WORKING
PAPERS
W

Lasse Koskinen, Tapio Nummi,
Janne Salonen

Modelling and Predicting Individual Salaries: A Study of Finland's Unique Dataset

Finnish Centre for Pensions

WORKING PAPERS

Lasse Koskinen, Tapio Nummi,
Janne Salonen

Modelling and Predicting Individual Salaries: A Study of Finland's Unique Dataset

Finnish Centre for Pensions

WORKING PAPERS

ABSTRACT

This paper models wages by exploiting unique Finnish dataset. The data is divided into four subgroups - for both genders and income quartiles. Panel data model with a range of variables - functional forms of age, duration of employment, and the GDP growth - as explanatory variables is estimated. The model provides predictors for the future individual wages. Out of sample predictions are made for a normal growth and a deep depression period. The results show that individual aspects and wage group may play a significant role in modelling and prediction. This is potentially useful information when designing and developing pension scheme.

ACKNOWLEDGEMENTS

Lasse Koskinen[1], Tapio Nummi[2] and Janne Salonen[3]

---

[1] Insurance Supervisory Authority, Finland.
[2] University of Tampere, Finland.
[3] The Finnish Centre for Pensions, Finland.

CONTENTS

1   Introduction

Different types of models have been proposed in the literature for describing average salary profiles, but individual profiles are rarely modelled. However, a moderate average wage is not equivalent to moderate pension for all individuals. A balance should to be found between the "social side" and the financial viability of the scheme, in particular with an evaluation of the repercussions of pension policies for the situation of individuals.

Clearly individual wage profiles are needed in pension system development but the modelling is often limited by lack of adequate data. For example Carrier & Sand (1998) argued that the challenges surrounding the projection of salaries are numerous. One challenge is that employers do not easily volunteer salary data. Bone and Mitchell (1997) presented a strong case for obtaining more data and constructing better models of retirement income elements, so that actuaries can make better estimates and policymakers can make better choices.

Shiller (2003) also stressed the importance of good data sources. He claimed that present-day insurance and financial systems, so vital to individual well-being and social welfare, would be impossible without suitable data sources. If we are to expand the risk management institutions we will need better, more encompassing databases and data management technologies to cover these risks.

Recently empirical research in insurance has been enriched by the availability of a wealth of new source of data cross sections of individuals observed over time (cf. Plamondon, P. et al 2002). This provides several advantages over conventional cross-sectional or time-series approaches. This facilitates the construction and testing of more realistic behavioural models.

In this paper a unique individual dataset is exploited. It consists of all the participants of the Finnish private-sector statutory pension scheme who retired in 1998.

The ability to model and perform decision modelling and analysis is an essential feature of pension applications ranging from design to planning and control of systems. Models showing correlation or causation between variables can be used to improve decision-making. They can serve for instance as an early warning system or they can be used to explore the financial effects of alternatively policy options. However, modelling is fraught with dangers. A model, which heretofore was valid, may lose validity due to changing conditions, thus becoming an inaccurate representation of reality and adversely affecting the ability of the decision-maker to make good decisions.

Almost all policy decisions are based on prediction. Every decision becomes operational at some point in the future, so it should be based on prediction of future conditions. Predicting is a necessary input to planning, whether in business, or government. The

selection and implementation of the proper predicting methodology has always been an important planning and control issue for insurance companies and goverment's agencys.

The general objective of this study is to develop model that describes - at least in part - individual features of salary development and can be used for prediction purposes.

In order to properly analyse the data and to make a satisfactory prediction, it is essential to understand the environment in which the data has been collected. In particular how the environment has changed in the past. Here we studied the Finnish case as an example. During the years 1991-1993 Finland suffered from a depression that in many ways was at least as severe as the Great Depression in the 1930s. The losses of the Finnish private-sector statutory pension scheme were reported in Koskinen and Pukkila (2002). This demonstrates how difficult it is to make predictions for a deep recession period.

Panel data models are regression type models that have been used extensively in practical application. Frees et al (2001) demonstrated the use of panel data models through series of actuarial case studies. These case studies illustrate how a broad class of panel data models can be applied to different functional areas and to data that have different features. Here we model panel data of individual salaries by a mixed linear model. Their parameters contain both fixed and random effects (cf. e.g. Pinheiro and Bates 2000, Nummi and Möttönen 2000).

We divided the wage data into subgroups - for both sexes and income quartiles. Each subgroup was modelled by linear mixture model. These model specifications enable predicting future salaries on individuals given past salaries on the same and other individuals. These predictions are potentially important for pension plan design and pension system development since nearly all decisions must be based on quantitative predictions of future (Plamondon et al 2002).

2   Individual salary data

The data was collected as a part of Finnish pension reform package in 2001-2002 (etk.fi / Pension reform 2005). The unique dataset consists of all those people who retired in 1998. Pension companies and institutions provided the data with individual yearly wages and working time, and some background information such as date of birth and sex. The working hours data is recorded daily and is of good quality. This enabled us to calculate effective wage by dividing wages with work effort. We operated on monthly wages. The yearly working hours is also used as an indicator for career integrity. The duration of the career is used as an explanatory variable for wages. The whole panel of data consists of a group of people, whose information we have from 1962 to 1998. However we restricted the data for our needs.

