SSC Annual Meeting, June 1999
Proceedings of the Survey Methods Section

TREND CORRELATION OF LABOUR MARKET EARNINGS IN CANADA: 1982 TO 1995

E. Chen and G. Rowe[1]

ABSTRACT

Using Statistics Canada's Longitudinal Administrative Database (LAD) data, this paper uniquely contributes to the understanding of stability of the labour market real earnings in Canada by using Kendall's rank correlation and significance test to identify the presence of career trends in earnings streams of individual tax payers. Distribution of the Kendall's rank correlation for trend, lag correlation and partial correlation may be compared and will shed light on the stability of individual career in the labour market from 1982 to 1995. The availability of vast individual earnings data with over 1.5 million records of earnings information from the LAD database will provide useful information on rank measures and distributions on the individual careers in Canada.

KEY WORDS: Microsimulation; Kendall's tau correlation; Evaluation.

RÉSUMÉ

Le présent document, qui utilise les données de la banque des Données administratives longitudinales (DAL) de Statistique Canada, vise uniquement à mieux comprendre la stabilité des revenus réels du marché du travail au Canada en se servant de la corrélation des rangs de Kendall et d'un test de signification afin d'identifier la présence des tendances de carrière parmi les flux de revenus des contribuables. La distribution de la corrélation des rangs de Kendall des tendances, de la corrélation avec retard et de la corrélation partielle peuvent être comparées, ce qui éclaircira la stabilité des carrières individuelles du marché du travail entre 1982 et 1995. La disponibilité d'abondantes données sur les revenus des particuliers dont plus de 1,5 million d'enregistrements sur les revenus compris dans la banque DAL offriront des renseignements utiles sur les mesures de classement et la distribution des carrières individuelles au Canada.

MOTS-CLÉS : Microsimulation; corrélation du tau de Kendall; évaluation.

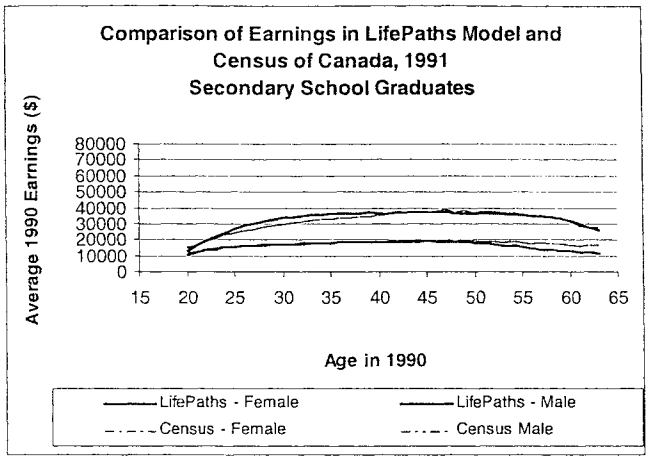
1. INTRODUCTION

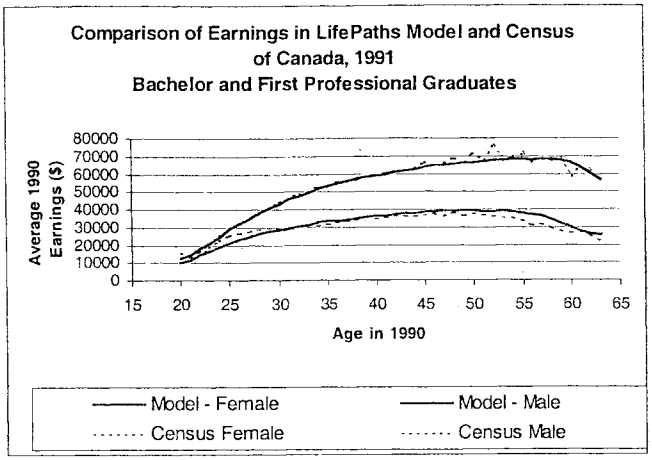
LifePaths is a longitudinal microsimulation model in which individual lifetimes are simulated. The model initializes a case by randomly generating an individual's sex, province of residence, age at immigration and year of birth. Individual's year of birth can range from 1892 to 2051; but by fixing mortality and immigration assumptions, births occurring in 1892-1991 will reproduce provincial age-sex structures as enumerated in the 1991 Census of Canada. The set of variables that describe the demographic, social and economic circumstances of the individual undergoes changes as he/she ages. These changes are dictated by the events in each individual life. The model comprises a set of events that can affect each case at appropriate points in their lifetime. These events include: entering the education system, education progression, graduation, entering the job market, gaining or losing a job, common law union formation, marriage, having children, separation, divorce and death. Ultimately, the individual's lifetime concludes when the death event occurs.

