WORKER FLOWS AND OCCUPATIONS IN THE CPS 1976-2010: A FRAMEWORK FOR ADJUSTING THE DATA

Alexandre Ounnas

DISCUSSION PAPER | 2020 / 08

UCLouvain

IRES | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

Worker Flows and Occupations in the CPS 1976-2010: A Framework for Adjusting the Data

Alexandre OUNNAS*

February, 2020

Abstract

This paper proposes a framework for adjusting issues affecting series of stocks and gross flows by occupations obtained from the Current Population Survey (CPS). Using data over the period 1976-2010 and the occupation classification of Autor and Dorn (2013) to rank occupations between high, medium and low skills, I adjust series for the 1994 redesign of the CPS questionnaire, changes in occupational classification and revisions in the size and composition of the US population. In a second step, I correct flow rates for the Time Aggregation bias. Due to constraints specific to flow rates by occupation, the correction proposed by Shimer (2012) and Elsby et al. (2015) cannot be applied. As a result, I use the bayesian estimation method of Bladt and Sørensen (2005).

Keywords: Current Population Survey; occupations; unobserved component model; time aggregation bias.
JEL-codes: C5, C8, J6.

*IRES - Université catholique de Louvain - alexandre.ounnas@uclouvain.be

1 Introduction

The macro-labor literature highlights the key role played by labor market flows in driving unemployment fluctuations (Darby et al. (1986), Blanchard and Diamond (1990), Fujita and Ramey (2009)). One of the leading data source to study these transitions is the monthly Current Population Survey (CPS) which is a representative survey of U.S households. Due to its sample size and the possibility it provides to follow individuals for consecutive months, the CPS is particularly well suited for the study of gross labor market flows.

The CPS also provides information on the occupations of (un)employed workers which offers the opportunity to analyze occupational mobility (Kambourov and Manovskii (2008), Kambourov and Manovskii (2009)). These occupations codes are the basis for the classification built by Autor and Dorn (2013) and a strand of the macro-labor literature has recently taken an interest in analyzing the impact of *Job Polarization* on labor market stocks (Jaimovich and Siu (2012), Foote and Ryan (2015)) and flows (Cortes et al. (2016)).

However, there are shortcomings in using the CPS, due to the fact it suffers from measurement errors. A well known example of these measurement problems regards unemployment and inactivity (Abowd and Zellner (1985), Poterba and Summers (1986)). Some evidence also suggests that these measurement errors affect the coding and assignment of occupations (Kambourov and Manovskii (2013)). Moreover, concepts, methods and definitions have been updated over the years, which created inconsistencies and breaks in many of the time series. More specifically, the CPS redesign of 1994, updates in occupational classifications and revision in the size and composition of the US population from the Census Bureau can generate issues for analysis of occupational data over extended periods of time. Furthermore, the measurement of gross flows suffers from a Time Aggregation problem stemming from the discrete nature of data collection (Perry et al. (1972), Shimer (2012)). This issue leads to an underestimation of flows from and to unemployment and a overestimation of flows between employment and inactivity.

This paper proposes a framework for adjusting CPS series of stocks and gross flows for all the above mentioned problems with a particular focus on series with an occupational dimension. In order to do so, I use CPS data for the 1976-2010 time period and the panel of occupations developed by Autor and Dorn (2013) that allows for the classification of occupations between high, middle and low skill.

In a first stage, I use an Unobserved Component model (Harvey (1990), Durbin and Koopman (2012)) to deseasonalize and adjust series from effects of the 1994 CPS redesign, the classification changes of 1976-1982 and 2003-2010 and the population updates of 1976-79, 1990 and 2003. The 1994 CPS redesign, in particular, is found to have substantial effects on unemployment stocks and flow series through a modification in the definition and measurement of *New Unemployed Entrants*, i.e. unemployed workers entering the labor market for the first time. The estimates for classification and population changes also lead to significant adjustments in most time series.

In a second step, I carry out the correction for Time Aggregation. Shimer (2012) has shown how to retrieve instantaneous transition rates from discrete time transitions. Due to restrictions specific to gross flow series by occupations, I use the Bayesian procedure of Bladt and Sørensen (2005) to estimate these hazard rates and compute adjusted transition probabilities. The estimates are consistent with those reported by Elsby et al. (2015) and I can additionally provide results in terms of flows rates by occupations.

The paper is organized as follows. In Section 2, I present the data and the occupation classification and review in more detail the problems affecting CPS time series. Section 3 focuses on the seasonal adjustment and the adjustments implemented for the 1994 redesign, classification changes and population revisions. This section presents the econometric set-up and the estimation results. Finally, the framework for correcting the Time Aggregation bias is presented and discussed in Section 4.

2 Data

2.1 Data Description and Skill Classification

The Current Population Survey (CPS) and the occupation-task classification developed by Autor and Dorn (2013) are the 2 main sources of data used in this work. The CPS is a representative survey of US households containing labor force information at the individual level. It is used to compute official statistics such as the monthly unemployment rate, and is particularly attractive due to its large sample size, its availability and the possibility to follow individuals for 4 straight months.[1]

I use data from 1976 to 2010 and restrict the sample to individuals aged 16 and over. In addition to basic labor force and demographic questions, the CPS asks interviewed workers to report -if they are employed- their current occupations, or -if they are unemployed- the last occupation in which they were employed. Unemployed who enter the labor market for the first time (*New unemployed entrants* in the CPS) and individuals for which occupation codes are missing are excluded from my analysis. Individuals outside of the labor force are not asked the occupation question except when they are in the outgoing rotation group (individuals in their fourth or eighth month of interview who exit the following month) which only represents one fourth of the total sample (See Cortes et al. (2016)).

I follow Autor and Dorn (2013) and use the 3-digits (or detailed) occupation codes provided by the CPS to classify occupations. Every ten years or so, the Bureau of Labor Statistics (BLS) adjusts the occupation classification to reflect structural changes in the occupational composition and account, for example, for the creation of new occupations.[2] These updates have to be accounted for to avoid breaks in series and Autor and Dorn (2013) have built a crosswalk (with the 1992-2002 classification as reference) that allows to obtain a fairly balanced panel of occupations across the 1976-2010 period. To rank occupations, Autor and Dorn (2013) build a measure (*the task value*) of the task content of each detailed occupation in their panel with respect to four dimensions: abstract versus cognitive and routine versus non routine. They subsequently aggregate detailed occupations into the 6 groups that can be found in the first column of Table 1. In this paper, I further aggregate their grouping into 3 categories: High skill occupations which require a high level of cognitive or abstract tasks, middle skill occupations in which tasks are primarily repetitive (routine) and low skill occupations which entail mostly manual tasks.

It is worth mentioning that Autor (2013) defines tasks as "a unit of work activity that produces output" while skills are "a stock of capabilities for performing various tasks". There exists an imperfect mapping between skills and tasks (e.g. cognitive tasks usually require higher skills) and it should be clear that the use of the terms high, middle and low skill stand for cognitive, routine and manual intensive occupations. Furthermore, the skill level assigned to all employed and unemployed workers in the sample depends only on the skill level (task contents) of occupations. In other words, an individual is classified as high skill because she works or used to work in a high skill occupation. This 3-groups classification implies that there are 3 employment states (E^h, E^m and E^l), 3 unemployment states (U^h, U^m and U^l) and inactivity I. These 7 states constitute the population stocks and there are 49 possible gross flows in total.[3]

[1] The CPS actually allows to follow individuals for 16 months with an 8 months gap in between (4-8-4).

[2] A new classification was introduced in January 2011 which explains why the sample is restricted to December 2010.

[3] The classification used in this work is slightly different than the one used by Jaimovich and Siu (2012) who consider four different occupational groups. They further disentangle middle skill occupations between routine cognitive and routine manual occupations. Contrary to the classification used in this paper, the *Transport/construct/mech/mining/farm* occupation group would be considered as middle skill. This 4 groups classification is used in most of the literature interested in interacting job polarization and jobless recoveries. However, the *routine task intensity index* of Autor and Dorn (not reported in Table 1) also suggests that the group *Transport/construct/mech/mining/farm* requires mostly manual tasks. I choose to only consider 3 groups as it is enough to capture the *Job Polarization* trend and adding an additional skill level results in multiplying the labor market states and flows (4 skill levels would imply 9 states and 81 flow rates).

	Abstract tasks	Routine tasks	Manual tasks	skill level
Managers/prof/tech/finance/public safety	+	-	-	*high*
Production/craft	+	+	-	*middle*
Transport/construct/mech/mining/farm	-	+	+	*low*
Machine operators/assemblers	-	+	+	*middle*
Clerical/retail sales	-	+	-	*middle*
Service occupations	-	-	+	*low*

The first 4 columns of this table are taken from Table 2 of Autor and Dorn (2013). A "+" indicates that the task value of a given occupation-group is above the task value averaged over all occupation-groups. The shaded cells give the maximum task value for each occupation-group. I assign a skill level to each groups of occupations according to whether the task value of the occupation-group she belongs to is more abstract (high skill), routine (middle skill) or manual (low skill).

Table 1: Skill classification

In order to compute gross flows, monthly files have to be matched for 2 consecutive months. The rotating structure of the CPS implies that one fourth of the sample exit the survey in the following month (the *outgoing rotation group*). Therefore only three fourth of the original sample can be matched across 2 months. Individuals are linked longitudinally by using the 2 households identifiers and the person's line number. I follow Madrian and Lefgren (1999) and apply a criterion that checks for identical race, sex and age of matches across 2 consecutive months.

The CPS data suffer from various issues that need to be addressed before computing stocks and flow rates. Once these adjustments (presented in Section 2.2) have been implemented, information on individual characteristics such as age or education is lost. Hence, I briefly present some descriptive statistics based on flow rates computed from the uncorrected data for the period 1994-2010. The sample is restricted to this specific period to avoid the break implied by the 1994 redesign of the CPS questionnaire. Descriptive statistics for stocks are computed from the unmatched data. Monthly flow rates, p_t^{ij}, at time t from state i to state j are obtained by summing up workers transitioning from i in $t-1$ to j in t.[4] Doing so allows to obtain gross flows which are then divided by the stock of individuals in state i in $t-1$. For instance, the flow rate from high skill unemployment U^h to high skill employment E^h at time t is given by

$$p_t^{U^h E^h} = \frac{U^h E_t^h}{U_{t-1}^h} \tag{1}$$

where $U^h E_t^h$ is the number of individuals moving from U^h in $t-1$ to E^h in t (gross flow) and U_{t-1}^h is the stock of high skill unemployed individuals in $t-1$. This expression is the maximum likelihood estimator for a multinomial distribution with 7 possible outcomes (the states E^h, E^m,...).
Tables 2, 3 and 4 display these descriptive statistics for stocks and flow rates. Tables 3 and 4 focus on monthly average flow rates from unemployment. Average flow rates from employment and inactivity can be found in Appendix A.1.1.

From Table 2, it is worth pointing the increasing relationship between age and the skill level of the occupations. We can observe a higher share of older workers (compared to the entire sample) in high skill occupations and a higher share of younger workers in low skill occupations.

[4] Applying weights provided by the CPS. I follow Shimer (2012) and average individual final weights across 2 consecutive months.

	Employment			Unemployment			Inactivity	
	E^h	E^m	E^l	U^h	U^m	U^l	I	*Total*
Age								
<25	18.9	36.6	38.9	18.5	40.4	39.6	42.1	36.3
25-50	57.7	47.4	45.9	60.1	47.0	49.1	41.5	47.4
>50	23.4	16.0	15.1	21.4	12.6	11.3	16.4	16.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Marit. Status								
married	55.3	42.4	37.6	50.0	34.2	31.8	37.9	40.3
not married	44.7	57.6	62.4	50.0	65.8	68.2	62.1	59.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Race								
White	72.3	64.1	61.2	71.7	56.6	55.2	62.1	62.8
Hisp.	8.5	14.0	17.8	8.0	15.2	19.0	14.7	14.7
Black	12.1	15.5	15.5	13.0	21.9	20.3	16.4	16.2
Others	7.0	6.4	5.6	7.3	6.4	5.4	6.9	6.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Educ.								
< HS	5.4	18.8	31.1	5.5	23.5	36.2	28.8	24.2
HS	20.9	34.0	36.2	18.1	37.4	38.7	28.5	31.4
> HS	73.7	47.1	32.7	76.4	39.0	25.2	42.6	44.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
U. type								
Laid-off				21.5	22.2	31.4		26.6
Job loser				57.5	56.8	51.6		54.3
Job leaver				21.0	21.0	17.1		19.1
Total				100.0	100.0	100.0		100.0
U. duration								
<6 m				77.6	79.4	80.8		79.8
6-12 m.				11.6	10.7	9.7		10.4
>12 m				10.8	9.9	9.5		9.9
Total				100.0	100.0	100.0		100.0
N	128,842	167,826	212,318	38,447	70,513	105,464	287,533	1,010,943

Populations stocks for various characteristics computed from the raw CPS files for the periods Jan. 1994 to Nov. 2010. All observations are weighted using (final) weights provided by the CPS. The last row gives the total number of (unweighted) observations. The last column displays the share of total population with a given characteristic. Exceptions are for the last rows (U.type and duration) where percentages are expressed in terms of the total population in unemployment. Figures for unemployment exclude *New Unemployed Entrants*. HS stands for High School. Laid-off worker are expected to be recalled within a month, Job loser correspond to the category *other job loser* in the CPS and Job leaver are voluntary quits.

Table 2: Average Stocks over the period January 1994 - November 2010

The same holds in regard to the skill level of occupations and the educational attainment although, one third of individuals in low skill occupation hold a postsecondary degree.[5] Furthermore, low skill unemployment is characterized by a much higher share of laid-off workers whereas high and middle skill occupations have a higher share of job losers and leavers.

Tables 3 and 4 give an idea of the extent of the transitions between occupations of different skill levels. These transitions are not negligible as they represent around 11% of exits from high skill unemployment to employment ($p^{U^h E^m} + p^{U^h E^l}$). For middle and low skill unemployed individuals,

[5]Educational attainment is usually considered as a proxy for skills and this observation can therefore be linked to the comment made previously on the difference between skills and tasks.

these transitions represent 10% and 7% of exits from unemployment respectively. We also observe that young and less educated workers (compared to older and more educated workers) have higher transition rates to lower skilled occupations ($p^{U^hE^m}, p^{U^hE^l}$...) while the opposite applies to upward transitions ($p^{U^mE^h}, p^{U^lE^m}$...).

