1} = 1)} \tag{21}$$

and since $N_{i2} = 1$ implies $D_{i1} = 1$, one can rewrite R_1 as

$$= \frac{Prob(N_{i2} = 1)}{Prob(D_{i1} = 1)} \tag{22}$$

$$= \frac{E(N_{i2})}{E(D_{i1})} \quad \blacksquare \tag{23}$$

A.2

Proposition 2 *Assuming iid samples for each year that are drawn from a population cohort, then* $\hat{R}_1 \rightarrow R_1$

proof: Apply the Lindberg-Levy Central Limit Theorem

$$\sum_{i=1}^{n_2} N_{i2}/n_2 \xrightarrow{p} E(N_{i1}) \tag{24}$$

$$\sum_{i=1}^{n_1} N_{i2}/n_1 \xrightarrow{p} E(N_{i1}) \tag{25}$$

and use the result of Proposition 1

$$\frac{\sum_{i=1}^{n_2} N_{i2}/n_2}{\sum_{i=1}^{n_1} N_{i2}/n_1} \xrightarrow{p} Prob(N_{i2} = 1 | D_{i1} = 1) \quad \blacksquare \tag{26}$$

A.3

Proposition 3 *Assuming iid samples for each year that are drawn from a population cohort, independence across years, and* $lim_{n_1,n_2 \rightarrow \infty} \frac{n_1}{n_2} = 1$, *then* $\sqrt{n}_1(\hat{R}_1 - R_1) \xrightarrow{d} N(0, V)$ *where* V *is*

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) \tag{27}$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})} \tag{28}$$

proof: For ease of notation let $\hat{N}_j = n_j^{-1} \sum_{i=1}^{n_j} N_{ij}$, $N_j = E(N_{ij})$ and $V_{N_j} = V(N_{ij})$, and define $\hat{D}_j$, D_j and V_{D_j} in a similar fashion.

$$\sqrt{n}_1(\hat{R}_1 - R_1) = \sqrt{n}_1 \left(\frac{\hat{N}_2}{\hat{D}_1} - \frac{N_2}{D_1} \right) \tag{29}$$

$$= \sqrt{n}_1 \frac{(\hat{N}_2 - N_2)D_1 - (\hat{D}_1 - D_1)N_2}{D_1 \hat{D}_1} \tag{30}$$

$$= \frac{\sqrt{\frac{n_1}{n_2}}\sqrt{n}_2(\hat{N}_2 - N_2)D_1 - \sqrt{n}_1(\hat{D}_1 - D_1)N_2}{D_1^2} + o_p(1) \tag{31}$$

$$= -\phi_1 \sqrt{n}_1(\hat{D}_1 - D_1) + \phi_2 \sqrt{n}_2(\hat{N}_2 - N_2) + o_p(1) \tag{32}$$

$$\xrightarrow{d} N\left(0, \phi_1^2 V_{D_1} + \phi_2^2 V_{N_2}\right) \quad \blacksquare \tag{33}$$

A.4

Proposition 4 *Assume that the composition of a country's population changes from period 1 to period 2. Further assume that these compositional changes break (or interrupt) the spell of interest. The continuation probability for the period 1 population can be expressed as*

$$R_1 = adj_1 \frac{E(N_{i2})}{E(D_{i1})} \tag{34}$$

where adj_1 is the population growth factor.

proof: To ease the presentation, I assume that the change in population is due to the arrival of one new immigrant in year 2. Similar arguments would hold true for other population changes. Without loss of generality, Assume a population of size n in year 1, and $n+1$ in year 2. Order the year 2 populations so that the new immigrant in year 2 is last. By Proposition 1, the continuation probability of the

year 1 population is

$$R_1 = \frac{\sum_{i=1}^{n} N_{i2}/n}{\sum_{i=1}^{n} D_{i1}/n} \tag{35}$$

By assuming that the change in population results in breaks in the spell of interest, one can conclude that $\sum_{i=1}^{n} N_{i2} = \sum_{i=1}^{n+1} N_{i2}$. As a result, R_1 can be rewritten as