The LifePaths model undergoes continuous updating and refinement by using different data sources available in Statistics Canada. To date, LifePaths has been employed in a broad range of policy analysis and research activities. Examples of LifePaths applications include: analysis of Canada Student Loan policy options (under contract to HRDC and the Government of Ontario), study of returns to education, examination

[1] Edward Chen and Geoff Rowe, Statistics Canada , Socio-Economic Modeling Group, 24th Floor R.H.Coats Building, Ottawa, Ontario, Canada, K1A 0T6. e-mails – chenedw@statcan.ca, rowegt@statcan.ca

Figure 1





of time use over the life course (Wolfson and Rowe 1996; Wolfson, 1997; Wolfson and Rowe 1998a) and simulating future prospects for the tax-transfer system and pensions (Wolfson, Rowe, Gribble and Lin 1998; Wolfson and Rowe 1998b). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1997; Rowe and Chen 1998; Plager and Chen 1999).

2. OBJECTIVES

In its current state of development, LifePaths contains relatively mature models of secondary school and post-secondary educational careers at the provincial level. In addition to accounting for life course events, LifePaths also imputes various individual characteristics: in particular, hourly equivalent wage rates (conditional on sex, education level, field of study and duration since graduation) and usual hours of work (conditional on education level, sex, age and hours last year). These characteristics together with employment status determine individual earnings.

Currently, the earnings module in the LifePaths is derived from 1991 Census of Canada earnings information. Figure 1 displays the comparisons of calibrating results of LifePaths and Census, 1991 for both the Secondary School Graduates and Bachelor and First Professional Graduates.

In this study, we examine the current earnings module in the LifePaths and compare the results to another data source the Longitudinal Administrative Databank (LAD). Specifically, we try to answer the following questions:

1. What were the trends in individual earnings in Canada for all earners in their prime age? The prime age/sex groups of interests are males and females at aged 30-34, aged 35-39 and aged 40-44 in 1982 and their earnings observed over a 14-year period from 1982 to 1995;
2. What were the levels and volatility of individual earnings in these age groups (as measured by averages and standard deviations) over the 14-year period?
3. How do the results from the LifePaths Model compare to the LAD income file? And how can the differences be explained or how can the LifePaths be improved?

3. LAD INCOME DATA

3.1 LAD Income Tax File

Statistics Canada's Longitudinal Administrative Databank (LAD) income tax file is a 10 per cent sample of Canadian income tax filers on an annual basis. Individual information on earnings, taxes, and basic demographic characteristics provide a large sample size, dynamic structure and detailed individual income information to conduct both dynamic analyses of earnings and income mobility in Canada as well as the provinces. Being a longitudinal data file, LAD contains individual income information over time and enables us to explore the questions mentioned earlier. The LAD file had personal income tax information from year 1982 to 1995 at the time of the study. Future years of income tax data are added as they become available.

Individual income tax filers are selected into the LAD database based on a random selection of Social Insurance Numbers. If the individual is selected into the database in any given year, that person will be also selected in all other years in which a tax form is filed. There are also selection procedures in place to ensure a

representative cross section as well as a representative sample of the longitudinal aspect of the entire population. These procedures include the selection of new income tax filers who are either young people or immigrants, for example.