	FS.	Age			Marit. status		Race			
		<25	**25-50**	**50<**	**M.**	**NM.**	**W.**	**H.**	**B.**	**O.**
High Skill										
$p^{U^hE^h}$	16.17	14.51	17.13	14.64	17.89	14.44	17.38	13.25	11.70	14.66
$p^{U^hE^m}$	6.37	11.28	6.01	4.14	5.43	7.30	6.34	7.69	6.19	5.44
$p^{U^hE^l}$	4.83	9.72	4.42	2.72	3.75	5.91	4.67	6.34	5.45	3.65
$p^{U^hU^h}$	53.87	40.26	55.43	58.00	54.11	53.60	54.10	51.30	54.22	53.55
$p^{U^hU^m}$	0.68	0.82	0.67	0.58	0.58	0.77	0.59	1.14	0.93	0.71
$p^{U^hU^l}$	0.67	1.06	0.63	0.51	0.52	0.81	0.56	1.06	1.06	0.48
p^{U^hI}	17.42	22.35	15.70	19.41	17.72	17.16	16.35	19.22	20.45	21.51
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Middle Skill										
$p^{U^mE^h}$	2.80	2.31	3.16	2.68	3.44	2.46	3.35	1.94	1.96	2.95
$p^{U^mE^m}$	14.98	15.83	14.55	14.26	16.08	14.38	16.53	14.63	11.57	14.33
$p^{U^mE^l}$	7.59	10.20	6.57	4.48	6.26	8.28	7.98	8.56	6.45	5.99
$p^{U^mU^h}$	0.40	0.19	0.52	0.53	0.50	0.36	0.45	0.33	0.33	0.44
$p^{U^mU^m}$	50.20	43.01	53.74	55.02	51.07	49.72	50.21	47.32	52.27	48.65
$p^{U^mU^l}$	1.08	1.05	1.18	0.81	0.94	1.16	0.87	1.48	1.36	1.05
p^{U^mI}	22.96	27.41	20.28	22.22	21.71	23.65	20.63	25.74	26.05	26.60
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Low Skill										
$p^{U^lE^h}$	1.60	1.50	1.71	1.42	1.84	1.49	1.95	1.13	1.16	1.53
$p^{U^lE^m}$	5.23	7.35	4.19	2.90	4.53	5.55	5.93	4.47	4.15	5.06
$p^{U^lE^l}$	22.12	20.60	23.20	22.13	25.20	20.67	23.60	26.56	15.38	19.94
$p^{U^lU^h}$	0.22	0.12	0.26	0.34	0.26	0.20	0.24	0.17	0.20	0.25
$p^{U^lU^m}$	0.78	0.77	0.79	0.72	0.70	0.82	0.67	0.90	0.98	0.74
$p^{U^lU^l}$	47.89	43.28	50.45	50.80	48.48	47.62	47.36	44.38	52.12	47.04
p^{U^lI}	22.15	26.38	19.40	21.69	19.00	23.65	20.25	22.38	26.01	25.43
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Flow rates from unemployment $p^{U\cdot X}$ expressed in percentage, computed from the raw matched CPS files and averaged over the period Feb. 1994 to Dec. 2010. All observations are weighted using weights provided by the CPS. The sample for unemployed excludes *New Unemployed Entrants*. FS stands for full sample, M for Married and NM for not married. The races W., H., B. and O. stands for white, hispanic, black and others.

Table 3: Average Flow Rates from Unemployment over the period February 1994 - November 2010 (1)

Finally, it can be interesting to comment on transitions between unemployment states with different skill level (e.g. U^hU^m,U^hU^l ...). In theory, these transitions should not be observed and they can be interpreted in 2 ways: they either capture misreporting/misassignment of occupation codes, or a Time Aggregation problem in the form of an employment spell in a different occupation between 2

surveys which is not recorded by the CPS.[6] In support of the latter view, it is worth mentioning that these transition rates are always higher to states of a lower skill level (i.e The U^mU^l flow rate is twice as big as the U^mU^h flow rate, on average) which can suggest a very short employment spell in a lower skilled occupation. In Table 4, it is also shown that these flow rates are only marginally affected when removing individuals in the rotation group (1rst or 5th month of interview, see Fallick and Fleischman (2004)) which are known to be more likely to make mistakes when answering labor force questions. The Time Aggregation correction implemented in Section 4 brings these flows to 0 but this correction assumes that all of these transitions actually missed an intervening spell in employment. If these transitions were merely the results of mistakes, the importance of these misreported flows should be limited as they represent, respectively, 1.35% (0.68+0.67), 1.48% (1.08+0.40) and 1.00%(0.22+0.78) of total transitions for high, middle and low skill unemployment. It is, however, likely that these flows capture both mistakes from respondents/interviewers, and missed transitions.

	FS	Educ.			U. type			U. duration			Rota. Gr.	
		<HS	**HS**	**HS<**	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**1**	**0**
High Skill												
$p^{U^hE^h}$	16.17	7.38	9.47	18.30	40.84	12.95	15.83	18.35	10.72	8.34	15.76	16.38
$p^{U^hE^m}$	6.37	6.87	8.35	5.87	5.75	5.29	8.75	7.07	5.15	3.08	6.60	6.26
$p^{U^hE^l}$	4.83	11.06	7.96	3.70	6.15	3.85	5.86	5.35	3.65	2.80	4.93	4.78
$p^{U^hU^h}$	53.87	44.05	52.92	54.69	28.88	64.58	52.06	51.17	61.18	61.63	52.72	54.45
$p^{U^hU^m}$	0.68	0.82	0.87	0.62	1.87	0.56	0.40	0.70	0.58	0.68	0.73	0.65
$p^{U^hU^l}$	0.67	1.58	1.25	0.47	1.89	0.53	0.35	0.69	0.69	0.61	0.75	0.62
p^{U^hI}	17.42	28.24	19.18	16.36	14.61	12.25	16.75	16.67	18.03	22.87	18.51	16.87
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Middle Skill												
$p^{U^mE^h}$	2.80	0.67	1.64	5.20	2.96	2.61	3.71	3.04	2.27	1.59	2.70	2.85
$p^{U^mE^m}$	14.98	12.61	15.27	16.12	35.10	11.81	15.56	16.91	9.39	6.45	14.61	15.17
$p^{U^mE^l}$	7.59	9.49	8.11	5.94	8.49	6.79	9.46	8.27	6.14	4.60	7.77	7.48
$p^{U^mU^h}$	0.40	0.13	0.28	0.68	0.77	0.45	0.31	0.41	0.43	0.40	0.52	0.34
$p^{U^mU^m}$	50.20	46.49	51.70	50.75	35.64	60.44	49.40	48.05	57.31	56.20	49.38	50.64
$p^{U^mU^l}$	1.08	1.50	1.20	0.73	2.97	0.90	0.63	1.11	1.02	1.04	1.16	1.04
p^{U^mI}	22.96	29.10	21.80	20.58	14.07	17.01	20.93	22.21	23.44	29.71	23.85	22.48
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Low Skill												
$p^{U^lE^h}$	1.60	0.50	1.26	3.80	1.71	1.46	2.11	1.76	1.20	0.86	1.58	1.62
$p^{U^lE^m}$	5.23	4.29	5.19	6.67	3.82	4.44	7.34	5.77	3.72	2.44	5.24	5.22
$p^{U^lE^l}$	22.12	22.27	22.50	21.62	41.97	19.19	21.54	24.71	14.27	10.42	21.61	22.40
$p^{U^lU^h}$	0.22	0.07	0.19	0.46	0.34	0.25	0.14	0.23	0.23	0.18	0.23	0.21
$p^{U^lU^m}$	0.78	0.69	0.84	0.80	1.31	0.78	0.38	0.76	0.88	0.81	0.87	0.73
$p^{U^lU^l}$	47.89	6.07	50.06	46.60	38.66	55.88	47.48	45.74	55.83	55.01	47.03	48.34
p^{U^lI}	22.15	26.11	19.96	20.04	12.19	18.00	21.01	21.03	23.87	30.27	23.43	21.49
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

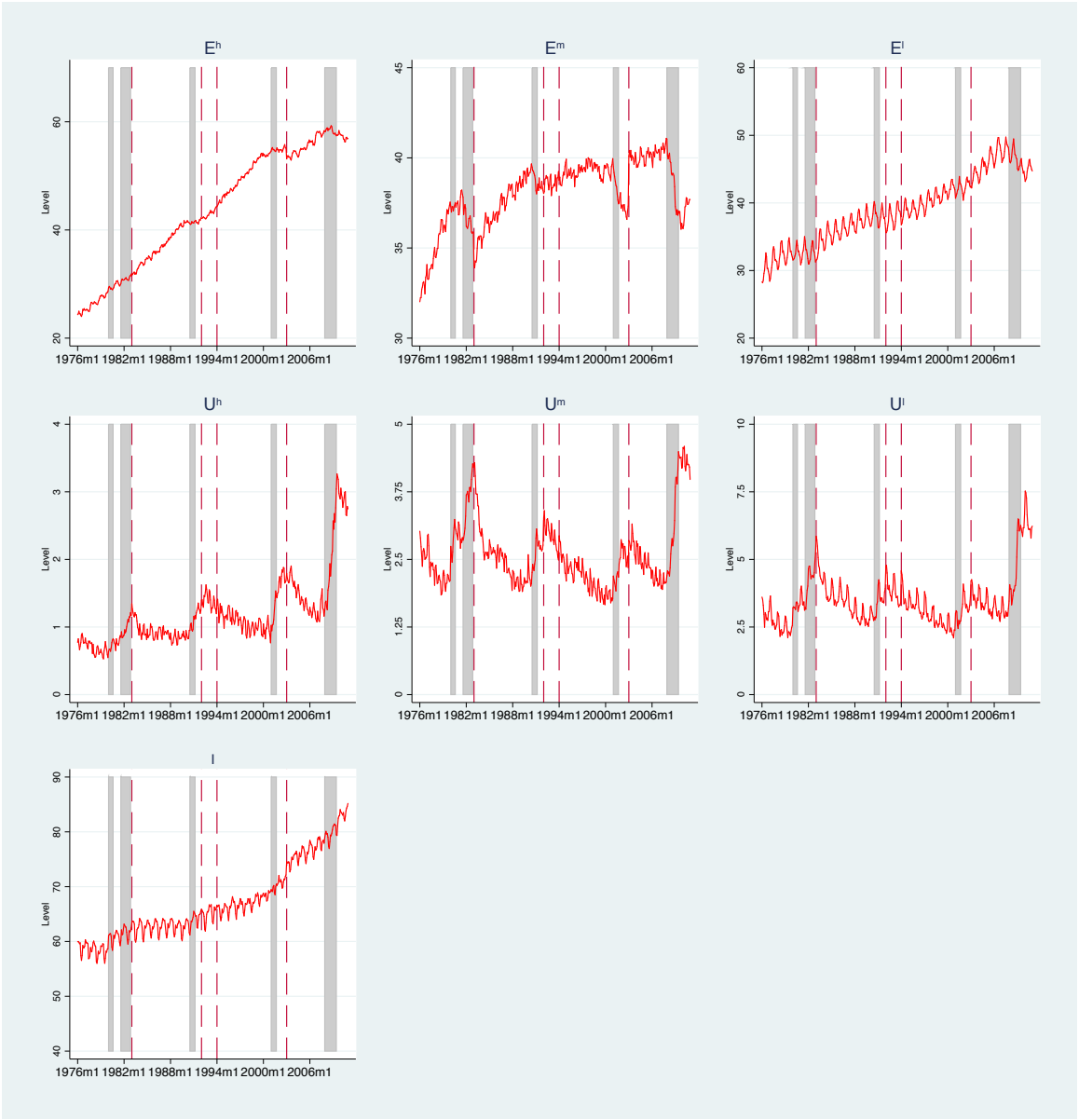
Flow rates from unemployment $p^{U\cdot X}$ expressed in percentage, computed from the raw matched CPS files and averaged over the period Feb. 1994 to Dec. 2010. All observations are weighted using weights provided by the CPS. The sample for unemployed excludes *New Unemployed Entrants*. (1), (2), (3) correspond to laid-off workers, job loser and job leaver. Duration in unemployment is grouped according to unemployed with a duration lower than 6 months (4), those with 6 to 12 months duration (5) and unemployed for more than a year (6). The last 2 columns display results for individuals in the rotation group (1) or not (0).

Table 4: Average Flow Rates from Unemployment over the period February 1994 - November 2010 (2)

[6]Note that this issue arises also for transition from employment such as E^hU^m or E^mU^h displayed in Appendix A.1.1.

2.2 Data Issues

As many researchers have highlighted, working with CPS (occupational) data requires an important pre-treatment of the original data in order to render them suitable for analysis. The issues corrected in this paper are reviewed below.[7] In order to help visualize some of the problems to be corrected, Figure 1 for stocks and Figures 2 and 3 for flows display some selected monthly series. It should be noted however, that the raw data are quite noisy meaning only substantial breaks can be identified from these figures.



Monthly population stocks expressed in millions. Unemployment series exclude *New Unemployed Entrants*. From left to right, the vertical dashed lines represent the 1983 and 1992 classification changes, the 1994 redesign and the 2003 classification change. Shaded areas correspond to recession periods as defined by the NBER.

Figure 1: Labor Market Stocks

[7] Adjustments for misclassifications between unemployment and inactivity are not be considered here as the focus is mostly on issues arising from the occupation dimension.

The main adjustements that are applied to stocks and gross flow series relate to the followings issues:

1 - Missing values/Outliers/Seasonality.

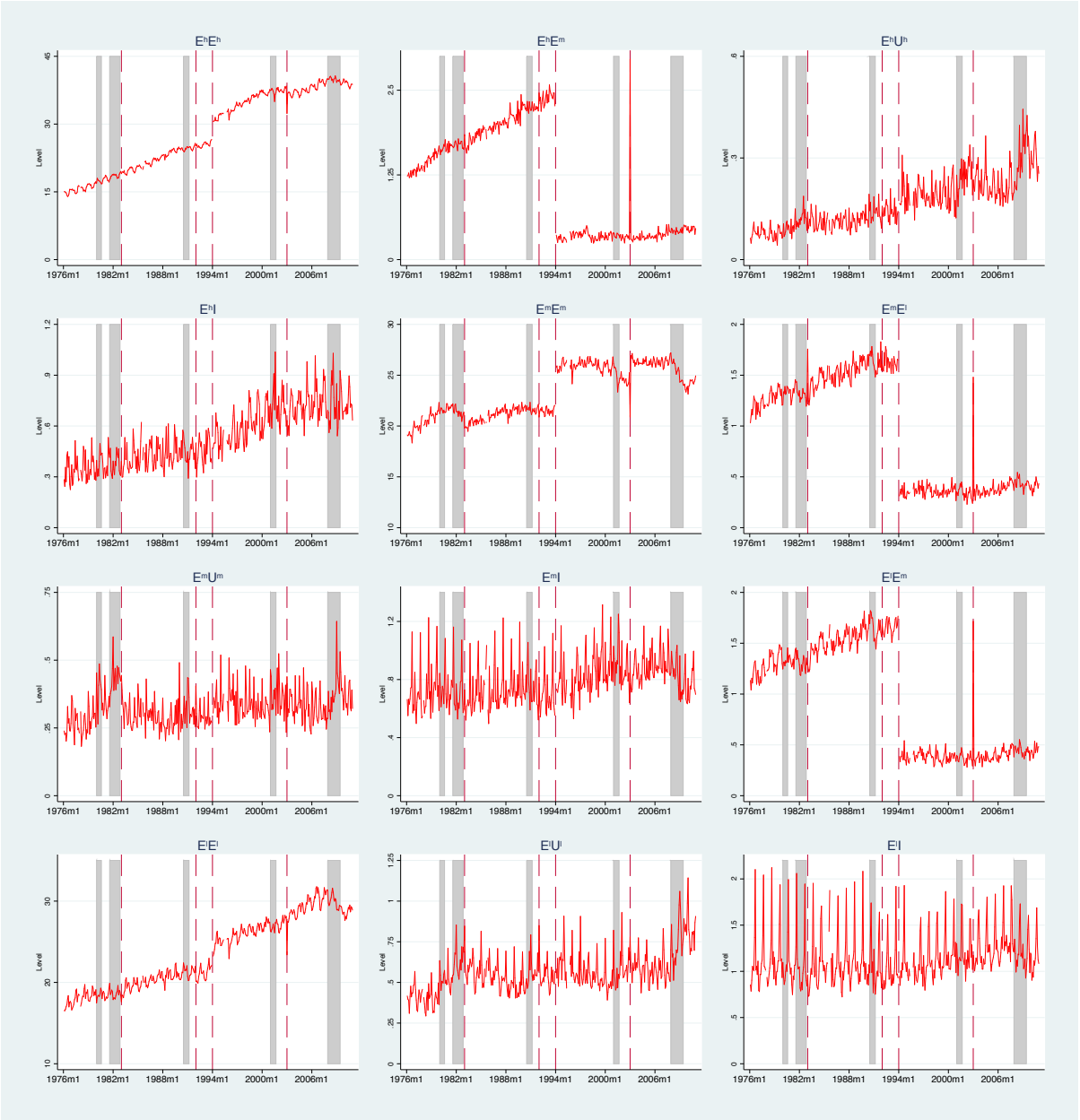
The changes in the construction of household identifiers over the years result in missing values. More specifically, these changes prevent the matching of individuals across 2 consecutive months for certain time periods, namely January 1978, July and October 1985, January 1994, and June to September 1995. This, in turn, results in missing values in the series of gross flows. Since population stocks are computed from the original unmatched samples, missing values are not an issue for these series. Moreover, some papers (i.e. Moscarini and Thomsson (2007) and Kambourov and Manovskii (2013)) have pointed to the fact that the occupation code assignment is subject to a substantial share of coding errors. Occupational data should therefore be used with caution. In the context of this paper, potential errors in this regard should be mitigated by the fact that I aggregate occupations in terms of only 3 groups. Nevertheless, series could be subject to potential outliers, and consequently a procedure will be applied to correct for these. Lastly, series related to inactivity and unemployment usually exhibit a strong seasonal component. We can further observe differences between skill groups as low skill flows and stocks seem to exhibit a stronger seasonal component.