$$= \left(\frac{n+1}{n}\right) \frac{\sum_{i=1}^{n+1} N_{i2}/n + 1}{\sum_{i=1}^{n} D_{i1}/n} \tag{36}$$

$$\equiv adj_1 \frac{E(N_{i2})}{E(D_{i1})} \quad \blacksquare \tag{37}$$

A.5

Proposition 5 *Given repeated cross sections where the base weights sum up to the target population, then* $\tilde{Q}_1 = \hat{adj}_1 \tilde{R}_1$.

proof:

$$\tilde{Q}_1 = \frac{\tilde{n}_2^{s+1,c}}{\tilde{n}_1^{s,c}} \tag{38}$$

$$= \frac{\sum_{i=1}^{n_2} bw_{i2} N_{i2}}{\sum_{i=1}^{n_1} bw_{i1} D_{i1}} \tag{39}$$

$$= \frac{\sum_{i=1}^{n_2} bw_{i2}}{\sum_{i=1}^{n_1} bw_{i1}} \cdot \frac{\sum_{i=1}^{n_2} \frac{bw_{i2}}{\sum_{i=1}^{n_2} bw_{i2}/n_2} N_{i2}/n_2}{\sum_{i=1}^{n_1} \frac{bw_{i1}}{\sum_{i=1}^{n_1} bw_{i1}/n_1} D_{i1}/n_1} \tag{40}$$

$$= \frac{\sum_{i=1}^{n_2} bw_{i2}}{\sum_{i=1}^{n_1} bw_{i1}} \cdot \frac{\sum_{i=1}^{n_2} nw_{i2} N_{i2}/n_2}{\sum_{i=1}^{n_1} nw_{i1} D_{i1}/n_1} \tag{41}$$

$$= \frac{\sum_{i=1}^{n_2} bw_{i2}}{\sum_{i=1}^{n_1} bw_{i1}} \cdot \tilde{R}_1 \tag{42}$$

$$= \hat{adj}_1 \tilde{R}_1 \quad \blacksquare \tag{43}$$

A.6

Proposition 6 *Assume iid samples for each year that are drawn from a population where compositional changes break (or interrupt) the spell of interest. Further assume independence across years,* $lim_{n_1,n_2 \to \infty} \frac{n_1}{n_2} = 1$*, and where* $\hat{adj}_1$ *is the true*

population growth factor. Then, $\sqrt{n}_1(\tilde{Q}_1 - \tilde{Q}_1) \xrightarrow{d} N(0, V)$ *where* V *is*

$$V = adj_1^2 \left[\phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2})\right] \tag{44}$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})} \tag{45}$$

proof:

$$\sqrt{n}_1(\tilde{Q}_1 - \tilde{Q}_1) = \sqrt{n}_1(\hat{a}dj_1\tilde{R}_1 - adj_1 R_1) \tag{46}$$

$$= adj_1\sqrt{n}_1(\tilde{R}_1 - R_1) \tag{47}$$

$$\xrightarrow{d} adj_1 N\left(0, \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2})\right) \quad \text{(by Proposition 3)} \tag{48}$$

$$\xrightarrow{d} N\left(0, adj_1^2[\phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2})]\right) \quad \blacksquare \tag{49}$$

A.7

Proposition 7 *Assuming iid samples for each year, samples of equal size, independence across years, and no change in population, then* $\sqrt{n}((\hat{R}_j - \hat{R}_1) - (R_j - R_1) \xrightarrow{d} N(0, V)$ *where* V *depends on* j*, an integer greater than or equal to 2. Case* 1*:* $j = 2$

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) + \phi_3^2 V(D_{i2}) + \phi_4^2 V(N_{i3}) + 2\phi_2\phi_3\mu Cov(D_{i2}, N_{i2}) \tag{50}$$