First restriction is that part time workers have been excluded. We limited our analysis to those who have been working their carriers on relatively stable jobs in private sector. Second restriction follows from the cross section effect that data consists of those retired in 1998. Naturally this peaks some cohorts. Here peak cohorts born in 1933 and in 1938.

Bearing these facts in mind we focused to cohorts born between 1933 and 1938 (see table I) and years from 1975 to 1994. These limitations mean that data was restricted to 2986 individuals. There are 57% women and 43% men in the data. Genders were analysed separately.

One should notice that the data here covers people aged 37 to 52 years. The rate of growth of wages was found to be strongly linked to the working hours in the case of lowest quartiles for male and two lowest quartiles for female. This is rather expected result since the working hours for lowest quartiles are low both for men and women.

Table 1. *The size of cohorts.*

Birth year	Number of individuals
1933	683
1934	133
1935	206
1936	249
1937	415
1938	1300

Since wages have tendency to follow overall economic performance, an indicator on the GDP was derived. The yearly change of consumer price deflated GDP was used here. The data were in 1998 level. The recession pattern is shown in figure 1. The GDP fell sharply

from 1990-1993. In the context of non-life insurance Hartwig et al (1997) studied worker's compensation and economic cycle. They found a strong association between economic growth and rising frequency, severity and loss ratio.

Figure 1. *Annual growth of real GDP (%).*
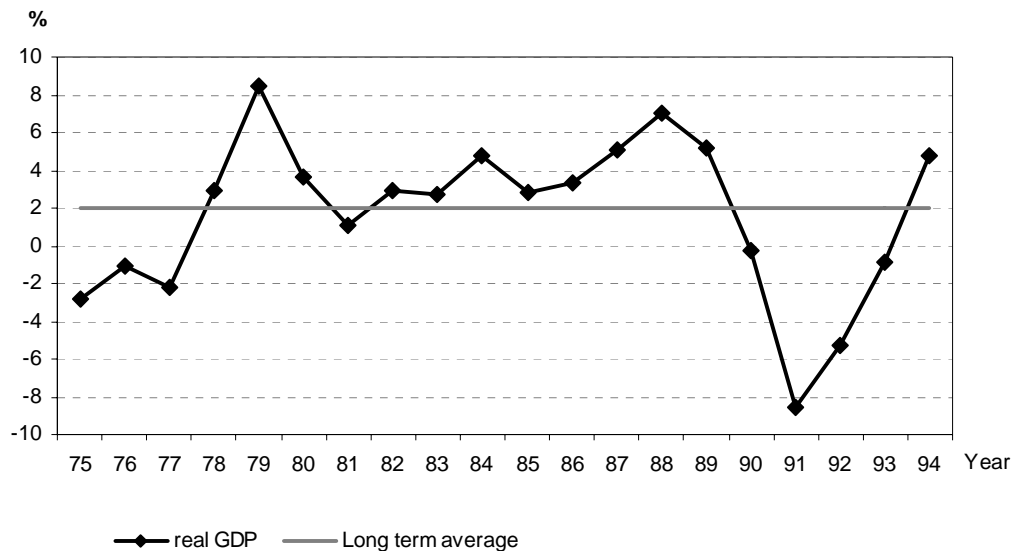
Finally the effective wages were deflated to 1998 level using average wages. The data were also real in terms of general earnings. It is known that the level of income has an effect on future wage growth (see f.ex. Murphy K. & Welch F.1990). We examined that by dividing male and female data to income quartiles.

The development of average wages for different male quartiles possesses a gentle negative trend which means that these cohorts fall behind general earnings development. In this respect the subgroups do not provide extra information. By contrast the percentual change varies from quartile to quartile (see figures 2 and 3). This demonstrates why separate model spefications might be needed for different subgroups.