2 - 1994 redesign of the CPS questionnaire.

A major redesign of the CPS survey, which introduced important changes to the questionnaire, took place in 1994. Polivka and Rothgeb (1993) and Polivka and Miller (1998) provide a detailed review of these changes and their potential effects on labor market statistics. 2 major modifications in the questionnaire need to be addressed.
The first modification relates to the use of computerized data collection technique through the introduction of the *dependent interviewing technique* which affected the occupational code assignment. Prior to 1994, individuals were asked to report their occupations each month they were interviewed. Because of the detailed level of the classification, respondents or interviewers/coders were likely to report/assign different occupations in 2 consecutive months while no real transition actually occurred. After 1994, the CPS started to ask the question on occupations only in the first month individuals were interviewed ($mis = 1$ or $mis = 5$ in the CPS). In the subsequent months, individuals are provided with a description of the occupation they reported the previous month and asked whether this description still corresponded to their current job. Only a negative answer would then trigger a new occupation question. As a result, the number of spurious transitions between occupations substantially decreased after 1994, generating a break both in the mean and the variance of some flow rates series.
The second important modification is related to changes in the questionnaire made to better identify the labor force status of respondents.[8] These modifications turned out to have a minor impact on stocks (see Polivka and Miller (1998)) but a more substantial one on flow series in particular on series between unemployment and inactivity (see Abraham and Shimer (2001) or Cortes et al. (2016)). It should further be noted that the 1994 redesign resulted in a significant decrease in *New Unemployed Entrants* stemming from minor changes to the questionnaire and the definition of this reason for unemployment.[9] Because an occupation code is not available for this group of unemployed individuals, they are dropped from the sample used in this paper. Therefore, the pool of unemployed for which an occupation is available significantly increases after 1994. This turns out to have important effects on unemployment stocks and flow series.

[8]For instance, individual waiting to start a new job that were considered as unemployed prior to 1994 must now make a search effort in the previous four weeks to be registered as unemployed. An effort was also made to better identify unpaid family work.

[9]The CPS classifies unemployed according to the reason for unemployment. The 6 groups are *Job Loser/On Layoff*, *Other Job Loser*, *Temporary Job Ended*, *Job Leaver*, *Rentrant* and *New Entrant*.

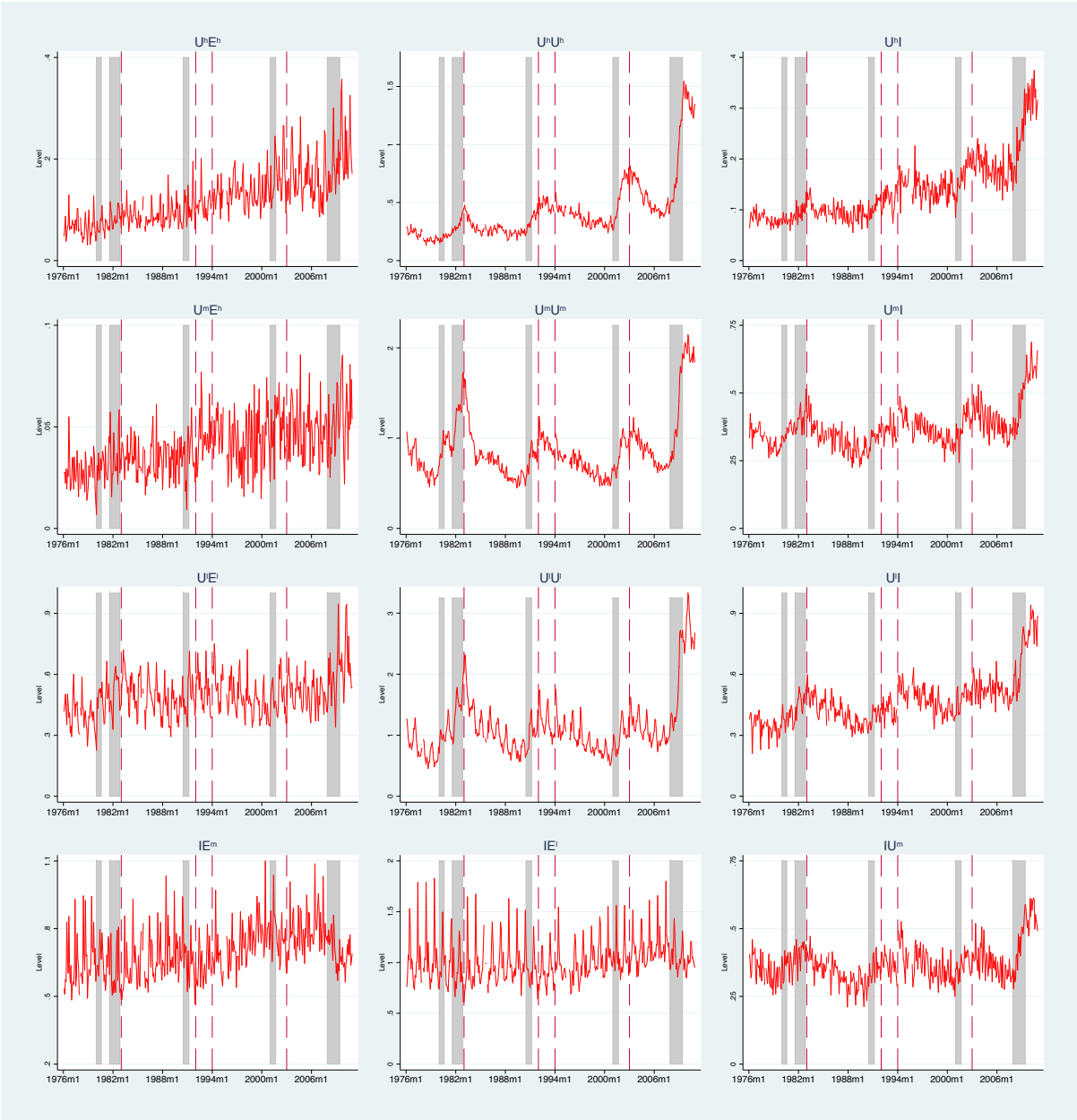


Monthly gross flows expressed in millions. Unemployment series exclude *New Unemployed Entrants*. From left to right, the vertical dashed lines represent the 1983 and 1992 classification changes, the 1994 redesign and the 2003 classification change. Shaded areas correspond to recession periods as defined by the NBER.

Figure 2: Gross flows (1)

Figures 1, 2 and 3 can help to assess the impact of the 1994 redesign on series of stocks and gross flows. Figure 1 seems to display no evident breaks, which appears to confirm that the 1994 redesign had little impact on stock series. Note however, that the effect of dropping *New Unemployed Entrants* can be inferred from Figure 5, which shows a clear break in 1994 for unemployment and none for employment and inactivity. Regarding gross flow series (Figures 2 and 3), a break in the mean can be seen in EE, EU (Figures 2) and possibly in UU series (Figures 3). In order to gain an idea of potential changes in the variance, I execute F-tests for equality of variance on log differenced series. The results of these tests (not reported here) indicate that for 61% of gross flows series, the null hypothesis of homoskedasticity can be rejected at a 5% significance level. For stock series, the null is rejected only

for high skill employment.[10] Furthermore, the 1994 redesign seem to have an effect on the trend of some series. We can refer to the $E^m E^l$ or $E^l E^m$ gross flows in Figure 2 in this regard. Overall, Figures 1, 2 and 3 indicate that series of stock and flows are affected in a heterogeneous way by the redesign. Figure 4 plots the ratio of aggregate employment, unemployment and inactivity obtained by adding the relevant gross flows to the same aggregate series obtained from stocks. These ratios can be understood as matching rates for CPS files (which is why they evolve around .7) and seem to display a break prior to and after the 1994 redesign.



Monthly gross flows expressed in millions. Unemployment series exclude *New Unemployed Entrants*. From left to right, the vertical dashed lines represent the 1983 and 1992 classification changes, the 1994 redesign and the 2003 classification change. Shaded areas correspond to recession periods as defined by the NBER.

Figure 3: Gross flows (2)

[10] The results of these tests serve only to indicate a potential break. These tests being quite sensitive to the normality assumption, their results should be taken with care.

3 - Classification changes.

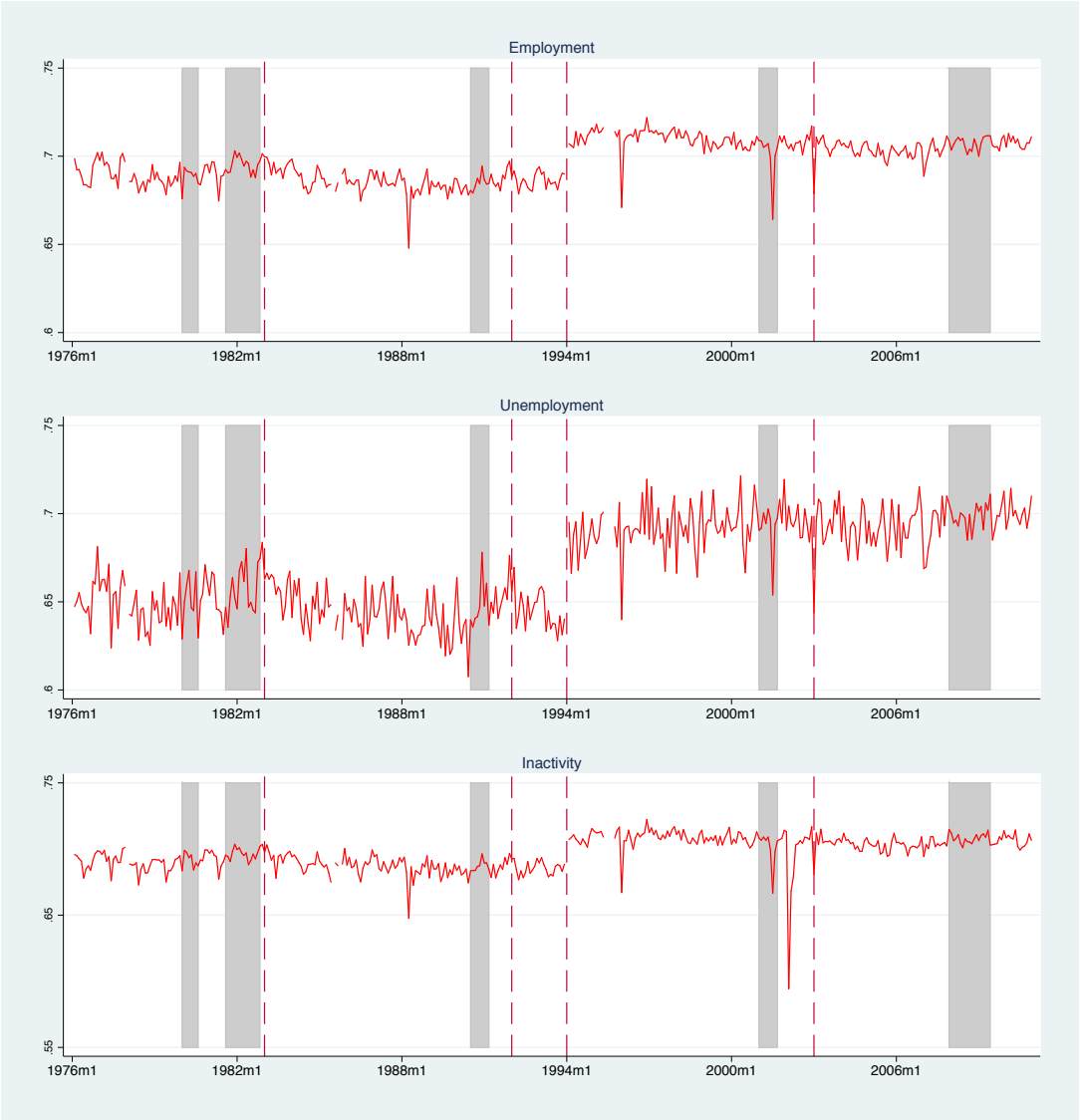
Every 10 years or so, the CPS revises and updates its occupational classification to reflect changes in the occupational composition of the labor market. For 1976-2010 time period, these classification changes occurred in 1983, 1992 and 2003 with 1983 and 2003 constituting major changes whereas the 1992 revision only implemented minor changes (see Cortes et al. (2016)).
The crosswalk built by D. Dorn is meant to correct for these classification changes by taking the 1992-2002 classification as a reference and building a balanced panel of occupations across the 1976-2010 period. This panel is, however, not sufficient to avoid breaks in stock and flow series. According to Cortes et al. (2016), this crosswalk still leads to a substantial drop in middle skill occupations in 1983 and the 2003 classification change also appears to have resulted in a permanent reallocation between high and middle skill employment series. These 2 effects are illustrated in Figure 1.
The effect on gross flow series is similar, with the E^mE^m series displaying shifts in 1983 and 2003 (2nd row in Figure 2). Moreover, flow series are further affected by an impact effect that originates from matching 2 consecutive months with 2 different classifications (e.g. December 1982 and January 1983). For instance, the E^mE^l series in Figure 2 shows a clear spike in January 2003. Similar spikes can be observed in U^mU^m and U^hU^h series in 1983 (Figure 3). Similarly to the 1994 redesign, these graphics reveal that not all series are equally affected by these classification changes.