LAD file is thus a representative sample of the adult population in Canada. This data source is reported to cover over 95 per cent of the target adult population (especially in the prime age groups) in Canada. It has good coverage of income tax filers and good data quality of income related variables. On the other hand, the limitation of the LAD file is the lack of the education-related variables, such as highest education attainment or any of the labour market activity variables. The variables included in LAD file are only what the information was available on the income tax return forms.

3.2 Sample Selection Restrictions

This analysis includes all sample individuals available in the LAD file with several restrictions: 1. The individual must be still alive in year 1995; the last year LAD has income data at the time of study; 2. The individual has at least three years of non zero earnings over the 14-year period. The latter is required for measures of volatility and/or correlation to be estimable.

3.3 Confidentiality

The Statistics Act precludes the release of tax data, which would make it possible to identify the individuals in the LAD income file. Data and results released in this study followed rules that are applied to prevent data disclosure. In particular, the confidentiality rules regarding suppression with minimum sample size, residual disclosure and rounding are all applied. The grouping of results in the Atlantic provinces is one example of confidentiality requirements.

4. DEFINITIONS AND CORRELATION MEASURES

4.1 Definitions

As mentioned, the LAD data file provides a large number of individual income information with over 1.5 millions records with a 14-year period. Cox (1986) outlined an approach to analysis distribution information with large numbers of observations over a short observation period. In this study, several concepts are defined and applied to both data sources when retrieving and processing information.

Earnings: The earnings refer to the total dollars received by the individual during a calendar year from wages, salaries and commissions and self-employment. The annual earnings are adjusted by the constant average industrial wages in 1995 to account for wages increases over the years.

Age/Sex: Age represents the age at end of 1982. Individual age is reported in three prime earnings age groups: Age 30-34, Age 35-39 and Age 40-44 by sex in 1995.

Province: The province is the province that the individual resided in December 1995. If the individual is a non-filer in 1995, province is then the last province that the individual resided in the immediate previous year. The results are reported by the Atlantic provinces (include Newfoundland, Prince Edward Island, Nova Scotia and New Brunswick); Ontario and Canada (excluding persons residing in Yukon and Northwest Territories in 1995).

Years: The years are the income years that the individual earnings occurred. The years are from 1982 to 1995 calendar years. The years span over a 14-year period in which the individual must have at least three years of non-zero earnings.

Zeros: Zero earnings in any reference year can occur either due to the individual having no earnings or due to the individual not filing an tax return in that year.

3.2 Statistical Correlation Measures

Statistical correlation measures provide information on the closeness of a relationship between two variables. If the earnings variable can be expressed exactly as a linear function of another variable, for example year, then the correlation coefficient is either 1 or –1 depending on whether the real earnings increase with time or decline with time. Where in this study takes the variable year into consideration, the correlation coefficient measures the regularity and direction over time of individual earnings trends.

Several different correlation measures are used to evaluate the individual earnings patterns over the 14-year period. The measures adopted include the commonly used Pearson correlation coefficient as well

as Kendall's tau, lagged tau and partial tau correlations.

Pearson coefficient:

The Pearson product moment correlation coefficient ρ is a widely used measure that assesses linear association between variables. We have used it to measure trends in individual annual earnings. The coefficient is easily influenced by the presence of extreme values, such as an individual having no earnings in one year or very high earnings in another year.

Trend tau:

Kendall's tau correlation measures the agreement between two sets of rankings. It can also be described as measuring the degree of concordance or discordance between two variables. Variables are in concordance for any pair or observations (x_j, x_i) if the sign of the difference (y_j, y_i) tends to be the same as the sign (x_j, x_i). Variables are discordant if the signs differ. The trend tau is derived from the probability of individual annual earnings increasing by a positive amount over any two years. The excess of concordance over discordance is

$$S_{trend} = N_c - N_d = \sum_{i<j}\sum \mathrm{sgn}(X_{Y_j} - X_{Y_i})\,\mathrm{sgn}(Y_j - Y_i),$$

where X denotes earnings and Y denotes year. The trend coefficient is: $\tau_{trend} = \dfrac{2S_{trend}}{n(n-1)}$.