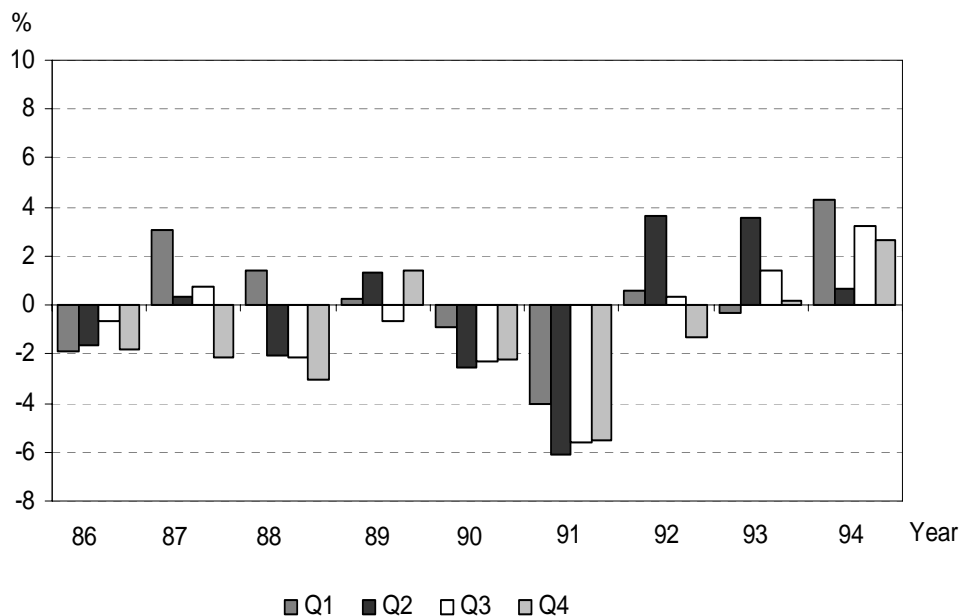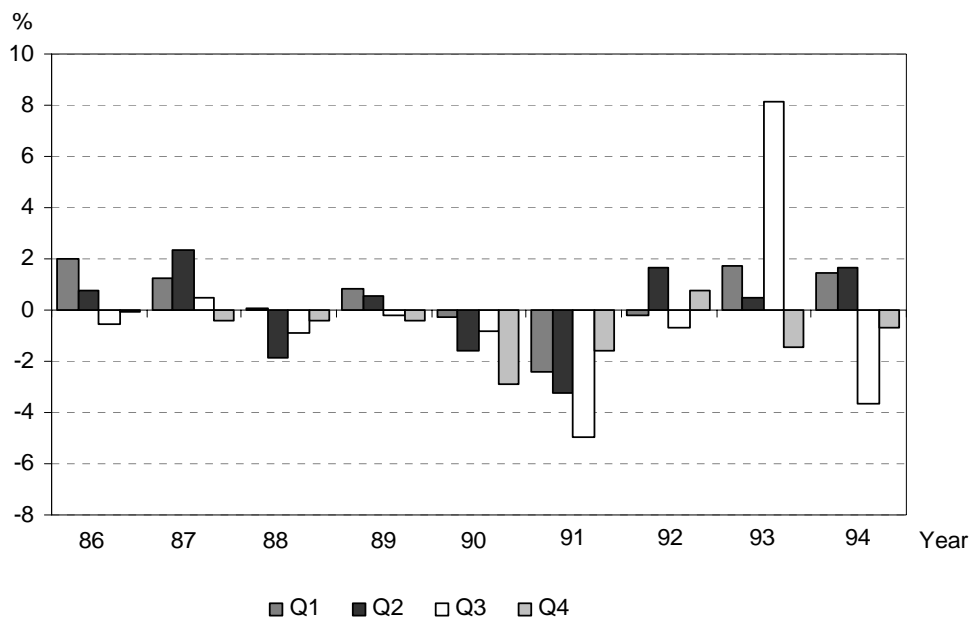*Figure 2.* Annual change (%) of mean wage in each quartile (men).[*]

Figure 3. Annual change (%) of mean wage in each quartile (women).[*]

* Q1=lowest, Q2=2nd lowest, Q3=2nd highest, Q4=highest.

3   Model

The used model is an extension of the basic linear model that allows some model parameters to be drawn from a probability distribution. This kind of model is usually referred to as mixed model since the model parameters contain both fixed and random effects. The following individual specific and economy wide variables are used

- z_{ij} age of an individual i at time j
- d_i duration of the career of an individual i
- b_j GDP at the time j

In economics the human capital model is dominant in explaining the variation of wages over life cycle. In short it means that young individuals invest heavily in their schooling and job training. This predicts that their wages will rise with age as their human capital increases, but wages will eventually fall with depreciating human capital. This theory has been tested in econometric studies using suitable age polynomials. Most specifications of the earnings function include linear and quadratic terms for age. The functional form may lead to overstated age-earnings profile (cf. e.g. Murphy and Welch 1990, and Johnston and Neumark 1996). Panel data allowed us to account for individual variation.

   We examined a number of specifications especially with age variable. It looks like that the usual age and age squared is not enough but also a cubic term is needed. The linear mixed model we ended up in our analysis is of the following form

$$y_{ij} = \beta_0 + u_{i0} + (\beta_1 + u_{i1})z_{ij} + \beta_2 z_{ij}^2 + \beta_3 z_{ij}^3 + \beta_4 d_i + \beta_5 b_j + \varepsilon_{ij},$$

The parameters $\beta_0 \beta_1,...,\beta_5$ are coefficients associated with the entire population. A cubic polynomial regression is fitted to age variable. Random parameters u_{i0} and u_{i1} are associated with an individual under consideration. They are used to model the subject individual specific development of a salary curve. The random errors ε_{ij} are independently and normally distributed. Here we assume that the joint distribution of u_{i0} and u_{i1} is

- multivariate normal with the expected value zero
- independent of the random errors ε_{ij}

The estimates of the model parameters are obtained by using the so-called REML (Restricted Maximum Likelihood) method, which maximizes the modified likelihood function of the normally distributed observations. For more details see e.g. Pinheiro, J. and Bates, D. (2000).