4 -Population revisions

Over the 1976-2010 period, the BLS updated its population estimates according to new estimates produced by the Census Bureau and a detailed review for most of these modifications can be found in the following document: `https://www.bls.gov/cps/eetech_methods.pdf`. These updates usually result in increases in the level of series of interest. According to the BLS, most of these modifications lead to a proportional increase (or decrease) in employment, unemployment and inactivity meaning that rates are only marginally affected. Although, the effect of these population updates on series with occupational dimension is unknown, it is likely to be different across series with different skill levels. For instance, new population estimates in 1986 resulted in an increase of the total population due -to a large extent- to new estimates of the Hispanic population. This population is more likely to work in middle and low skill jobs (see Table 2), which suggests there is a more noticeable effect on these series as opposed to those for high skill jobs. Should this be the case, it would be important to account for these population changes.
Furthermore, the BLS adjusted series retroactively for some population revisions. This was the case for the 1982 update for which the BLS revised series from 1970 onwards, the 1985 and 1986 updates (which were applied to series from 1980), the 1994 population change (applied to series from 1990) and the 2003 update (for which series from 2000 were revised). The new population estimate applied in 1997, 1998, 1999 and after 2003 led to no revision in series.[11] The differences between these population revisions should be considered when modelling these effects. Finally, it is worth mentioning that the micro data downloaded from the NBER website do not provide revised population weights for the time period of 1976-79 and 2000-02.[12] This can be inferred from Figure 5 that plots the ratio of aggregate employment, unemployment and inactivity computed from the NBER micro data to the same official series released by the BLS. This ratio is close to 1 for employment and inactivity except for the 2 time periods mentioned above. For unemployment, the ratio is lower than 1 since *New Unemployed Entrants* are dropped from the sample.

[11] In January 2003, on top of the update applied to series from 2000 onwards, there was an additional increase in population that lead to no revision in past series. This month also saw the introduction of a new classification for occupations.

[12] Revised weights for January 2000 to December 2002 are actually available in separate files.



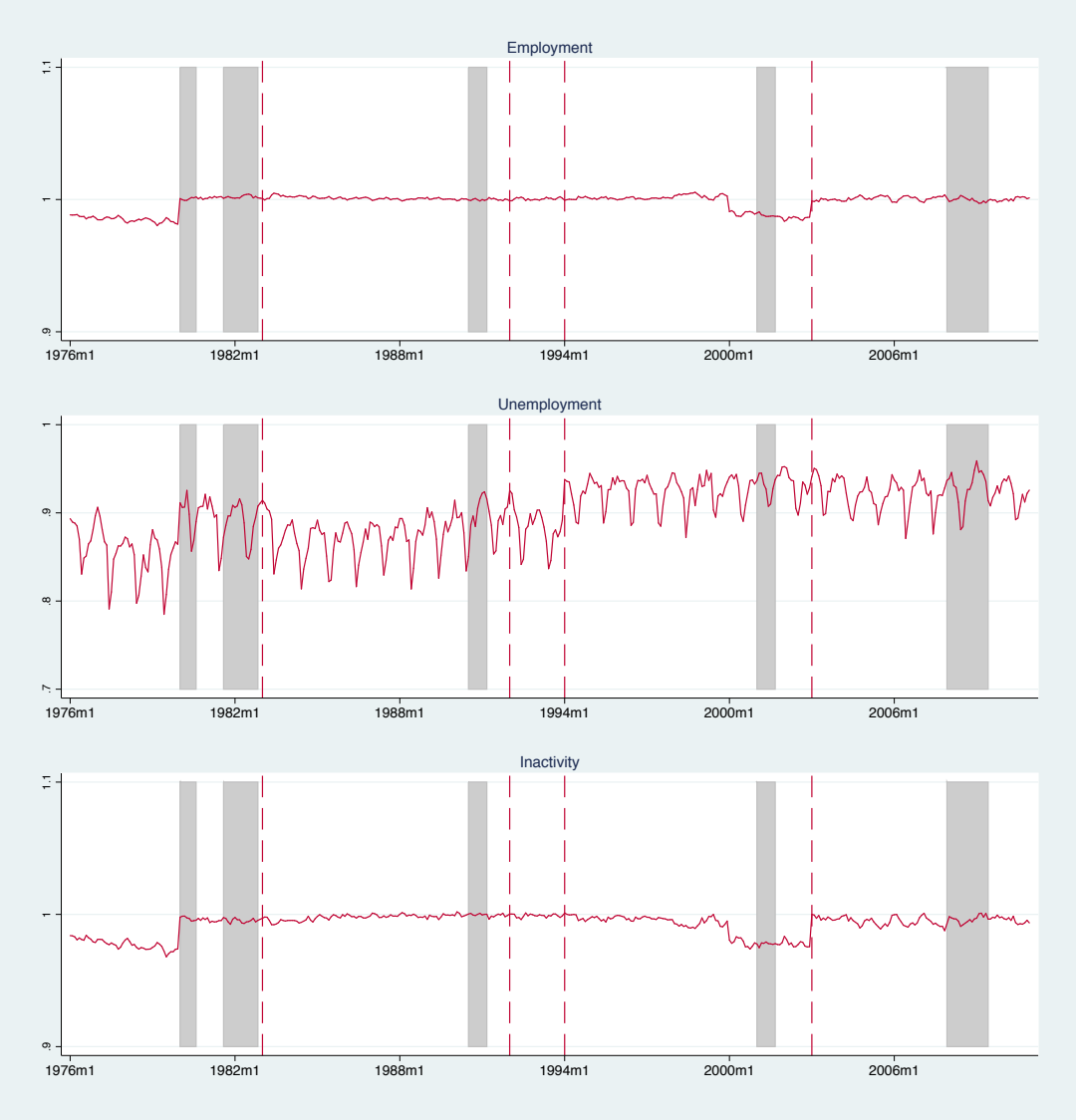
Ratio of aggregate employment, unemployment and inactivity obtained from flows (e.g. $E_{t-1} = E^h E_t^h + E^h E_t^m + ... + E^m E_t^h + E^m E_t^m + ... + E^l E_t^h + E^l E_t^m + ...$) to the same aggregate series computed from stocks (e.g $E_{t-1} = E_{t-1}^h + E_{t-1}^m + E_{t-1}^l$). This ratio can be understood as a matching rate of CPS files since stocks obtained from flows are obtained from matched CPS files. Unemployment series exclude *New Unemployed Entrants*. From left to right, the vertical dashed lines represent the 1983 and 1992 classification changes, the 1994 redesign and the 2003 classification change. Shaded areas correspond to recession periods as defined by the NBER

Figure 4: Ratio of Flows to Stock series

5 -Margin of Adjustment and Time Aggregation bias

The Margin of Adjustment issue (Abowd and Zellner (1985) and Poterba and Summers (1986)) relates to the fact that flow rates do not allow for the perfect reproduction of the evolution of population stocks due to mortality, retirement or migration.
On the other hand, the Time Aggregation bias originates from the discrete nature of data collection. This bias has been acknowledged in the literature on gross flows since the work of Kaitz (1970) and Perry et al. (1972). The labor market status is determined according to the labor market activities during the week prior to the interview (the reference week) but no questions are asked about activities

in the 3 weeks prior to the reference week.[13] It is therefore possible for a worker to be recorded as employed in month $t-1$, to experience a short spell in unemployment and to find quickly a new job such that when interviewed in month t, her status would be employed. The recorded transition would therefore be an EE transition instead of an EUE one.



Ratio of seasonally unadjusted series for aggregate employment, unemployment and inactivity obtained from micro CPS files to series obtained from the BLS. Unemployment series from micro CPS files exclude *New Unemployed Entrants*. From left to right, the vertical dashed lines represent the 1983 and 1992 classification changes, the 1994 redesign and the 2003 classification change. Shaded areas correspond to recession periods as defined by the NBER

Figure 5: Ratio of Stock Series From Micro Data to Officially Released Series

As Elsby et al. (2015) points out, considering that around 50% of unemployed transition to employment or inactivity on average each month (see Table 3 or Table 4), recorded transitions are likely to miss an intervening spell in unemployment. Moreover, Feldstein (1975) shows that workers on lay off account for a substantial share of unemployed (around 26% according to Table 2) and are expected to be recalled within thirty day. These unemployed could potentially also contributes to missing EUE

[13] This not fully correct for unemployed who need to have made a search effort in the four weeks prior to the interview and be currently available to take a job.

transitions. Using a different data source (the Survey of Income and Program Participation (SIPP) for the period 1984-1985), Ryscavage (1989) finds that around 10% of unemployed individuals experienced an unemployment spell shorter than one month. This transition would thus be missed by the CPS. Furthermore, there could be a skill dimension to the Time Aggregation bias. Intuitively, workers in low skill occupations are likely to experience more labor market transitions than those in high skill occupations which would result in a higher number of missed transitions for these workers.
A last remark on this topic regards the fact that most of the previously mentioned papers adopt a 2 states framework (E and U). Their focus is therefore on missed EUE and UEU transitions. As pointed by Shimer (2012) and Elsby et al. (2015), in a 3 states framework, transitions involving inactivity (IUE or EUI transitions) could also be subject to this Time Aggregation bias.

All the issues which have been presented above are adjusted in the following sections in 2 steps. Firstly, the adjustments for the 1994 redesign, classification changes and population revisions are presented in Section 3. Secondly, corrections for the Margin of Adjustment and the Time Aggregation bias rely on a different framework, which is presented as part of a second step in Section 4.

3 A Framework for Seasonality, Outliers and Intervention Effects

Missing values, outliers and seasonality are issues found in most time series and are usually dealt with the *X-12ARIMA* procedure used by the Census Bureau or the *TRAMO-SEATS* procedure developed by the Bank of Spain.[14] These 2 procedures correct time series and perform the seasonal adjustment in 2 steps. First, the series are corrected for deterministic effects such as trading days, holidays, or any other interventions that could affect the series. An outlier detection procedure is also carried out during this first step. Secondly, the seasonal adjustment is performed using moving averages (*X-12ARIMA*) or an unobserved component model (*TRAMO-SEATS*).

These 2 methods are widely used and effective, but they are not fully able to adjust the problems at hand here. In particular, the 1994 redesign generates a potential change in the variance of some stock and flow series. This effect cannot be addressed by these 2 procedures which only allow for the correction of deterministic effects such as a change in mean or trend. Furthermore, having to correct 56 series that are not equally affected by the issues presented in Section 2.2, makes it worth developing a set-up flexible enough to adjust each series individually.

In the next section, I develop a framework based on Unobserved Component (UC) models (Harvey (1990), Durbin and Koopman (2012)) rather than Autoregressive Integrated Moving Average (ARIMA) models which are used in the first step of the *X-12ARIMA* and *TRAMO-SEATS* procedures. Compared to $ARIMA$ models, UC models present the advantage of not relying on the assumption of stationarity. In this way, the transformation of the series, usually taking differences, can be avoided and, thereby, the potential loss of information coming from these transformations can be prevented. Furthermore, UC models are based on a structural decomposition in which all components are modelled to capture specific aspects of the observed series. The framework presented in Section 3.1 also allows for the execution of the seasonal adjustment in one step, which would not be possible using the *X-12ARIMA* and *TRAMO-SEATS* procedures. Moreover, UC model rely on the Kalman Filter and Smoother which allows for a straightforward account of missing values. Diagnosis tests to assess the performance of a given specification can be constructed from the output of the filter which in turn permits the detection of outliers without requiring extensive additional computations. For a more detailed discussion on the advantages of UC models, the reader is referred to Durbin and Koopman (2012).

[14] See *www.census.gov* for the X-12ARIMA procedure (actually X-13 now) and *www.bde.es* for the TRAMO-SEATS one.

3.1 An Unobserved Component model

In UC models, the baseline specification decomposes the univariate series (y_t) into a mean (μ_t), a seasonal (γ_t) and an irregular component (ε_t). To this standard decomposition, I add an additional component χ_t to capture the effect of the 1994 redesign and the effects of exogenous variables (X_t) collected in the vector β. These exogenous variables are meant to capture the effects of classification and population changes:

$$y_t = \mu_t + \gamma_t + \varepsilon_t + \chi_t + X_t\beta \tag{2}$$

with

$$\mu_t = \mu_{t-1} + \nu_t + \eta_{\mu,t}, \quad \eta_{\mu,t} \sim \mathcal{N}(0, \sigma^2_\mu) \tag{3}$$

$$\nu_t = \nu_{t-1} + \eta_{\nu,t}, \quad \eta_{\nu,t} \sim \mathcal{N}(0, \sigma^2_\nu) \tag{4}$$

$$\gamma_t = -\sum_{k=1}^{11} \gamma_{k,t} + \eta_{\gamma,t}, \quad \eta_{\gamma,t} \sim \mathcal{N}(0, \sigma^2_\gamma) \tag{5}$$

The mean and trend components μ_t and ν_t, are both assumed to evolve according to random walks. As pointed by Durbin and Koopman (2012), setting $\sigma^2_\nu = 0$ implies $\nu_t = \nu_{t-1} = \nu$ which corresponds to a deterministic linear trend. The seasonal component γ_t is modelled as dummy variables for each of the twelve months plus a random noise $\eta_{\gamma,t}$ that allows the seasonal component to vary over time. This specification ensures that the expected value of the sum of the seasonal effect over a year is zero.[15]

For the irregular component ε_t, I assume an $ARMA(p, q)$ specification:

$$\varepsilon_t = \sum_{j=1}^{p} \rho_j \varepsilon_{t-j} + \eta_{\varepsilon,t} + \sum_{j=1}^{q} \theta_j \eta_{\varepsilon,t-j}, \quad \eta_{\varepsilon,t} \sim \mathcal{N}(0, \sigma^2_\varepsilon). \tag{6}$$

In UC models, the irregular component is usually assumed to be a simple white noise ($p = q = 0$). However, this specification can leave some autocorrelation in the residuals.[16] It is therefore useful to allow for an ARMA specification to capture this autocorrelation. The random noises $\eta_{\mu,t}, \eta_{\nu,t}, \eta_{\gamma,t}$ and $\eta_{\varepsilon,t}$ are assumed to be uncorrelated.

The effect of the 1994 break is modelled as an $ARMA(p_\chi, q_\chi)$ process with mean which affects the series from 1976 to 1993. This assumption allows to account for changes in the variance that could affect some series. In Section 2.2, it was also mentioned that the 1994 break could potentially have an effect on the trend. To capture this aspect, I allow the mean to evolve as a random walk with a linear trend:

$$\chi_t = \mu_{\chi,t} + \varepsilon_{\chi,t} \tag{7}$$

$$\mu_{\chi,t} = \mu_{\chi,t-1} + \nu_\chi + \eta_{\mu_\chi,t}, \quad \eta_{\mu_\chi,t} \sim \mathcal{N}(0, \sigma^2_{\mu_\chi}) \tag{8}$$

$$\varepsilon_{\chi,t} = \sum_{j=1}^{p_\chi} \rho_{\chi,j} \varepsilon_{\chi,t-j} + \eta_{\varepsilon_\chi,t} + \sum_{j=1}^{q_\chi} \theta_{\chi,j} \eta_{\varepsilon_\chi,t-j}, \quad \eta_{\varepsilon_\chi,t} \sim \mathcal{N}(0, \sigma^2_{\varepsilon_\chi}) \tag{9}$$

The disturbances $\eta_{\mu_\chi,t}$ and $\eta_{\varepsilon_\chi,t}$ are assumed to be mutually uncorrelated.