Case 2*:* $j \geq 3$

$$V = \phi_1^2 V(D_{i1}) + \phi_2^2 V(N_{i2}) + \phi_3^2 V(D_{ij}) + \phi_4^2 V(N_{ij+1}) \tag{51}$$

with

$$\phi_1 = \frac{E(N_{i2})}{[E(D_{i1})]^2}, \quad \phi_2 = \frac{1}{E(D_{i1})}, \quad \phi_3 = \frac{E(N_{ij+1})}{[E(D_{ij})]^2}, \quad \phi_4 = \frac{1}{E(D_{ij})} \tag{52}$$

and μ is the probability that a random chosen person in the population aged 16 years and up, will be 17 or more years of age.

proof: For ease of notation let $\hat{N}_j = n_j^{-1} \sum_{i=1}^{n_j} N_{ij}$, $N_j = E(N_{ij})$ and $V_{N_j} = V(N_{ij})$, and define $\hat{D}_j$, D_j and V_{D_j} in a similar fashion. Finally, let $C_2 = Cov(D_{i2}, N_{i2})$

Case 1: $j = 2$

$$\sqrt{n}((\hat{R}_2 - \hat{R}_1) - (R_2 - R_1)$$

$$= \sqrt{n}\left(\left(\frac{\hat{N}_3}{\hat{D}_2} - \frac{N_3}{D_2}\right) - \left(\frac{\hat{N}_2}{\hat{D}_1} - \frac{N_2}{D_1}\right)\right) \tag{53}$$

$$= \sqrt{n}\frac{(\hat{N}_3 - N_3)D_2 - (\hat{D}_2 - D_2)N_3}{D_2\hat{D}_2} - \sqrt{n}\frac{(\hat{N}_2 - N_2)D_1 - (\hat{D}_1 - D_1)N_2}{D_1\hat{D}_1} \tag{54}$$

$$= \sqrt{n}\frac{(\hat{N}_3 - N_3)D_2 - (\hat{D}_2 - D_2)N_3}{D_2^2} - \sqrt{n}\frac{(\hat{N}_2 - N_2)D_1 - (\hat{D}_1 - D_1)N_2}{D_1^2} + o_p(1) \tag{55}$$

$$= \phi_1\sqrt{n}(\hat{D}_1 - D_1) - \phi_2\sqrt{n}(\hat{N}_2 - N_2) - \phi_3\sqrt{n}(\hat{D}_2 - D_2) + \phi_4\sqrt{n}(\hat{N}_3 - N_3) + o_p(1) \tag{56}$$

$$\xrightarrow{d} N\left(0, \phi_1^2 V_{D_1} + \phi_2^2 V_{N_2} + \phi_3^2 V_{D_2} + \phi_4^2 V_{N_3} + 2\phi_2\phi_3\mu C_2\right) \tag{57}$$

Case 2: $j \geq 3$. The proof is similar to Case 1, with one exception. Since the four components of the test statistics, i.e. $\hat{N}_{j+1}$, $\hat{D}_j$, $\hat{N}_2$ and $\hat{D}_1$ are functions of different yearly samples when $j \geq 3$, the covariance term is zero. ∎

Replacing the population moments with corresponding sample analogs generates a consistent estimator for each asymptotic variance.

References

Baker, M. (1992): "Unemployment Duration: Compositional Effects and Cyclical Variability," *American Economic Review*, 82(1), 313–321.

Brochu, P. R. (2006): "An Exploration in Job Stability," PhD Thesis, University of British Columbia.

Diebold, F. X., D. Neumark, and D. Polsky (1997): "Job Stability in the United States," *Journal of Labor Economics*, 15(2), 206–233.

Güell, M., and L. Hu (2006): "Estimating the Probability of Leaving Unemployment Using Uncompleted Spells from Repeated Cross-section Data," *Journal of Econometrics*, 133(1), 307–341.