Parameters

The starting point for group specific modelling is that the same model is estimated with every group - for each quartile (Q1-Q4) and for both genders. The model specifications satisfied the usual statistical criteria. The estimation results are presented in Table 2. The results indicate that there is substantial variatiability both between wage groups and genders.

We also tested the approach where a separate model for every wage group was estimated. Somewhat surprisingly, the fitted model specifications, especially for men, did not differ much.

The wage quartile was reflected in the model specification in number of ways. The factors influencing the salary varied significantly between quartiles. Table 2 indicates that the quartiles have common features, but also some notable differences. A highly significant relationship was found between age (or its transformation) and wages in both genders, whereas the GDP growth was statistically significant only in the case of women's quartile 2. For women the significance and effect varied from quartile to quartile. For men age had positive effect and the square of age has negative effect on wages. Yet the quartile specific parameters differed significantly.

The duration of the career had significant effect only for the lowest wage quartiles. The link with wages in other groups was much weaker. This is rather expected result since lowest quarters have highest unemployment risk.

For men the cubic of age was significant factor for two lowest quartiles. For women it was the only factor contributing in every quartile. It had negative effect for women and positive for men (Q1-Q2).

Vartiainen (2002) studied gender wage differential in Finland. He found that the women with high wage predictors drag most behind their male colleagues. This is compatible with the differences between gender model specificatios for highest quartiles. The negative age effect among women can be partly explained by the fact that the wages were deflated using earnings and salary index.

Table 2. *Model specification for wage quartiles and genders.*

	Men	Women
Q1 (lowest)		
Age	77.9 ***	-25.9
Agesq	-2.1***	1.868**
Ageq	0.02***	-0.02**
Duration	77.5***	32.2***
GDP	-32.1	-104.3
Q2 (2nd lowest)		
Age	172.9***	-31.2***
Agesq	-4.6***	2.6***
Ageq	0.04***	-0.03***
Duration	8.4	11.07***
GDP	-20.1	171.0**
Q3 (2nd highest)		
Age	131.1***	-1.8
Agesq	-1.7*	2.0***
Ageq	0.002	-0.03***
Duration	0.52	5.8
GDP	-95.6	39.7
Q4 (highest)		
Age	221.0***	20.4
Agesq	-2.3***	1.4
Ageq		-0.02**
Duration	-58.7	30.0
GDP	- 19.6	0.5

***=1%, **=5% and *=10% siginificance level
Q4: Ageq caused singularity.

4   Predictions

"Don't never prophesy: If you prophesies right, ain't nobody going to remember and if you prophesies wrong, ain't nobody going to let you forget" - Mark Twain

When assessing the solvency of a scheme the pension experts are mainly interested in predicting average wages. Instead, in system development, individual variation in wages is essential - a high average wage does not guarantee adequate pension for all members of the group. Hence predicting individual salary growth is very important for planning purposes and risk assessment.

In the course of time the factors influencing the wages are subject to change and the estimated model specifications might lose their prediction power. One would expect that economic conditions are reflected in wages. However, the GDP was statistically significant factor only in one case (Q2 women). It is not unthinkable that rather than being dependent on every movement of the GDP, wages react only on recessions and booms.

When a time dependent model is formulated and fitted to the same data, then inferences and forecasts made from the fitted model will be biased and seriously over-optimistic. Hence genuine out-of-sample prediction tests are needed (cf. e.g. Chatfield 2001). We test the prediction of model specifications in two very different time periods. From 1986 to 1990 was "business as usual" period and from 1991 to 1994 was a deep depression period. This kind of exceptional slump in economy is a heavy test on prediction.

The model specifications presented in Table 2 were the basis for our prediction task. We show that the fitted models can be used to produce individual wage predictions as well as standard errors of estimates.

First predictions were needed for the exogenous variable GDP - age can be predicted without problem and we assumed constant duaration. Holt-Winters predictions for GDP were made. It resulted in 2.3–2.5% annual increase in 1986-89 and 2.6–2.7% increase in 1991-94. Figures 4 and 5 illustrate the individual variation in predictions for wages.