The effects of classification and population changes are modelled using dummy variables. Classification changes are assumed to generate a transitory effect on the mean of the series. These effects are therefore captured by variables, taking the value 1 when the classification is active (e.g. 1976-1982) and 0 otherwise. The 1992-2002 classification used by Autor and Dorn (2013) to build their panel of occupations is assumed to be the reference classification. On the other hand, population changes are

[15] This restrictions is one way to ensure identification of this component w.r.t to the mean component. See Harvey (1990).

[16] For instance outliers, residual seasonality or the presence of a cycle could explain why some autocorrelations could be left in residuals.

modelled as permanent change in the mean of the series (or level shift). This implies the addition of exogenous variables taking the value 1 from the period when the population adjustment happens onwards.[17] One exception regards the 1976-79 population change. This effect is likely to originate from the population adjustments of the 80's that led to the revision in series from 1980 (and not from 1976). To capture this effect, I model an additional transitory change in the mean over this specific period. Furthermore, it should be noted that without additional restrictions (discussed in Section 3.3.2), it is not possible to disentangle the classification effect from the population effect taking place in 2003. As a result, only one exogenous variable captures both effects. Finally, the impact effects generated by classification and population changes on gross flow series are captured by variables taking the value 1 for the month in which the change occurs (e.g. January 1983 for the first classification change) and 0 otherwise.

Model (2)-(9) constitutes the general specification and in practice, a simplified version of this set-up is generally be estimated. As an example of specification and for most series exhibiting a change in their variance prior and after 1994, assuming a constant mean plus noise for the 1994 component χ_t ($p_\chi = q_\chi = 0$) and an autoregressive process for the irregular component ($p \neq 0$ and $q = 0$) will be enough to correct these series.

Model (2)-(9) can then be written in state space form and estimated using the Kalman Filter:

$$y_t = Z_t \alpha_t \tag{10}$$

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t \ , \qquad \eta_t \sim \mathcal{N}(0, Q_t). \tag{11}$$

The fact that some of the components are non-stationary (e.g. the mean component in (3)) implies that the standard recursions cannot be initialized in the usual manner through the use of the unconditional mean and variance for α_t. I will use the *Augmented Kalman Filter* of De Jong (1991) to account for the non-stationarity of some of the state variables. All these specificities are discussed in more details in Appendix A.2.1 through A.2.3.

3.2 Outlier Detection

In the context of time series, outliers can generate a substantial bias on estimated parameters through their impact on the autocorrelation function of the data. As a result, locating and correcting outliers is an important step when working with time series (see Fox (1972), Tsay (1988) or Chen and Liu (1993)). Both *X12-ARIMA* and *TRAMO-SEATS* implement an outlier detection procedure inspired by the work of Chen and Liu (1993). However, their proposed procedure is developed within the $ARIMA$ framework and it does not fit directly into the UC framework.

De Jong and Penzer (1998) demonstrate how outliers can be detected through a simple modification of the state space (10)-(11) by introducing shocks to both equations. The attractive feature of the framework developed by De Jong and Penzer (1998) is that the potential effect of an outlier can be estimated directly from the output of the Kalman Filter and Smoother (see Appendix A.2.2 for the Kalman Smoother). Furthermore, it is possible to perform several tests to check for the statistical significance of the estimated effects and an analogue to the standard t-statistics can be used to select outliers. All the relevant derivations and proofs can be found in the work of De Jong and Penzer (1998). These statistics can be computed for all time periods and for various types of outliers which leads to 2 common problems in outlier detection.

First, I need define which types of outliers are to be considered in the analysis. I follow the literature and consider additive outlier (AO), level shift outlier (LS) and outliers to the seasonal component. AO and LS capture, respectively, a one time and a permanent change in the series (see Chen and Liu (1993)).[18] For seasonal outliers, I follow Penzer (2006) who present an application of

[17] The assumption of a transitory or permanent change in the mean of the series can be relaxed by adding a random component to these effects. A change in trend could also be captured by adding a trend break.

[18] These 2 outliers are also considered in the *X12-ARIMA* procedure.

the framework developed by De Jong and Penzer (1998) to seasonal series. He considers shocks to each of the seasonal dummies individually. Overall, for each period t in the sample, 13 statistical tests are performed (AO, LS and the 11 seasonal components).

The second issue relates to the identification of the number, the type and the location of the potential outliers. One way to proceed is to use an iterative procedure as proposed by Chen and Liu (1993). This procedure allows for the joint estimation of the model's parameters and the effects of potential outliers through 3 stages. In the first stage, outliers are found starting from the first period in the sample ($t = 1$) and the t-statistics are computed for all type of outliers (AO, LS and seasonals). If the maximum of these t statistics is greater than some predefined critical value C, there is the possibility of an outlier. This first stage serves to identify the number, date and type of potential outliers one by one. In the second stage, the potential outliers found in stage 1 are estimated jointly and those with a t statistics smaller than C are removed one by one, starting with the smallest and re-estimating parameters each time an outlier is removed. In the third stage, the first 2 stages are repeated until no outliers are found. This section only sketches the main aspects of the outlier detection procedure and more information and details can be found in Appendix A.2.4.

3.3 Estimation and Diagnostic Checks

3.3.1 Specification and Diagnosis tests

The log likelihood function can be evaluated through one run of the Augmented Kalman Filter presented in Appendix A.2.2. For most series, a multiplicative model is assumed by specifying the dependent variable in log. An additive model (dependent variable in level) is assumed for series exhibiting a change in trend before and after 1994 (e.g. E^hE^m, E^mE^l or E^lE^m in Figure 2). All corrected series (and their log difference) are then visually inspected and I look for a specification of model (2)-(9) leading to no autocorrelation and normality of the innovations, $v_t = y_t - Z_t a_t$ (see equation (10)).

In order to determine the specifications for the irregular component (6) and for the 1994 redesign component (7)-(9), I start by estimating a simpler version of model (2)-(9) with the irregular component assumed to be a white noise ($p = q = 0$) and the 1994 redesign component χ_t, a constant ($p_\chi = q_\chi = 0$, $\sigma^2_{\varepsilon_\chi} = 0$ and $\sigma^2_{\mu_\chi} = 0$) without trend. Once this specification has been estimated, I perform the above mentioned tests as well as a heteroskedasticity test on the innovations by splitting the sample in 1994. Doing so can provide indication on whether the specification captures the potential effect of the 1994 break on the variance of the series. The starting specification is then adjusted according to the test results. If there is potentially autocorrelation in the innovations, I add AR and/or MA terms to the irregular specification (6). If the innovations show signs of heteroskedasticity, I adjust the specification for the 1994 components. For most series failing this test, it suffices to allow for a mean plus noise specification. For some other series (like E^mE^l in Figure (2)), a (linear) trend must be added to the specification of the 1994 redesign component. If some problems still persists, I attempt to adjust the critical value C to check whether any (additional) outliers can be detected.

When a satisfactory specification has been obtained, the corrected series is computed by subtracting from the original series, the Kalman Smoother estimates for the seasonal components, the 1994 redesign and the classification and population changes.[19] The effects of classification and population changes are subject to a different selection process discussed in the following section.

[19]For some series estimated using an additive model, the 1994 component is, in a very few cases, quite large leading to over smoothed corrected series before 1994. One way to avoid this problem is to allow for a correlation between the disturbances of the irregular components (6) and (9).

3.3.2 Estimating Classification and Population effects

As described in Section 3.1, classification and population effects are modelled as temporary or permanent changes in the level of series. This implies that caution is warranted with regard to the estimated results as these could capture a change in mean resulting from factors unrelated to classification/populations changes.[20] A second reason for treating the results with care regards the number of effects which have to be estimated, in particular for flow series. If all population changes reported by the BLS are corrected, there is a total of 16 effects to estimate (2 classifications + 13 populations changes + 2003 change which is both a classification and a population change) for stock series. This number doubles for gross flows, since additional impact effects have to be estimated for these series.

In order to ensure that the estimated effects actually capture classification/population changes, I try to exploit additional information. I start by checking the individual statistical significance of each effect by computing t-statistics similar to the ones used for outliers in Section 3.2. However, using only statistical tests is not enough to determine whether a given classification or population change does or does not affect a series. In particular, the statistical power of a tests is influenced by many factors, such as the sample size or the magnitude of the estimated effects, and type II errors (failing to reject the null hypothesis of no effect while it is wrong) cannot be ruled out.

To reinforce the belief that I actually correct population and classification, I use information provided by the BLS on the impact of population adjustments on employment, unemployment and inactivity. These estimates are usually based on computing labor market statistics applying the new population weights in the month prior to their introduction (usually December) and comparing these with the results obtained with the old population weights. These estimates from the BLS can be used to indicate the likely sign and magnitude of the population changes.

Furthermore, I exploit the fact that for employment, classification changes constitute a reallocation between occupation groups (i.e. between E^h, E^m and E^l) that leave the aggregate level of employment (i.e. $E = E^h + E^m + E^l$) unchanged. This fact can be inferred from Figure 5, since the ratio of the micro data series ($E = E^h + E^m + E^l$) to the officially released series is close to 1 except for periods associated with population changes (1976-79 and 2000-2003). For unemployment, Figure 5 could suggests a potential effect of the 1976-1982 classification.[21] For inactivity, Figure 5 shows no effect of classification as it should be the case.

This additional information implies that I can start to estimate population changes from aggregate series as these series should not be affected by classification changes. Subsequently, I can check their statistical significance and whether their signs and magnitudes are consistent with those reported by the BLS. The estimates that are not consistent are discarded. This lowers the number of population effects and allows to select those that are significant and/or consistent with the BLS estimates. The selected effects are then estimated at the occupational level with classification changes. For classification changes, I further check that estimated effects compensate each other across occupation group of a given state. For instance, if I estimate an increase in E^l for a specific classification, a similar decrease in E^h and/or E^m must be obtained to ensure that the aggregate employment level is left (almost) unchanged.[22] I also check the statistical significance of these effects. Similarly to population changes, classification estimates that would lead to inconsistencies are discarded.

Once the populations and classifications effects have been estimated, I perform some checks by comparing the estimates from correcting aggregate series (e.g. E) with those obtained at the disaggregated level (e.g. $E = E^h + E^m + E^l$). This selection process is applied to both stock and flow series. Furthermore, I check that estimates for stock series are consistent with those obtained for flow series. This

[20] An example of such factors could be a recession. For instance, the 1976-1982 classification affect series for 7 years, 2 of which corresponds to the double dip recessions of 1980.

[21] This effect can be seen more easily from Figure 13 in Appendix A.2.5 which presents this selection process in more details.

[22] Note that this restriction could be used to disentangle the population and classification effects happening both in January 2003. However, this would require switching to a multivariate framework.

is done by exploiting the fact that the sum of gross flows at time t from a given state should be equal to the stock of this state in period $t-1$ (e.g. $E^h_{t-1} = E^hE^h_t + E^hE^m_t + E^hE^l_t + E^hU^h_t + E^hU^m_t + E^hU^l_t + E^hI_t$).

The results of this selection process are presented in detail in Appendix A.2.5. From these results, I correct the 1976-79, 1990, 2000 and 2003 population changes[23] as these appear to be the changes statistically significant and/or in line with the BLS results. Moreover, these population updates correspond to the largest ones according to the BLS. For classification changes, I correct the 1976-1982 and the 2003 classification changes. There is no variable meant to capture the 1983-1991 classification change which is line with the observations made by Cortes et al. (2016) on the fact that the 1983 change was a major one (compared to the period 1976-1982) whereas the change from the 1983-1991 classification to the 1992-2002 one (which serves as reference) was a minor one. Note that these effects are not estimated for all series by occupations. In particular, the 1990 population revisions and the 1976-82 classification change are usually not considered in the specification of high skill stocks and flows as these effects are estimated to be close to 0 (or of the wrong sign). See Appendix A.2.5.

It is worth emphasizing that all these efforts are undertaken to ensure that no differences and inconsistencies between adjusted series are created by correcting estimated effects that could be capturing other factors affecting series. This does not imply that all the effects related to the issues presented in Section 2.2 are corrected but those found to be important (statistically significant) and in accordance with external evidences are. Furthermore, the framework presented so far could be extended and it should be noted that a pre-treatment of the data in the line of Moscarini and Thomsson (2007) could also be applied before these corrections are performed.

3.4 Corrected Series

This section discusses the estimation results obtained for the 1994 redesign, and the classification and population changes. These results are displayed in Table 5 for stocks and in Tables 6 and 7 for flows. These tables show effects that need to be subtracted from the original series. Thus, a negative sign implies that the original series increases. Effects are computed by taking their average values over the time period during which they affect series (e.g the average effect over the time period 1976-1982 for the first classification change, over 1976-1994 for the CPS redesign, ...) and expressed in millions. Note that classification changes, population revisions and the 1994 redesign effects are not estimated for all series. See Appendix A.2.5. These tables also display the effects obtained by directly correcting aggregate stock and flows. As explained in the previous section, these aggregate series are used to compare estimates with those obtained from disaggregated series. Furthermore, some evidences on aggregate series are available for the 1994 redesign from the work of Polivka and Miller (1998) for stocks, and Abraham and Shimer (2001) and Cortes et al. (2016) for flows. Figures 6, 7 and 8 exhibit a subset of the corrected series for stocks and flows and Figure 9 plots stocks normalized by total population and computed from the corrected stocks series:

$$\begin{aligned} e^i_{stock} &= \frac{E^i}{E^h + E^m + E^l + U^h + U^h + U^m + U^l + I} \\ u^i_{stock} &= \frac{U^i}{E^h + E^m + E^l + U^h + U^h + U^m + U^l + I} \end{aligned} \tag{12}$$

with the employment to population ratio $e = e^h + e^m + e^l$.

[23] Note that for the 2000 population change, the correction is performed by simply applying the unrevised weights available in the CPS micro data for this period. Hence, there is no need to model an additional permanent change for this population change.

The same stocks can be obtained from corrected gross flow series:

$$e^i_{flows} = \frac{\sum E^i j}{\sum E^h j + \sum E^m j + \sum E^l j + \sum U^h j + \sum U^h j + \sum U^m j + \sum U^l j + \sum I j}$$
$$u^i_{flows} = \frac{\sum U^i j}{\sum E^h j + \sum E^m j + \sum E^l j + \sum U^h j + \sum U^h j + \sum U^m j + \sum U^l j + \sum I j} \quad (13)$$

Additional details on the specifications retained for all series, the tests carried out and the outlier detection can be found in Appendix A.2.6.