Lagged tau:

Kendall's tau can equally be applied to assess lagged agreement. Here one measures rankings on a pair representing the earnings in current year and in previous year. This correlation measures the "smoothness" in variation of earnings from one year to another.

$$S_{Lagged} = N_c - N_d = \sum_{i<j}\sum \mathrm{sgn}(X_{Y_j} - X_{Y_i})(X_{Y_{j-1}} - X_{Y_{i-1}})$$

We have the lagged tau correlation: $\tau_{lagged} = \dfrac{2S_{lagged}}{n(n-1)}$.

Partial tau:

The partial tau defined in this study measures the smoothness of the individual earnings over time after the trend is removed. We define the partial tau as

$$\tau_{partial} = \frac{\tau_{lagged} - \tau_{trend,y}\tau_{trend,y-1}}{\sqrt{(1-\tau^2_{trend,y})(1-\tau^2_{trend,y-1})}} = \frac{\tau_{lagged} - \tau^2_{trend}}{1-\tau^2_{trend}}$$

Statistical hypothesis test:

Kendall's tau can be used to test the null hypothesis of no trend with two possible alternatives; the individual earnings have increased with time or have decreased with time. Small sample quantiles of the test statistic can be found in Conover (1980). If the empirical result exceeds the 95% quantile, we can reject the null hypothesis of no trend and conclude the individual earnings show evidence a systematic increase with time. On the other hand, if the empirical result is less than the 5% quantile, the individual's earnings show a systematic decrease with time.

4. EMPIRICAL FINDINGS AND DATA ANALYSIS

4.1 Correlation Measures

Table 1 (in Appendix) presents the comparison of correlation coefficients from the LAD income file with the results from the LifePaths Model. The results include both the averages of the computed individual correlation coefficients as well as the median estimates by the selected prime age/sex groups.

As expected, the Pearson correlation coefficients are larger in absolute values than the Kendall's Tau trend coefficients. The tau lagged coefficients indicate that individual earnings over the 14-year period have tended to have low year to year correlation. Low correlation tends to imply instability in earnings from one year to the next with females having slightly higher stability than males. The lagged correlations may confound trends with year to year stability. The partial tau coefficients remove the trend effect showing virtually no difference between males and females. In other words, tau statistics indicate that the males and females had similar year to year stability in earnings in the 14-year period, but that they may differ in respect to stability of career trends.

In most respects, the results from the LAD income file are in agreement with the results from the LifePaths Model. Differences that exist might be explained by the fact that the LifePaths Model was developed using 1991 Census information and 1990's educational progression (i.e., in order to project into the future outcomes of current education cohorts). Individuals in the LAD file, however, showed patterns earnings consistent with the lower historical education participation and graduation rates of the 60's and 70's.

Changes in female participation in the labour force since the 1970's may also account for some of the differences.

Table 2 presents the comparison of career trends in earnings streams of individual taxpayers. A significance test on the trend Tau is used to determine whether individual real earnings have changed within the 14-year period. The test results are classified into three categories: Positive (if the individual real earnings have increased); Negative (if the individual real earnings have decreased); or no real change (if none of the above).

Only about one third of the prime age individuals in Canada had their real earnings increase over the 14-year period. Another 20-25% had no real change in their earnings after adjusting for the increases in the average industrial wage. Less 10% experienced a real earnings decrease. The age group, which showed the greatest proportion with increased earnings, was the age group 30-34 for both males and females. Residents of Ontario tended to have much higher chance of increase compared to the residents in the Atlantic region.