Figure 4. *Examples of individual wage predictions and actual values (men).*
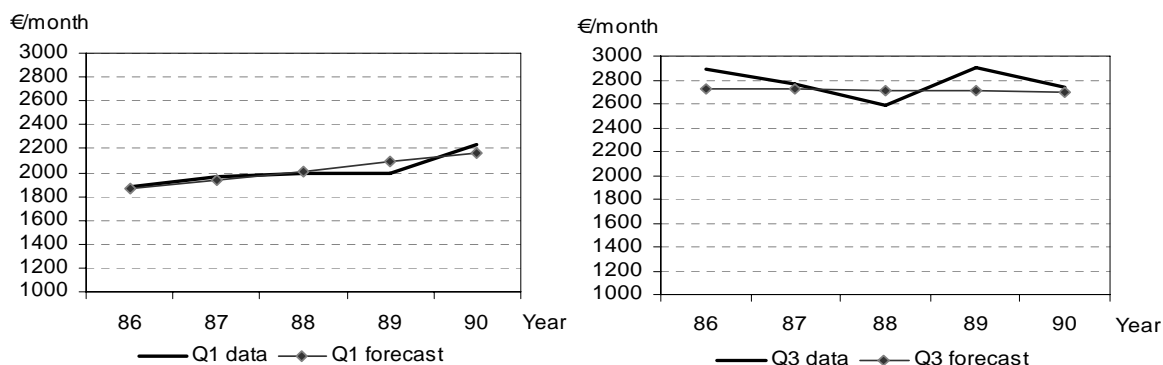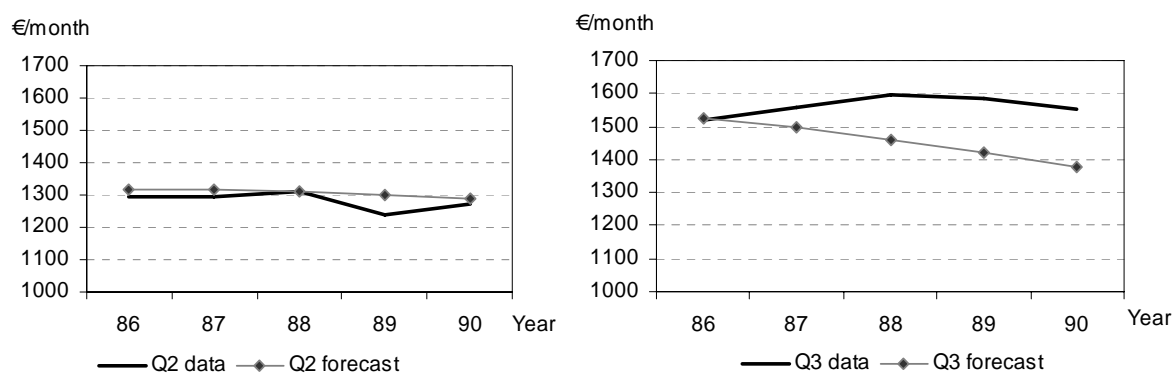

Figure 5. *Examples of individual wage predictions and actual values (women).*

Next we consider group level predictions. Figures 6–9 facilitate the prediction comparison between wage quarters and periods. The errors are presented as percentage. The figures reveal several interesting features.

Under normal period (1986–1990) the models slightly overestimate the real wage development. The middle quartiles (Q2 and Q3) are well predicted. The first and fourth quartiles are rather more challeng-ing to predict. This holds for both men and women.

Under recession period the prediction error increases with time. This in no surprise since the Finnish recession was unexpectedly prolonged. The business cycle is certainly a factor affecting wage risk. The predictions seriously overestimate low-paid wages. In fact, the low-paid employees have difficulties on maintaining their wages. Low wage is hence associated with high wage risk. The model works very well for Q3, whose error is minimal for both men and women. The difference between men and women seems to be in high-paid group (Q4), where the errors have different directions.

Generally one could say that these models work better for women. The figures show that the prediction errors for men are considerably larger than for women. This holds doth

during recession period and during normal period. The low-paid quartile is the most difficult to predict. One reason for this could be un-employment, which is more common in this quartile than in others. The figures reveal no clear bias in predictions. Related to the sign of the error, there are two distinct types of development. Most wages are somewhat overestimated, but for example high-paid men are slightly underestimated (Q4 in figure 7). We think these results are an illustration that studying averages is not enough.

Figure 6. *Absolute prediction error as percentage of mean wage 1986-1990 (men).*