Stock	Redesign	Classification	Population		Classif. and pop.
	1976-93	1976-82	$1976-79$	1990	2003-10
Aggregate					
E	-	-	-1.62^{***}	0.98^{***}	3.19^{***}
U	-0.59^{***}	0.3^{*}	-0.43^{***}	0.07	0.13
I	0.4	-	-1.42^{***}	-	1.75^{***}
Skills					
E^h	-0.17	-	-0.65^{***}	0.07	-1.85^{***}
E^m	-0.25	1.27^{***}	-0.36^{*}	0.11	3.33^{***}
E^l	0.4^{*}	-1.2^{***}	-0.43^{**}	0.68^{**}	1.27^{***}
$\sum E^j$	-0.01	0.07	-1.43	0.86	2.75
U^h	-0.04	-	-0.01	-	-
U^m	-0.1	0.26^{***}	-0.22^{**}	0.07	0.18
U^l	-0.54^{***}	-	-0.2^{**}	-	0.19
$\sum U^j$	-0.68	0.26	-0.43	0.07	0.36

Average effects of the 1994 redesign, the 1976-1982 classification change, the 1976-79 and 1990 population changes and the 2003 classification/population change over the periods during which they affect series (e.g. 1976-1994 for the redesign). These are expressed in millions and such that they are the quantities that need to be subtracted from the original series (a negative effect implies that the series needs to be increased). $*$, $**$, $***$ indicate the significance level at 10%, 5% and 1%. A "-" means that the effect was not included into the model specification.

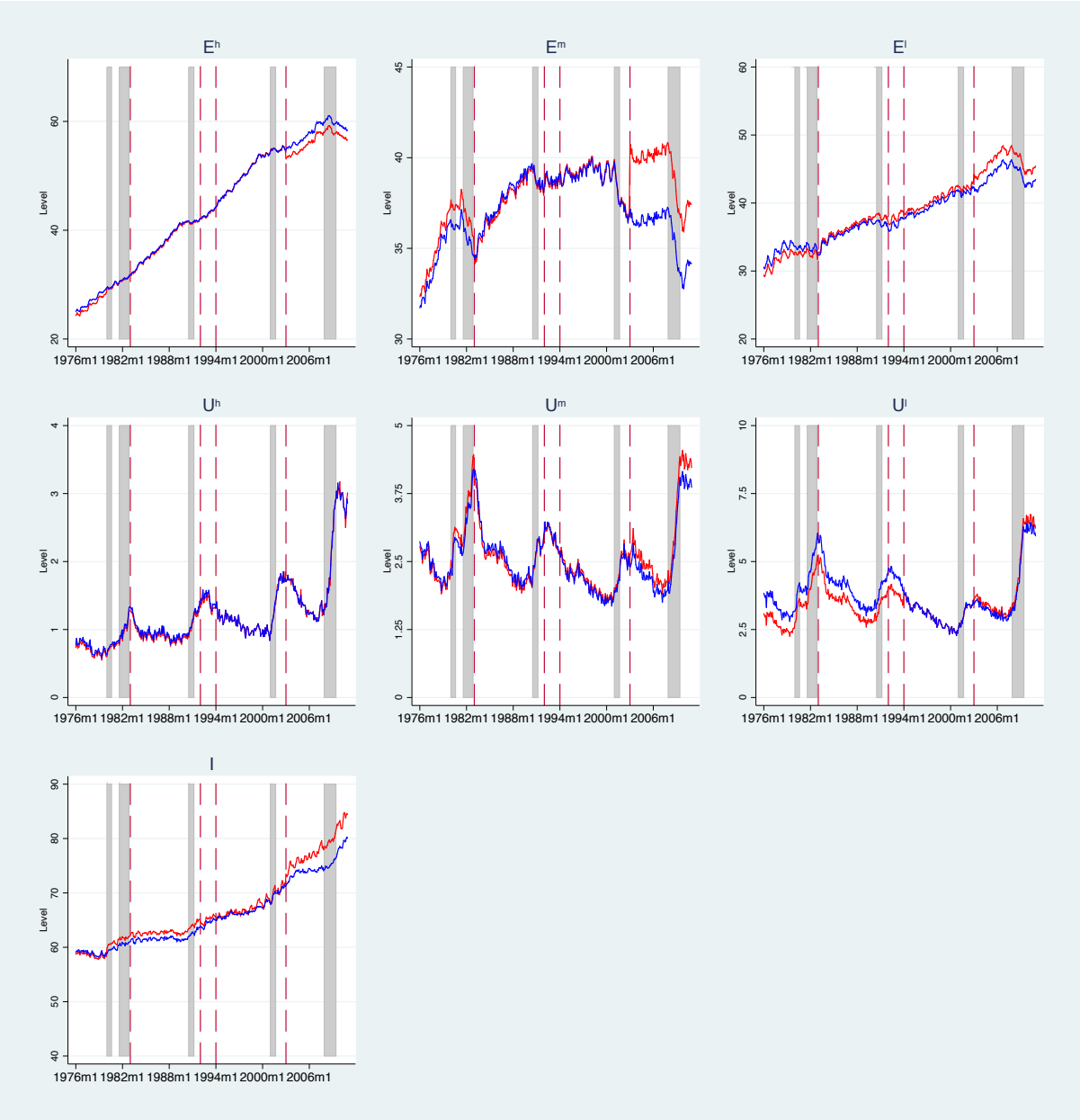
Table 5: Stocks : 1994 Redesign, Classification and Population Changes.

Table 5 shows that the correction for the 1994 redesign increases unemployment, decreases inactivity and leaves the employment level unchanged. The adjustment for unemployment originates from low skill unemployment and the results for employment point to a reallocation between low skill and high/medium skill occupations.

These observations are in line with the results reported by Polivka and Miller (1998) who find increases in terms of employment after the redesign in the manufacturing industry (middle skill) and in the finance, insurance and real estate industries (high skill).[24] Moreover, they point to significant decreases of employment in the construction and the transportation and utilities industries (low skill). They also find an increase in unemployment of 9% in the service industry (low skill) after the redesign. In addition, Polivka and Miller (1998) show that after the redesign, the share of *New Unemployed*

[24]Polivka and Miller (1998) results are based on the administration of a parallel survey from July 1992 to December 1993 using the new questionnaire to be introduced in 1994, and using the old questionnaire from January 1994 to may 1994. Note that industries do not correspond perfectly to occupations but, for instance, the manufacturing industry is likely to have a high share of middle skill occupation jobs. Polivka and Miller (1998) also give results in terms of 6 occupational groups but these differ from the ones presented in Table 1.

Entrants in total unemployment dropped by approximately 40%. This decrease is mostly compensated by an increase in *Re-Entrants* for which information on occupation is available. Therefore, the decrease in the number of *New Unemployed Entrants* after 1994 implies that the pool of unemployed with an occupation code increases and it is required to increase unemployment over the 1976-1993 time period to account for this change. Moreover, the increase in unemployment in the service industry reported by Polivka and Miller (1998) is consistent with the significant increase in low skill unemployment shown in Table 5.



Monthly population stocks in millions for the period Jan-1976 to Dec-2010. The corrected series are displayed in blue and the original series in red. The estimated seasonal component is also removed from the original series. Shaded areas show recessions periods as defined by the NBER.

Figure 6: Corrected Stocks

The decrease in the number of *New Unemployed Entrants* also explains why the percentage adjustment in the aggregate unemployment rate is much greater (not displayed in Table 5 but equal to 8% on average) than reported by Polivka and Miller (1998) who find a non significant increase of around 1%. In order to provide further support to the estimated effects obtained in this paper, I also downloaded the seasonally unadjusted unemployment rate series from the BLS website. I then estimated a multiplicative UC model assuming that the 1994 redesign χ_t, is simply a constant and without any classification and population changes (see Section 3.1). The estimate for the 1994 redesign indicates a 4.1% increase in the unemployment rate after the redesign with a t-statistic of -1.89. This estimate is still slightly greater than found by Polivka and Miller (1998) but consistent with their findings, the effect is only statistically significant at a level of 10% or more.[25] Polivka and Miller (1998) also estimate a significant increase in the employment to population ratio and in the labor force participation rate of 0.5% and 0.6% while I find an average decrease of .1% in the employment to population ratio and an average increase of .4% in the labor force participation rate. The decrease in the employment to population ratio stems only from the large increase in unemployment since the decrease in inactivity has a positive effect on the employment to population ratio.

For classification and population changes, Table 5 shows that estimated effects vary in signs and magnitude between occupations. This suggests a non proportional effect of classification and population changes on occupation-specific series and highlights the relevance of performing these corrections. The 1976-82 classification change leads to a reallocation between middle and low skill employment while the adjustment on unemployment affects only middle skill occupations (see See Appendix A.2.5). With regard to population changes, the 1976-79 and 1990 changes have larger effects on, respectively, high and low skill employment. The 1990 population effect is modelled for high skill employment but the average effect reported in Table 5 is quite small. Consistent with the evidence displayed in Figure 1, the 2003 classification and population change results in an increase of high skill employment and a decrease in middle/low skill employment. The net effect of these corrections is however, a bit smaller than the estimate obtained from aggregate series for employment (2.75 vs 3.19 millions on average) but higher for unemployment (0.36 vs 0.13 millions).

Turning to gross flows from employment (Table 6), the aggregate results indicate that the 1994 redesign has a significative effect on EE and EI flows which both need to be increased by 2 and 0.29 millions on average from 1976 to 1993. The sum of this effect (-2.33 millions) should correspond to around 70% of the effect found for the employment stock since $E_{t-1} = EE_t + EU_t + EI_t$ and 30% of the CPS sample is loss due to the matching required to compute flows. However no effect of the redesign is estimated for the aggregate employment stock and for stocks by occupations, the estimated effects compensate each other such that the net effect is close to 0 too (Table 5). This difference can be understood from Figure 4 which shows that stocks obtained from flows represent around 65% of actual stocks before 1994 and around 70% after. The larger increase in flow series brings the percentage before 1994 to 70% consistent with what is observed after 1994. The population changes of 1990 and 2003 are only estimated for EE gross flows whereas the 1976-79 correction results in an increase in both the EE and EU gross flows.

For employment flows by occupation, the correction for the 1994 redesign leads to a decrease in flows between different occupations (e.g. in E^hE^m, E^hE^l, E^hU^m ...) mostly compensated by an increase in flows to the same occupations (E^hE^h, E^hU^h ...). This correction likely captures the effect of the *dependent interviewing technique* which reduced spurious transitions between occupations after 1994. The increase in the EE gross flow originates primarily from middle and high skill occupations. This is consistent with the results obtained for stocks which show increases in high and middle skill employment. The increases in flows to inactivity are significant for all occupations. The 1976-1982 classification change implies a decrease in flows from middle skill employment, and an increase in

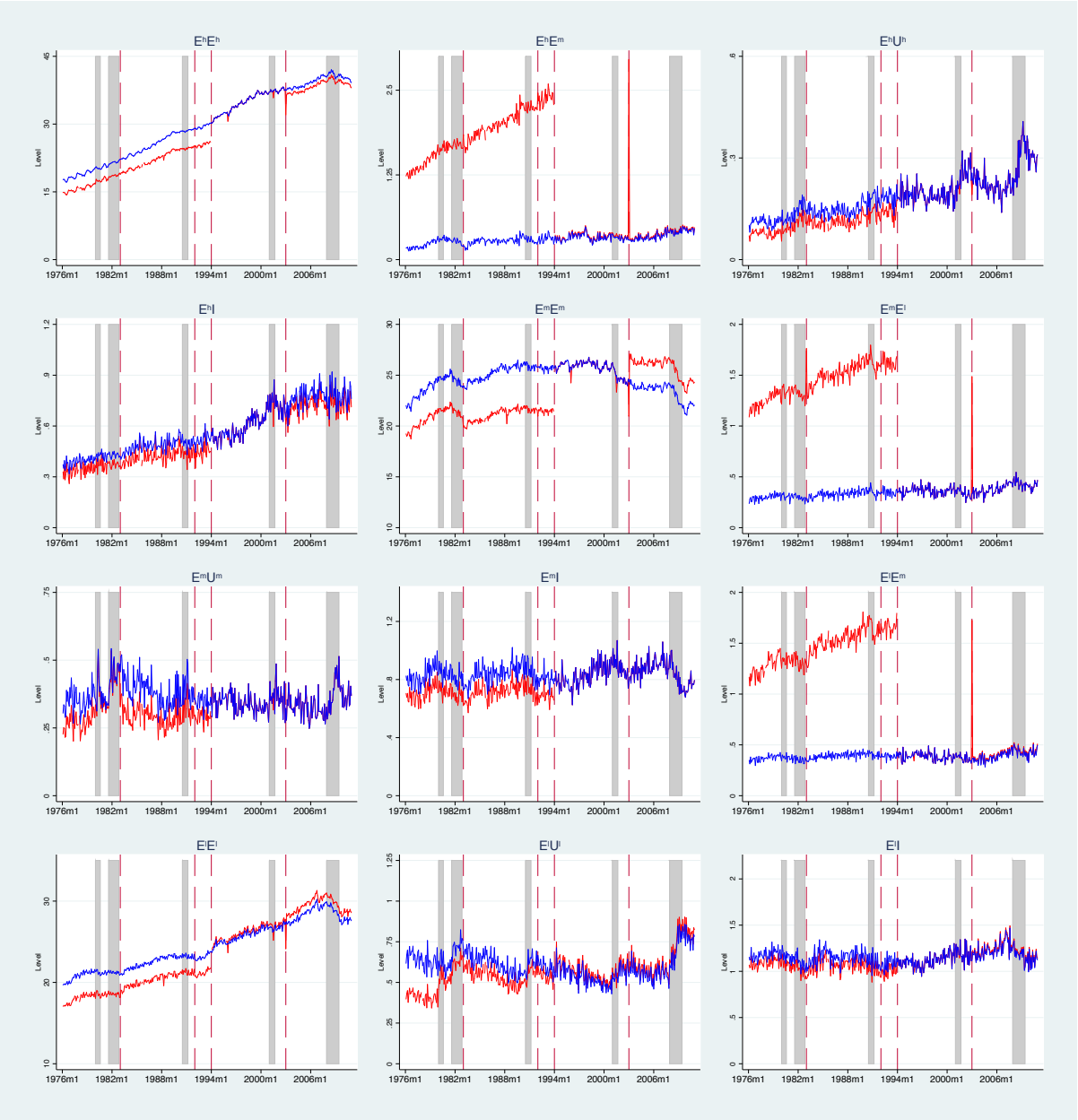
[25]These results are not reported in this paper but are available. Furthermore, I seasonally adjust this unemployment rate series from the BLS and compare the resulting series with the official seasonally adjusted series released by the BLS. Figure 18 in Appendix A.2.7 shows that both series are very similar which suggests that the seasonal adjustment procedure applied in this paper does not lead to inconsistent results.

flows from low skill employment. However, the net effect for this correction (0.22=-0.06+0.93-0.65) is marginally positive whereas it is zero for employment stocks. For population revisions, the 1976-79 correction has larger effects on high skill flows, the 1990 correction affects mostly lows skill flows and the 2003 classification/population change increases flows from high skill employment and decreases those from middle/low skill employment.