Table 3 presents an interest comparison of averages of earnings and averages of standard deviations. Higher earnings are shown to be associated with higher standard deviations and with higher coefficients of variation (C.V.). This appears to indicate degrees of fluctuation in annual earnings are higher for high income individuals and/or may increase over individual careers. The average earnings and its standard deviation from the LifePaths Model show very similar results with the LAD income file. On the whole, the LifePaths Model, as expected, simulates higher average earnings since the education background in the Model tend to be higher than is found in the LAD population. The coefficients of variation are very comparable results between these two data source.

5. FUTURE WORK

The work presented here is a further stage in the construction, validation and refinement of the comprehensive microsimulation model. In the long term, one of our objectives is to put together a model that will simulate historical education and labour market participation. So that even closer agreement between LAD and LifePaths Model may be expected from furture versions.

LifePaths is and will remain, a work in progress, since the goal of the model is to encapsulate as much detail as possible on socio-economic processes in Canada as well as the historical patterns of change in those processes.

REFERENCES

Chen, E.J. and Oderkirk, J. (1998). "Varied Pathways: The Undergraduate Experience in Ontario", Feature article. Education Quarterly Review, Statistics Canada, 81-003-XPB, Vol. 4, No. 3, 1998.

Conover, W.J. (1980). Practical Nonparametric Statistics 2nd Edition, John Wiley & Sons.

Cox and Solomon (1986). Analysis of Variability with large number of small sample. Biometrika, 73, 3, pp 543-54.

LAD Dictionary, Longitudinal Administrative Data Bank Dictionary April, 1997, Version 1.0

Plager, L. and E.J. Chen (1999). "Student Debt from 1990/91 to 1995/96: An Analysis of Canada Student Loans Data". Major Release, Statistics Canada Daily, Education Quarterly Review, Statistics Canada, 81-003-XPB, Vol. 5, No. 4, 1999.

Rowe, G. and E.J. Chen (1998). "An Increment-Decrement Model of Secondary School Progression for Canadian Provinces", Symposium on Longitudinal Analysis for Complex Surveys, Statistics Canada, Ottawa.

Wolfson, M.C. (1997). "Sketching LifePaths™: A New Framework for Socio-Economic Statistics" in R. Conte, R. Hegselmann and P. Terna (Eds.), *Simulating Social Phenomena, Lecture Notes in Economics and Mathematical Systems 456*, Springer.

Wolfson, M.C. and G. Rowe (1996). "Perspectives on Working Time Over the Life Cycle", Canadian Employment Research Forum Conference on "Changes to Working Time", Ottawa.

Wolfson, M.C. and G. Rowe (1998a). "LifePaths – Toward an Integrated Microanalytic Framework for Socio-Economic Statistics", 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Wolfson, M.C. and G. Rowe (1998b). "Public Pension Reforms – Analyses Based on the LifePaths Generational Accounting Framework", 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Appendix Tables

Table 1 Correlation Statistics at Canada Level for the Prime Age Groups

Sex	Age Group	No. of Obs.	LAD Income File				LifePaths Model			
			Rho	Tau - Trend	Tau - Lagged	Tau - Partial	Rho	Tau - Trend	Tau - Lagged	Tau - Partial
AVERAGES										
Female	30-34	96,572	0.279	0.244	0.332	0.140	0.171	0.137	0.237	0.115
	35-39	78,550	0.220	0.197	0.325	0.142	0.123	0.101	0.223	0.106
	40-44	56,363	0.100	0.110	0.306	0.140	0.048	0.047	0.210	0.100
Male	30-34	105,129	0.137	0.135	0.294	0.141	0.232	0.190	0.264	0.117
	35-39	88,835	0.050	0.067	0.286	0.144	0.168	0.140	0.237	0.112
	40-44	67,159	-0.055	0.008	0.280	0.145	0.114	0.094	0.217	0.106
MEDIANS										
Female	30-34	96,572	0.404	0.297	0.364	0.162	0.221	0.154	0.282	0.133
	35-39	78,550	0.319	0.231	0.359	0.162	0.162	0.121	0.256	0.123
	40-44	56,363	0.131	0.121	0.345	0.163	0.067	0.050	0.231	0.121
Male	30-34	105,129	0.175	0.143	0.333	0.157	0.299	0.231	0.298	0.128
	35-39	88,835	0.049	0.055	0.318	0.160	0.222	0.165	0.265	0.123
	40-44	67,159	-0.108	-0.033	0.308	0.161	0.149	0.121	0.235	0.118