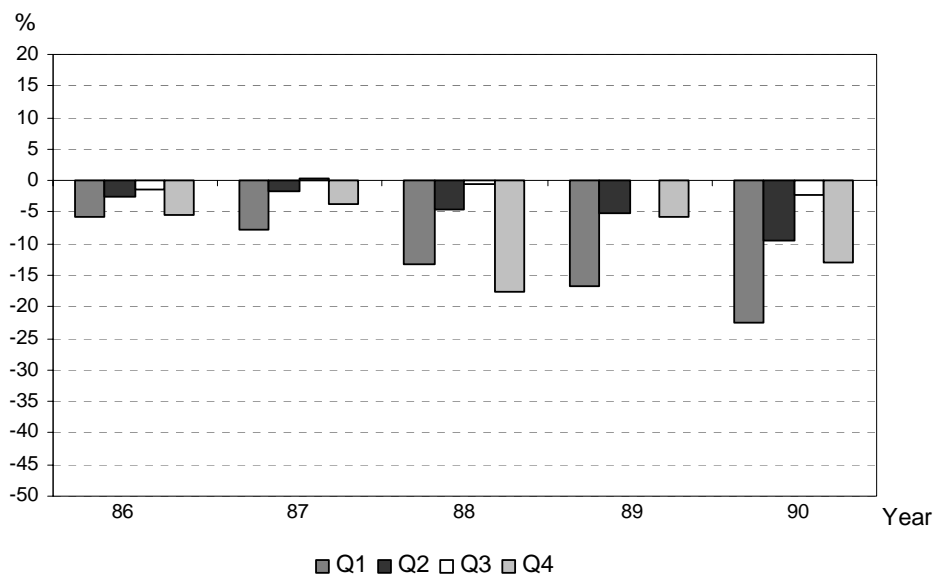***Figure 7.*** *Absolute prediction error as percentage of mean wage 1991-1994 (men).*
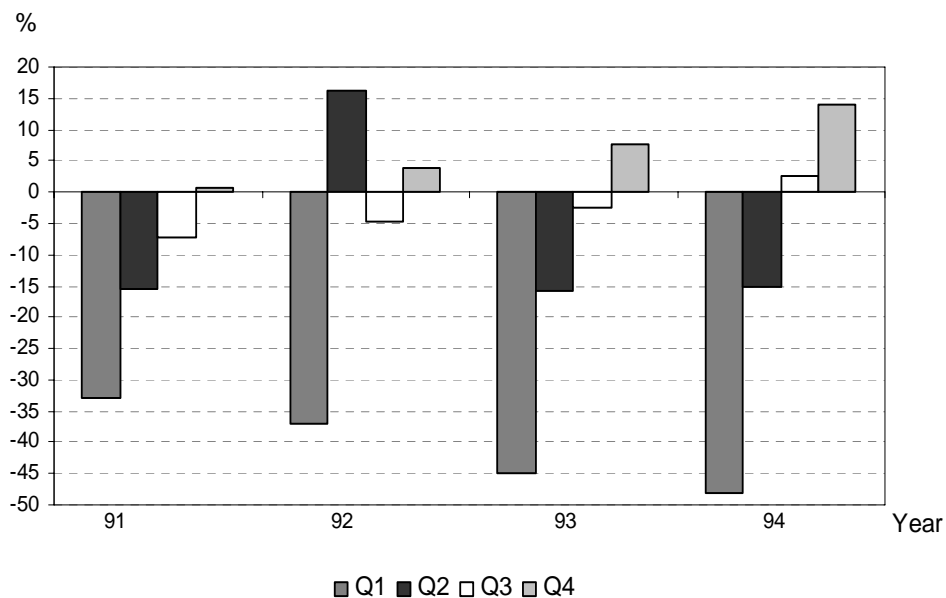
Figure 8. *Absolute prediction error as percentage of mean wage 1986-1990 (women).*
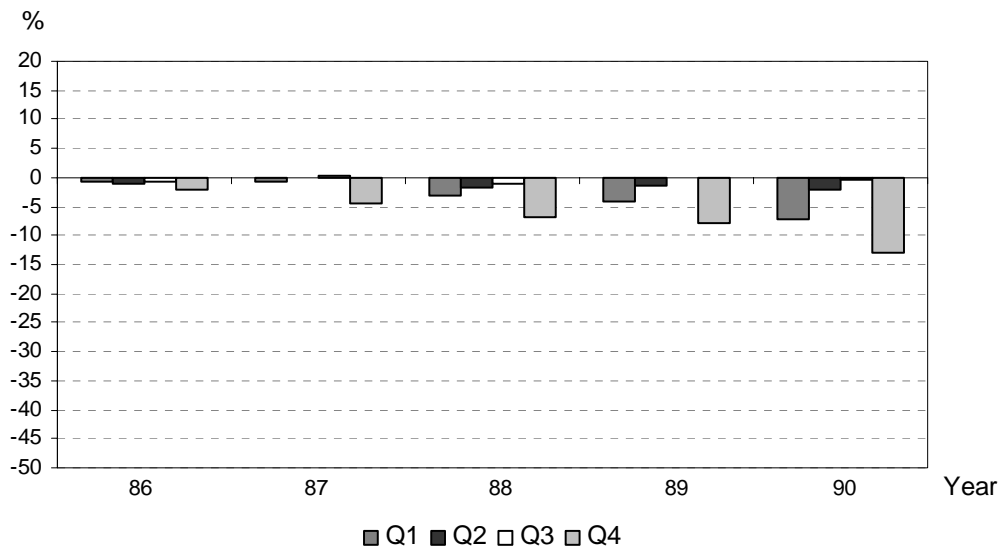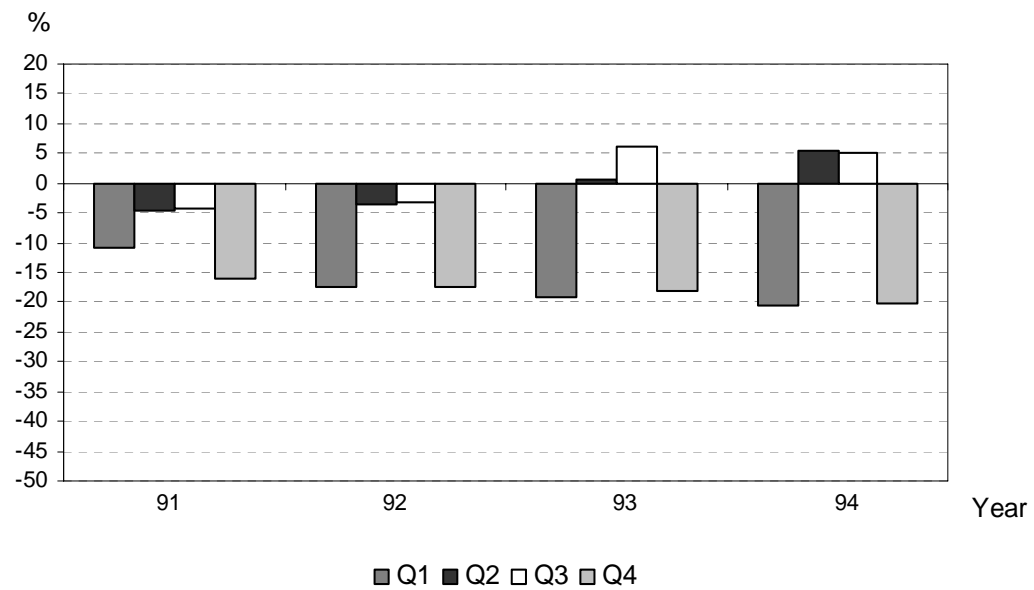