Flows	Redesign	Classification	Population		Classif. and pop.
	1976-93	1976-82	$1976-79$	1990	2003-10
Aggregate					
EE	-2^{***}	–	-1.1^{***}	0.43	2.1^{***}
EU	-0.04	0.11^{**}	-0.25^{***}	–	–
EI	-0.29^{***}	–	-0.01	–	–
$\sum Ej$	-2.33	0.11	-1.36	0.43	2.1
Skills					
E^hE^h	-3.2^{***}	–	-0.46^{***}	–	-1.2^{***}
E^hE^m	1.6^{***}	–	–	0.02	–
E^hE^l	0.83^{***}	-0.06^{**}	–	–	–
E^hU^h	-0.03^{***}	–	-0.01	–	–
E^hU^m	0.02^{**}	–	-0.01^{*}	–	–
E^hU^l	0.02^{***}	–	–	–	–
E^hI	-0.06^{***}	–	–	–	-0.05^{*}
$\sum E^hj$	-0.83	-0.06	-0.48	0.02	-1.25
E^mE^h	1.6^{***}	–	-0.13^{***}	0.08	–
E^mE^m	-4^{***}	0.88^{***}	–	–	2.4^{***}
E^mE^l	1.1^{***}	-0.05	-0.04	–	–
E^mU^h	0.02^{***}	–	-0.01^{***}	–	–
E^mU^m	-0.11	0.09^{***}	-0.05^{**}	–	–
E^mU^l	0.03	0.01^{*}	-0.01^{*}	–	–
E^mI	-0.12^{***}	–	–	–	–
$\sum E^mj$	-1.48	0.93	-0.23	0.08	2.4
E^lE^h	0.79^{***}	-0.07^{**}	–	–	–
E^lE^m	1.1^{***}	-0.03	–	–	0.01
E^lE^l	-2.1^{***}	-0.56^{***}	-0.15	0.28	0.63^{**}
E^lU^h	0.01^{***}	–	-0.01^{***}	–	$-^{**}$
E^lU^m	0.05^{***}	–	-0.01^{**}	–	–
E^lU^l	-0.08^{***}	–	-0.14^{***}	0.03	–
E^lI	-0.09^{**}	–	–	–	0.02
$\sum E^lj$	-0.31	-0.65	-0.31	0.31	0.67

Average effects of the 1994 redesign, the 1976-1982 classification change, the 1976-79 and 1990 population changes and the 2003 classification/population change over the periods during which these changes affect series (e.g. 1976-1994 for the redesign). These are expressed in millions and such that they are the quantities that need to be subtracted from the original series (a negative effect implies that the series needs to be increased). $*$, $**$, $***$ indicate the significance level at 10%, 5% and 1%. A "-" means that the effect was not included into the model specification.

Table 6: Flows : 1994 Redesign, Classification and Population Changes (1)



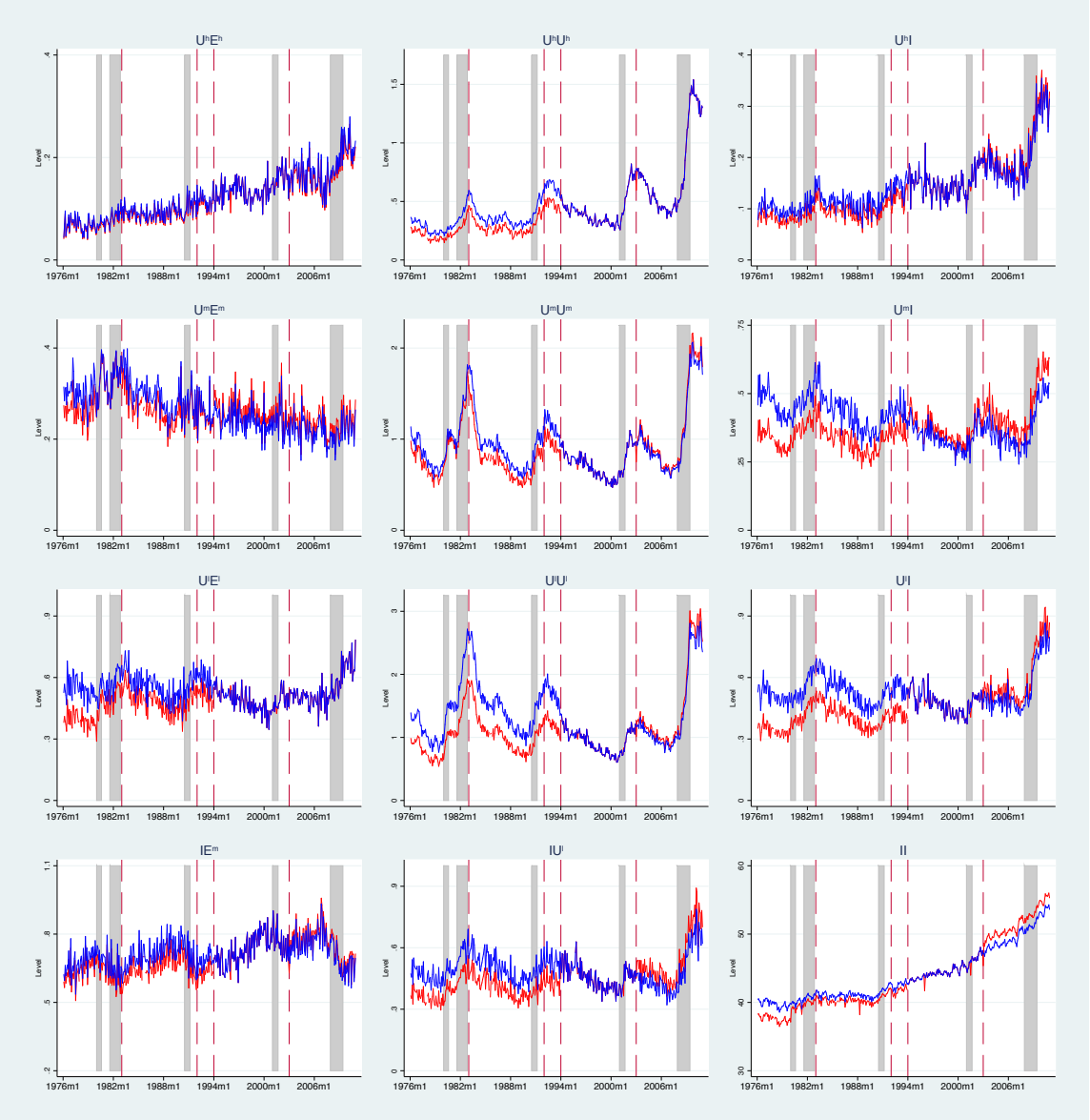
Monthly gross flows in millions for the period Jan-1976 to Dec-2010. The corrected series is displayed in blue and the original series in red. The estimated seasonal component is also removed from the original series. Shaded areas show recessions periods as defined by the NBER.

Figure 7: Corrected Flows (1)

For flows from Unemployment and from Inactivity (Table 7), the corrections for the 1994 redesign leads to significant increases in all aggregate gross flows. These increases are in line with observations made by Abraham and Shimer (2001) and Cortes et al. (2016) who claim that the 1994 redesign increased flows between unemployment and inactivity. In particular, Abraham and Shimer (2001) adjust their series by decreasing the UI flow rates by 3.15 percentage points (pp) after 1994 and compensating this by an increase in the UU flow rates. Additionally, they decrease the IU flow rate by 0.34 pp after 1994 which they offset with a raise of the II flow rate.[26] Similar computations using the results in

[26] Abraham and Shimer (2001) estimate regressions of detrended flow rates on detrented employment to population ratio. They claim that the residuals from these regressions for flow rates between unemployment and inactivity change levels after 1994. They then remove the average value of these residuals from the respective series.

Table 7 imply that on average, the redesign raises the UI flow rate by 1.70pp and decreases the UE and UU flow rates by 0.6pp and 1.1pp. Although the sign of the corrections are the same, the size of the adjustment is smaller and this increase is compensated by a decrease in both the UE and UU flow rates.[27] This could come from the differences in methodology and/or samples due to the absence of *New Unemployed Entrants* in my analysis. The estimated effects for flows from inactivity imply average increases in the IU and IE flow rates of 0.46pp and 0.28pp respectively, which are offset by a decrease in the II flow rate of 0.74pp. The effect for the IU flow rates is similar to the effect found by Abraham and Shimer (2001) although the IE flow rate is also adjusted resulting in a larger decrease in the II flow rate.



Monthly gross flows in millions for the period Jan-1976 to Dec-2010. The corrected series are displayed in blue and the original series in red. The estimated seasonal component is also removed from the original series. Shaded areas show recessions periods as defined by the NBER.

Figure 8: Corrected Flows (2)

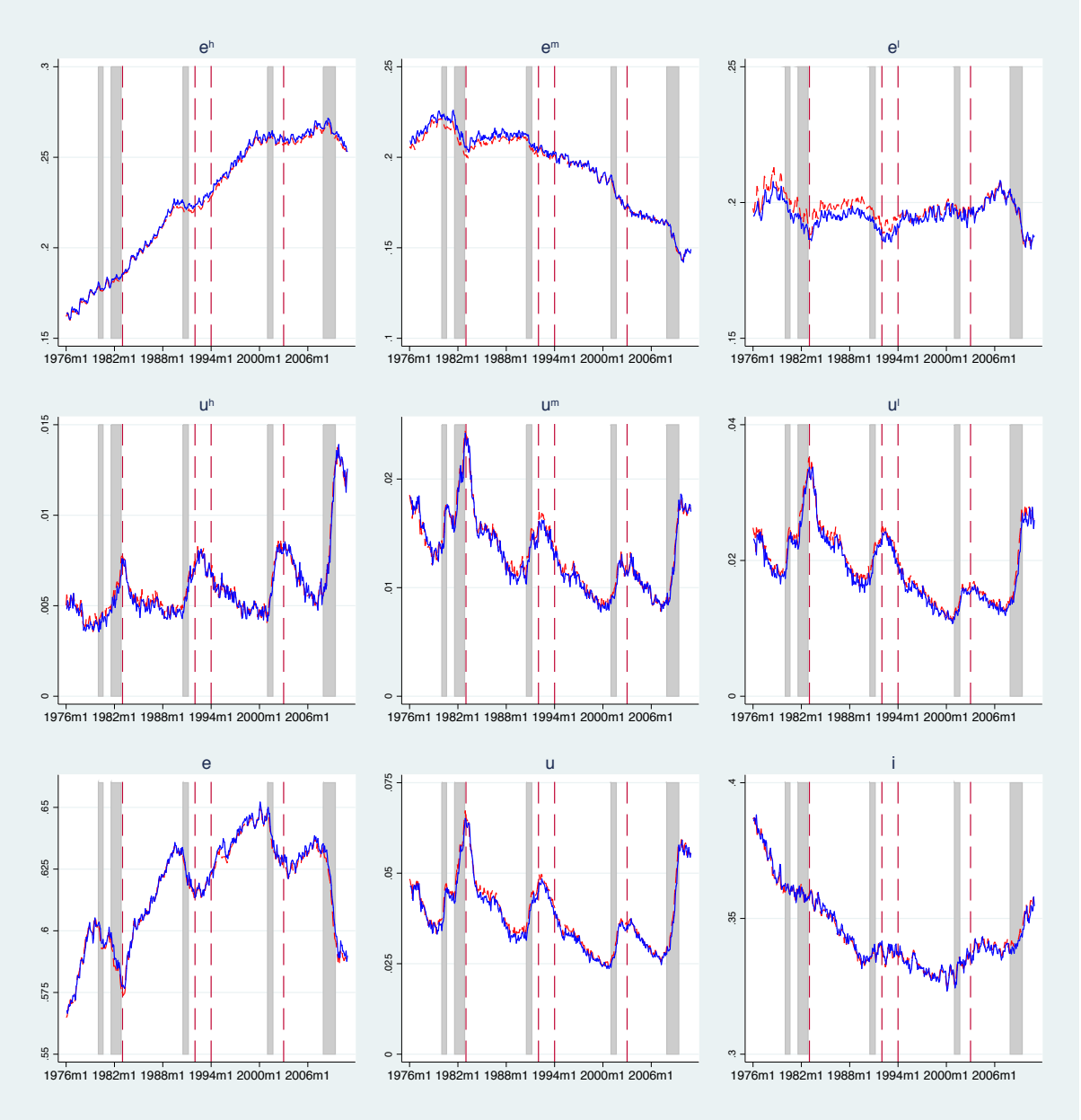
[27]Note that Abraham and Shimer (2001) correct the UI series after 1994 (by decreasing it), while the correction in this paper is applied to the series before 1994 (by increasing it).

Flows	Redesign	Classification	Population		Classif. and pop.
	1976-93	1976-82	$1976-79$	1990	2003-10
Aggregate					
UE	-0.17^{***}	0.04	-0.13^{**}	0.03	-
UU	-0.36^{**}	0.13	-0.09	-	0.08
UI	-0.23^{***}	0.01	-0.09^{*}	0.02	0.1
$\sum Uj$	-0.76	0.19	-0.31	0.05	0.18
Skills					
U^hE^h	-0.01	-	-	-	-0.01
U^hE^m	-0.01	-	-	-	-0.01
U^hE^l	0.00	-	-	-	-0.00
U^hU^h	-0.09^{***}	-	-	-	-
U^hU^m	0.04^{***}	-	-0.01	-	0.00
U^hU^l	0.02^{**}	-	-	-	-
U^hI	-0.02^{*}	-	-0.00	-	0.01
$\sum U^hj$	-0.05	-	-0.01	-	-0.01
U^mE^h	0.00	-	-0.00^{**}	0.00	-
U^mE^m	-0.03^{*}	0.03^{*}	-0.04^{**}	0.01	-
U^mE^l	-0.02^{*}	0.01	—	0.01	-
U^mU^h	0.04^{**}	—	-0.01	-	0.00
U^mU^m	-0.16^{**}	0.1	-0.07	-	0.05
U^mU^l	0.09^{***}	-	-	-	-
U^mI	-0.1^{***}	0.02	-0.04^{*}	0.03	0.03
$\sum U^mj$	-0.19	0.15	-0.16	0.05	0.09
U^lE^h	0.00^{***}	-	-0.01^{***}	-	-
U^lE^m	-0.02^{***}	-	-	-	-
U^lE^l	-0.07^{***}	-	-0.08^{***}	-	-
U^lU^h	0.02^{*}	-	-	-	-
U^lU^m	0.09^{***}	-	-0.03^{**}	—	0.01^{***}
U^lU^l	-0.43^{***}	-	-	—	0.1
U^lI	-0.12^{***}	-	-0.05^{***}	-	0.05
$\sum U^lj$	-0.53	-	-0.17	-	0.15
Aggregate					
IE	-0.17^{***}	-	-	-	-0.01
IU	-0.22^{***}	-	-0.03	-	0.05
II	-0.77^{***}	-	-1.4^{***}	-	1.6^{***}
Skills					
IE^h	-0.02^{*}	-	-	-	-0.02
IE^m	-0.05^{**}	0.01	-	-	0.02
IE^l	-0.09^{***}	-	-	-	0.02
IU^h	-0.02^{**}	-	-	-	-0.03
IU^m	-0.1^{***}	-	-0.02	-	0.02
IU^l	-0.09^{***}	-	-0.02	-	0.06^{*}
$\sum I^j$	-1.14	0.01	-1.44	—	1.68

Average effects of the 1994 redesign, the 1976-1982 classification change, the 1976-79 and 1990 population changes and the 2003 classification/population change over the periods during which these changes affect series (e.g. 1976-1994 for the redesign). These are expressed in millions and such that they are the quantities that need to be subtracted from the original series (a negative effect implies that the series needs to be increased). $*$, $**$, $***$ indicate the significance level at 10%, 5% and 1%. A "-" means that the effect was not included into the model specification.