Jones, S. R. G., and W. C. Riddell (1999): "The Measurement of Unemployment: An Empirical Approach," *Econometrica*, 67(1), 147–162, Notes and Comments.

Neumark, D., D. Polsky, and D. Hansen (1999): "Has Job Stability Declined Yet? New Evidence for the 1990s," *Journal of Labor Economics*, 17(4), S29–S64.

——— (2000): "Has Job Stability Declined Yet? New Evidence for the 1990s," in *On the Job: Is Long-Term Employment a Thing of the Past?*, ed. by D. Neumark, chap. 3, pp. 70–110. Russell Sage Foundation, New York.

Swinnerton, K. A., and H. Wial (1995): "Is Job Stability Declining in the U.S. Economy?," *Industrial and Labor Relations Review*, 48(2), 293–304.

Ureta, M. (1992): "The Importance of Lifetime Jobs in the US Economy, Revisited," *American Economic Review*, 82(1), 322–335.

Table 1: 1-year Employment-Continuation Probabilities: Canada, 2000

Group Specification	Proposed Method ($\tilde{R}_1$)	Existing Method ($\tilde{Q}_1$)	Population Growth Factor ($\hat{adj}_1$)
Overall	0.8040	0.8101	1.0076
	(0.0028)	(0.0029)	
Male	0.7975	0.8036	1.0076
	(0.0050)	(0.0051)	
Female	0.8108	0.8170	1.0076
	(0.0052)	(0.0053)	
Tenure less than 1 year	0.5517	0.5559	1.0076
	(0.0070)	(0.0071)	

Table 2: U.S. 4-year Male Employment-Continuation Probabilities - Time Differentials

Tenure Group Specification	1996	2000	Difference	Standard Errors Method
0-2	0.4775	0.4999	0.0224	
			(0.0076)**	DNP
			(0.0106) **	NPH
			(0.0116) **	proposed
			(0.0118) *	proposed (no covariance term)
3-6	0.4522	0.4658	0.0136	
			(0.0092)	DNP
			(0.0133)	NPH
			(0.0148)	proposed
			(0.0151)	proposed (no covariance term)
7-11	0.7069	0.6554	-0.0515	
			(0.0111)**	DNP
			(0.0197)**	NPH
			(0.0250)**	proposed
			(0.0244)**	proposed (no covariance term)
12+	0.7288	0.6927	-0.0362	
			(0.0083)**	DNP
			(0.0149)**	NPH
			(0.0227)	proposed
			(0.0194)*	proposed (no covariance term)
total	0.5798	0.5716	-0.0081	
			(0.0045)*	DNP
			(0.0060)	NPH
			(0.0083)	proposed
			(0.0073)	proposed (no covariance term)

** The estimated difference is significant at the 5% level

* The estimated difference is significant at the 10% level

Table 3: U.S. 4-year Female Employment-Continuation Probabilities - Time Differentials

Tenure Group	1996	2000	Difference	Standard Errors
Specification				Method
0-2	0.4294	0.4785	0.0491	
			(0.0074)**	DNP
			(0.0097)**	NPH
			(0.0111)**	proposed
			(0.0113)**	proposed (no covariance term)
3-6	0.4370	0.4253	-0.0117	
			(0.0092)	DNP
			(0.0125)	NPH
			(0.0148)	proposed
			(0.0150)	proposed (no covariance term)
7-11	0.6630	0.6134	-0.0496	
			(0.0106)**	DNP
			(0.0206)**	NPH
			(0.0270)*	proposed
			(0.0272)*	proposed (no covariance term)
12+	0.7264	0.6699	-0.0565	
			(0.0098)**	DNP
			(0.0172)**	NPH
			(0.0268)**	proposed
			(0.0231)**	proposed (no covariance term)
total	0.5355	0.5322	-0.0033	
			(0.0047)	DNP
			(0.0060)	NPH
			(0.0085)	proposed
			(0.0075)	proposed (no covariance term)

** The estimated difference is significant at the 5% level

* The estimated difference is significant at the 10% level