Figure 9. *Absolute prediction error as percentage of mean wage 1991-1994 (women).*

5   Conclusions

We have focused on the problem of modelling and predicting future salaries on individuals given past salaries on the same and other individuals. The cornerstone of this study was a dataset of individual careers, obtained from Finnish pension companies and institutions. Individual specific and economy wide variables have been used when a linear mixed model have been estimated for wage quarters and both genders. The objective has not been to present all the details of the Finnish data, but to demonstrate the main features and opportunities of the panel models combined with a high quality data.

Modelling is never an end itself. Here the interest has been in predicting for planning purposes. Wage forecasts may provide critical information about the pension to come and indivual risk. We have studied two fundamentally different out-of-sample periods of the economy of Finland - normal growth period 1986-1990 and deep depression period 1991-1994.

The model specifications and prediction results allow for the following general conclusions:

- The wage formation seems to be essentially different in different wage quartiles. Better forecast may be obtained by using quarter specific models.
- Individual variation among wage quartile is large and an important risk factor. In our opinion there is an urgent need to model it for pension planing purposes.
- Individual career forecasts are invaluable for pension system planners.
- The workers in the lowest quantile have difficulty in maintaining their wages in periods of deep depression. In this study the link with wages in other groups is much weaker.
- The prediction errors for the middle wage quarters seem to be considerably smaller than for the low and high wage salary groups. There is some indication that the middle quarters can be predicted quite accurately for several years ahead.
- For severe economic situations, scenario testing may still be invaluable additional tool.

References

Bone C. & Mitchell O. (1997). Building Better retirement Income Models. *North American Actuarial Journal*, **1**, 1-12.

Carrier J. & Shand K. (1998). New Salary Functions for Pension Valuations. *North American Actuarial Journal*, **3**, 18-28.

Chatfield C. (2001). *Time-series forecasting*, Chapman & Hall, London.

Frees E. & Young V. & Luo Y (2001). Case Studies Using Panel Data Models. *North American Actuarial Journal*, **5**, 24-42.

Graddy K. & Pistaferri L. (2000). Wage differences by gender: evidence from recently graduated MBAs. *Oxford Bulletin of Economics and Statistics*, **62**, 837-854.

Hartwig R., Kahley W. & Restrepo T. & Retterath R (1997). Worker's Compensation and Economic Cycles: A Longitudial Approach. *Proceedings of the Casualty Actuaries Society*, LXXXIV, 660-700.

Johnson R. W. & Neumark D. (1996). Wage Declines among Older Men. *The Review of Economics and Statistics*, **78**, 740-74.

Koskinen L. & Pukkila T. (2002). Credit insurance - Risk caused by the downturns of the national economy. 27th *International Congress of Actuaries,* Cancun, Mexico.

Murphy K. & Welch F.(1990). Empirical Age-Earnings Profiles. *Journal of Labor Economics,* **8**, 202-229.

Nummi T. & Möttönen J. (2000). On the Analysis of Multivariate Growth Curves. *Metrica* 52, 77-89.

Pinheiro, J. & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*, Springer, New York, 2000.

Plamondon, P.& Drouin A. & Binet G. & Cichon M. & McGillivary W. & Bedard M. & Perez-Montas H. (2002): *Actuarial practice in social security*, ILO and ISSA, Geneva.

Shiller R. (2003). *New Financial* Order, Princeton Press, Princeton and Qxford.

Vartiainen J. (2002), Gender wage differentials in the Finnish labour market. *Labour Institute for Economic Research*, Discussion papers 179.

Appendix

Appendix A1. *The parameters of the model: MEN.*

Wage Quartile 1.

	Value	Std.Error	DF	t-value	p-value
age	462.8131	90.1837	2581	5.131895	0.0000
age^2	-12.7797	3.9397	2581	-3.243829	0.0012
age^3	0.1172	0.0447	2581	2.619652	0.0089
duration	460.1976	34.9671	287	13.160862	0.0000
GDP	-232.3819	838.8313	2581	-0.277031	0.7818

Wage Quartile 2.