Table 2 Comparison of Presence of Career Trends Using Trend Tau Statistic

	Sex	Age Group	No. of Obs.	LAD Income file			LifePaths Model		
				Positive	Negative	No Change	Positive	Negative	No Change
				%			%		
Canada	Female	30-34	96,572	36.2	5.4	58.3	24.1	7.3	68.6
		35-39	78,550	33.2	7.3	59.5	20.8	8.8	70.4
		40-44	56,363	26.2	10.5	63.3	17.3	11.2	71.5
	Male	30-34	105,129	27.4	9.2	63.4	31.6	6.2	62.1
		35-39	88,835	22.3	12.6	65.1	25.1	7.7	67.2
		40-44	67,159	17.1	16.6	66.3	20.0	8.6	71.4
Atlantic	Female	30-34	7,724	31.4	7.3	61.3	20.0	9.0	71.7
		35-39	6,164	28.2	11.0	60.7	19.4	8.0	69.3
		40-44	4,279	19.3	16.4	64.4	22.7	11.0	71.4
	Male	30-34	8,798	22.1	12.7	65.2	17.6	11.0	66.9
		35-39	7,337	16.4	17.8	65.8	25.8	7.3	69.6
		40-44	5,411	12.5	24.7	62.8	23.3	7.1	71.5
Ontario	Female	30-34	36,916	40.6	4.2	55.1	25.1	5.8	69.1
		35-39	30,373	39.0	5.2	55.8	19.9	8.6	71.6
		40-44	21,977	33.1	7.5	59.5	17.5	10.7	71.8
	Male	30-34	38,770	30.2	7.2	62.6	30.1	6.4	63.5
		35-39	33,340	26.1	9.6	64.3	25.0	8.0	66.2
		40-44	25,389	20.8	12.4	66.8	19.2	8.9	71.9

Table 3. Comparison of Individual Average Earnings - LAD Income File and LifePaths Model

			LAD Income File			LifePaths Model		
	Sex	Age Group	Average Income	Std.	C.V.	Average Income	Std.	C.V.
Canada	Female	30-34	20,100	11,200	55.7%	23,655	16,504	69.8%
		35-39	20,700	11,600	56.0%	25,213	18,645	73.9%
		40-44	19,700	10,700	54.3%	23,720	13,507	56.9%
	Male	30-34	39,600	31,700	80.1%	39,344	31,428	79.9%
		35-39	42,400	34,300	80.9%	42,206	26,724	63.3%
		40-44	40,500	32,400	80.0%	41,998	36,600	87.1%
Altantic	Female	30-34	16,400	9,200	56.1%	20,367	8,624	42.3%
		35-39	16,300	7,500	46.0%	23,292	8,076	34.7%
		40-44	15,900	7,400	46.5%	18,088	6,946	38.4%
	Male	30-34	32,300	15,400	47.7%	32,711	13,212	40.4%
		35-39	34,300	18,100	52.8%	33,100	14,358	43.4%
		40-44	34,100	17,500	51.3%	35,196	52,035	147.8%
Ontario	Female	30-34	22,500	13,000	57.8%	26,352	13,586	51.6%
		35-39	23,100	13,500	58.4%	28,012	16,716	59.7%
		40-44	22,000	12,100	55.0%	25,121	9,845	39.2%
	Male	30-34	41,100	43,200	105.1%	43,649	32,993	75.6%
		35-39	45,600	41,100	90.1%	47,340	25,841	54.6%
		40-44	46,200	37,200	80.5%	45,025	18,907	42.0%