	Value	Std.Error	DF	t-value	p-value
age	1027.2990	97.7415	2808	10.510363	0.0000
age^2	-27.1411	4.2355	2808	-6.408031	0.0000
age^3	0.2314	0.0481	2808	4.807190	0.0000
duration	50.0082	45.9532	287	1.088243	0.2774
GDP	-119.3321	869.6535	2808	-0.137218	0.8909

Wage Quartile 3.

	Value	Std.Error	DF	t-value	p-value
age	779.0120	134.8131	2829	5.778461	0.0000
age^2	-9.9758	5.5587	2829	-1.794635	0.0728
age^3	0.0149	0.0626	2829	0.238290	0.8117
duration	3.1110	83.2749	286	0.037358	0.9702
GDP	-567.6486	1121.8180	2829	-0.506008	0.6129

Wage Quartile 4. (Cubic term caused singularity.)

	Value	Std.Error	DF	t-value	p-value
age	1312.5395	208.1829	2714	6.304742	0.0000
age^2	-13.9035	2.4935	2714	-5.575946	0.0000
duration	-348.5467	327.7432	287	-1.063475	0.2885
GDP	-420.3728	2672.4430	2714	-0.157299	0.8750

Appendix A2. The parameters of the model: WOMEN.

Wage Quartile 1.

	Value	Std.Error	DF	t-value	p-value
age	-154.0710	98.8757	2773	-1.558229	0.1193
age^2	11.1535	4.3324	2773	2.574430	0.0101
age^3	-0.1205	0.0484	2773	-2.488134	0.0129
duration	192.0661	21.6254	400	8.881493	0.0000
GDP	-619.2534	840.5685	2773	-0.736708	0.4614

Wage Quartile 2.

	Value	Std.Error	DF	t-value	p-value
age	-185.1191	56.5137	3757	-3.275653	0.0011
age^2	15.7225	2.5023	3757	6.283236	0.0000
age^3	-0.1871	0.0286	3757	-6.547215	0.0000
duration	65.7507	17.0701	400	3.851798	0.0001
GDP	1015.5011	504.0707	3757	2.014601	0.0440

Wage Quartile 3.

	Value	Std.Error	DF	t-value	p-value
age	-10.74638	64.2298	3899	-0.167311	0.8671
age^2	11.99497	2.7707	3899	4.329160	0.0000
age^3	-0.16084	0.0314	3899	-5.114904	0.0000
duration	34.51381	31.3520	400	1.100850	0.2716
GDP	234.92885	554.5775	3899	0.423618	0.6719

Wage Quartile 4.

	Value	Std.Error	DF	t-value	p-value
age	121.08294	153.2684	3947	0.7900058	0.4296
age^2	8.40591	5.2535	3947	1.6000451	0.1097
age^3	-0.11881	0.0552	3947	-2.1513045	0.0315
duration	178.54454	152.2858	400	1.1724304	0.2417
GDP	-116.30219	926.8640	3947	-0.1254792	0.9002

Appendix B1. MEN.

Mean prediction errors for years 1986-1990 ("Business as ususal").

Year	Wage Quartile 1	Wage Quartile 2	Wage Quartile 3	Wage Quartile 4
1986	-601.6218	-340.1501	-207.402	-1357.447
1987	-852.6275	-222.4521	48.86261	-894.9024
1988	-1458.492	-570.1517	-89.82407	-4086.366
1989	-1835.801	-654.6165	5.140832	-1357.466
1990	-2476.69	-1196.209	-318.8540	-2966.635

Mean prediction errors for years 1991-1994 (Depression).

Year	Wage Quartile 1	Wage Quartile 2	Wage Quartile 3	Wage Quartile 4
1991	-3602.322	-1926.827	-1065.834	146.1777
1992	-3881.002	-1907.422	-649.9709	836.9852
1993	-4764.451	-1904.471	-335.0946	1616.157
1994	-5053.863	-1903.357	366.421	3002.711

Appendix B2. WOMEN.

Mean prediction errors for years 1986-1990 ("Business as ususal" period).

Year	Wage Quartile 1	Wage Quartile 2	Wage Quartile 3	Wage Quartile 4
1986	-51.53458	-96.575	-82.84091	-276.2793
1987	-48.0526	-18.10262	18.66322	-630.6483
1988	-228.5557	-152.2645	-99.54979	-913.6822
1989	-308.7045	-128.44	-21.73826	-1084.961
1990	-543.185	-191.7194	-41.10984	-1711.855

Mean prediction errors for years 1991-1994 (Depression period).

Year	Wage Quartile 1	Wage Quartile 2	Wage Quartile 3	Wage Quartile 4
1991	-837.1783	-405.4364	-435.1827	-2124.923
1992	-1316.501	-315.8818	-330.8556	-2248.06
1993	-1438.329	57.46457	585.2582	-2378.899
1994	-1587.596	473.9739	530.1294	-2585.978