Table 7: Flows : 1994 Redesign, Classification and Population Changes (2).

The estimation results in terms of occupations are in line with the results obtained for stocks. The correction for the 1994 redesign leads to larger increases for gross flows from low skill unemployment and the 1976-1982 classification only decreases flows from middle skill unemployment. The 1976-79 population change seems to have a significant impact on the unemployment to employment (UE) gross flow, particularly on the U^lE^l and U^mE^m gross flows. The 1990 and 2003 corrections only appear to result in marginal adjustments to the original series. This matches the evidence presented in Table 5 which show no significant effects of these 2 population changes on unemployment stocks. The results for gross flows from inactivity indicate that the 1994 redesign mostly affects flows to middle and low skill occupations (both to employment and unemployment) as well as the II gross flow. The II gross flow is also the one mostly affected by the population changes of 1976-79 and 2003.



Monthly stocks obtained from correcting the stock series in blue (equation (12)) against the stocks obtained from the corrected flow series in red (equation (13)). These are expressed in terms of total population. The last row displays the (un)employment and inactivity to population ratio (e, u and i). Shaded areas show recessions periods as defined by the NBER.

Figure 9: Stock vs Flows

Overall, these results show that when working with occupation data from the CPS, the 1994 redesign has a significant effect on unemployment stocks through the changes in the definition of *New Unemployed Entrants*. This change in the composition of unemployment also significantly affect most gross flow series from unemployment. The introduction of the *dependent interviewing technique* further implies that the 1994 redesign has significant effects on flows from employment, in particular those between occupations. The results obtained for aggregate flows between unemployment and inactivity confirm the evidence in Abraham and Shimer (2001) on the impact of the 1994 redesign on these flows. The results also suggests that the use of Autor and Dorn (2013) classification requires adjusting employment series (both stock and flows) while effects on unemployment series are usually small. A reallocation between middle and low skill employment series has to be applied for the 1976-82 classification and the 2003-10 change requires increasing high skill employment series and decrease middle/low skill employment series. The 2003-10 classification change is also a large population revision which also requires to decrease inactivity series. The 1990 population revision leads to an adjustment, mostly in low skill employment stock and flows. From all the population revisions tested in this work, only the 1976-79 seems to have significant effects on all series of stocks and flows.

A last comment on these corrections regards the steps that were undertaken in order to obtain estimated effects consistent between stocks and flows series. Figure 9 compares the stocks computed from (12) to those obtained from (13) and based on gross flow series. This figure indicates that, for unemployment and inactivity (expressed in terms of total population), the stocks computed from corrected gross flows series are quite similar to the series obtained using corrected stocks. A slight difference can be seen for middle and low skill employment series (of about 0.5pp on average between flows and stocks series). This discrepancy primarily results from the estimates of 1994 redesign component. The correction for this effect leads to a larger increase in low skill employment when computed from gross flows than from the stocks and the opposite effect applies for middle skill employment. However, the employment to population ratio ($e = e^h + e^m + e^l$) displayed in the last row of Figure 9 is very similar regardless of whether it is computed from flows or stocks. Finally, the corrections for the Margin of Adjustment implemented in the next section will make the series of flows and stocks perfectly consistent.

4 Margin of Adjustment and Time Aggregation

In this section, I present the framework used to perform the last 2 corrections. The Margin of Adjustment correction was first studied by Abowd and Zellner (1985) and Poterba and Summers (1986) which inspired the correction proposed by Elsby et al. (2015). I only have to adapt their 3 states framework to the 7 states framework used in this paper. On the other hand, the correction for Time Aggregation relies on the link between Discrete Time Markov Chain (DTMC) and Continuous Time Markov Chain (CTMC). Most of the computations presented in the following sections can be found in Norris (1997) and Elsby et al. (2015). From now on, I work with flow rates p_t^{ij} and population stocks expressed in terms of total population (see equation (12)) computed after the corrections of Section 3.

4.1 Discrete Time Markov Chains

Both the Margin of adjustment and the Time Aggregation corrections assume that labor market stocks evolve according to a DTMC :

$$\underbrace{\begin{bmatrix} e^h \\ e^m \\ e^l \\ u^h \\ u^m \\ u^l \\ i \end{bmatrix}_t}_{s_t} = \underbrace{\begin{bmatrix} p^{E^hE^h} & p^{E^mE^h} & p^{E^lE^h} & p^{U^hE^h} & p^{U^mE^h} & p^{U^lE^h} & p^{IE^h} \\ p^{E^hE^m} & p^{E^mE^m} & p^{E^lE^m} & p^{U^hE^m} & p^{U^mE^m} & p^{U^lE^m} & p^{IE^m} \\ p^{E^hE^l} & p^{E^mE^l} & p^{E^lE^l} & p^{U^hE^l} & p^{U^mE^l} & p^{U^lE^l} & p^{IE^l} \\ p^{E^hU^h} & p^{E^mU^h} & p^{E^lU^h} & p^{U^hU^h} & p^{U^mU^h} & p^{U^lU^h} & p^{IU^h} \\ p^{E^hU^m} & p^{E^mU^m} & p^{E^lU^m} & p^{U^hU^m} & p^{U^mU^m} & p^{U^lU^m} & p^{IU^m} \\ p^{E^hU^l} & p^{E^mU^l} & p^{E^lU^l} & p^{U^hU^l} & p^{U^mU^l} & p^{U^lU^l} & p^{IU^h} \\ p^{E^hI} & p^{E^mI} & p^{E^lI} & p^{U^hI} & p^{U^mI} & p^{U^lI} & p^{II} \end{bmatrix}_t}_{P_t} \underbrace{\begin{bmatrix} e^h \\ e^m \\ e^l \\ u^h \\ u^m \\ u^l \\ i \end{bmatrix}_{t-1}}_{s_{t-1}}$$

with $e^h + e^m + e^l + u^h + u^m + u^l + i = 1$, P_t the 7×7 discrete time transition matrix (or stochastic matrix) and p_t^{ij} the flow rate from state i to state j. In a more compact notation, we have:

$$s_t = P_t s_{t-1}$$

$$\begin{aligned} p_t^{ij} &\geq 0 \\ p_t^{ii} &= 1 - \sum_{j \neq i} p_t^{ij} \end{aligned}$$

4.2 Margin of Adjustment

This correction is used to obtain flow rates $\tilde{p}_t^{ij}$, consistent with the evolution of the stock s_t. Elsby et al. (2015) propose to perform the required adjustment through a constrained minimization problem. Defining the 42×1 vectors of corrected and original flow rates $\tilde{p}_t$ and p_t,[28] we have the following minimization problem:

$$\min_{\tilde{p}_t} (\tilde{p}_t - p_t)' W_t^{-1} (\tilde{p}_t - p_t)$$

subject to

$$\begin{aligned} \Delta s_t &= A_{t-1} \tilde{p}_t, \\ \tilde{p}_t &\geq 0 \end{aligned}$$

To derive the constraints, we can use the fact that $p_t^{ii} = 1 - \sum_{j \neq i} p_t^{ij}$ to write the DTMC as:

$$\begin{bmatrix} \Delta e^h \\ \Delta e^m \\ \Delta e^l \\ \Delta u^h \\ \Delta u^m \\ \Delta u^l \\ \Delta i \end{bmatrix}_t = \begin{bmatrix} -\sum_{i \neq E^h} p^{E^h i} & p^{E^mE^h} & p^{E^lE^h} & p^{U^hE^h} & p^{U^mE^h} & p^{U^lE^h} & p^{IE^h} \\ p^{E^hE^m} & -\sum_{i \neq E^m} p^{E^m i} & p^{E^lE^m} & p^{U^hE^m} & p^{U^mE^m} & p^{U^lE^m} & p^{IE^m} \\ p^{E^hE^l} & p^{E^mE^l} & -\sum_{i \neq E^l} p^{E^l i} & p^{U^hE^l} & p^{U^mE^l} & p^{U^lE^l} & p^{IE^l} \\ p^{E^hU^h} & p^{E^mU^h} & p^{E^lU^h} & -\sum_{i \neq U^h} p^{U^h i} & p^{U^mU^h} & p^{U^lU^h} & p^{IU^h} \\ p^{E^hU^m} & p^{E^mU^m} & p^{E^lU^m} & p^{U^hU^m} & -\sum_{i \neq U^m} p^{U^m i} & p^{U^lU^m} & p^{IU^m} \\ p^{E^hU^l} & p^{E^mU^l} & p^{E^lU^l} & p^{U^hU^l} & p^{U^mU^l} & -\sum_{i \neq U^l} p^{U^l i} & p^{IU^h} \\ p^{E^hI} & p^{E^mI} & p^{E^lI} & p^{U^hI} & p^{U^mI} & p^{U^lI} & -\sum_{i \neq I} p^{Ii} \end{bmatrix}_t \begin{bmatrix} e^h \\ e^m \\ e^l \\ u^h \\ u^m \\ u^l \\ i \end{bmatrix}_{t-1} \tag{14}$$

[28] The length of the vector is 42×1 since we have $p_t^{ii} = 1 - \sum_{j \neq i} p_t^{ij}$.

and we can then rework this expression[29] to obtain the 7×42 matrix A_{t-1}:

$$A_{t-1} = \begin{bmatrix} \underbrace{-e^h}_{1\times6} & e^m & \underbrace{0}_{1\times5} & e^l & \underbrace{0}_{1\times5} & u^h & \underbrace{0}_{1\times5} & u^m & \underbrace{0}_{1\times5} & u^l & \underbrace{0}_{1\times5} & & i & & \underbrace{0}_{1\times5} \\ 0 & e^h & \underbrace{0}_{1\times4} & \underbrace{-e^m}_{1\times6} & 0 & e^l & \underbrace{0}_{1\times5} & u^h & \underbrace{0}_{1\times5} & u^m & \underbrace{0}_{1\times5} & u^l & \underbrace{0}_{1\times5} & i & \underbrace{0}_{1\times4} \\ \vdots & & & & & & & \cdots & & & & & & & \vdots \\ \underbrace{0}_{1\times5} & e^h & \underbrace{0}_{1\times5} & e^m & \underbrace{0}_{1\times5} & e^l & \underbrace{0}_{1\times5} & u^h & \underbrace{0}_{1\times5} & u^m & \underbrace{0}_{1\times5} & u^l & & \underbrace{-i}_{1\times6} & \end{bmatrix}_{t-1}$$

For the weighting matrix W_t, Elsby et al. (2015) propose to use the following block matrix:

$$W_t = \begin{bmatrix} W^{E^h} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & W^{E^m} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & W^{E^l} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & W^{U^h} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & W^{U^m} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & W^{U^l} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & W^{I} \end{bmatrix}_t$$

where W_t^i are 6×6 matrices obtained by using the Multinomial distribution assumption (see Section 2.1). $W_t^{E^h}$ for instance, is given by:

$$W_t^{E^h} = \begin{bmatrix} p^{E^hE^m}(1-p^{E^hE^m}) & -p^{E^hE^m}p^{E^hE^l} & -p^{E^hE^m}p^{E^hU^h} & -p^{E^hE^m}p^{E^hU^m} & -p^{E^hE^m}p^{E^hU^l} & -p^{E^hE^m}p^{E^hI} \\ -p^{E^hE^l}p^{E^hE^m} & p^{E^hE^l}(1-p^{E^hE^l}) & -p^{E^hE^l}p^{E^hU^h} & -p^{E^hE^l}p^{E^hU^m} & -p^{E^hE^l}p^{E^hU^l} & -p^{E^hE^l}p^{E^hI} \\ -p^{E^hU^h}p^{E^hE^m} & p^{E^hU^h}p^{E^hE^l} & p^{E^hU^h}(1-p^{E^hU^h}) & -p^{E^hU^h}p^{E^hU^m} & -p^{E^hU^h}p^{E^hU^l} & -p^{E^hU^h}p^{E^hI} \\ -p^{E^hU^m}p^{E^hE^m} & p^{E^hU^m}(1-p^{E^hE^l}) & -p^{E^hU^m}p^{E^hU^h} & p^{E^hU^m}(1-p^{E^hU^m}) & -p^{E^hU^m}p^{E^hU^l} & -p^{E^hU^m}p^{E^hI} \\ -p^{E^hU^l}p^{E^hE^m} & p^{E^hE^l}(1-p^{E^hE^l}) & -p^{E^hE^l}p^{E^hU^h} & -p^{E^hE^l}p^{E^hU^m} & -p^{E^hE^l}p^{E^hU^l} & -p^{E^hE^l}p^{E^hI} \\ -p^{E^hI}p^{E^hE^m} & p^{E^hE^l}(1-p^{E^hE^l}) & -p^{E^hE^l}p^{E^hU^h} & -p^{E^hE^l}p^{E^hU^m} & -p^{E^hE^l}p^{E^hU^l} & -p^{E^hE^l}p^{E^hI} \end{bmatrix}_t E_{t-1}^h$$

These quantities correspond to the variance-covariance matrix of flow rates computed from equation (1). As pointed by Elsby et al. (2015), this adjustment will only have a marginal impact on flow rates.

4.3 Time Aggregation bias

As is explained in Section 2.2, the Time Aggregation bias originates from the discrete nature of data collection. In a 2 state framework, Kaitz (1970) and Perry et al. (1972) show that it is possible to use unemployment spells with a duration smaller than 1 month to recover corrected weekly flow rates. Shimer (2012) proposes a continuous time set-up and uses the same intuition as Kaitz (1970) and Perry et al. (1972) to obtain hazard rates through short term unemployment spells. He further extends his framework to include a third state (inactivity) and proposes a correction also used by Elsby et al. (2015). This correction relies on the connection between Discrete and Continuous Time Markov chains.

[29] For instance the first line of (14) leads to $\Delta e_t^h = -\sum_{i \neq E^h} p^{E^h i} e_{t-1}^h + p^{E^m E^h} e_{t-1}^m + p^{E^l E^h} e_{t-1}^l + \cdots + p^{I E^h} i_{t-1}^h$. Using the vector $\tilde{p}_t$, we can write this equation in vector form as $[-e^h \ -e^h \ -e^h \ -e^h \ -e^h \ -e^h \ e^m \ 0 \ 0 \ 0 \ 0 \ 0 \ e^l \ 0 \ldots]_{t-1}\tilde{p}_t$.

4.3.1 Discrete and Continuous Time Markov Chains

A CTMC can be defined in the following way:

$$\dot{s} = Fs_t$$

with F_t, the 7×7 *infinitesimal generator* matrix of the CTMC satisfying

$$\begin{aligned} 0 \leq -f_t^{ii} \leq \infty \\ f_t^{ij} \geq 0 \\ \sum_j f_t^{ij} = 0 \end{aligned}$$

where f_t^{ij} are instantaneous transitions rates (hazard rates) of moving from state i to state j and $f_t^i \equiv -f_t^{ii}$ can be interpreted as the staying rate in state i.

The question of whether a valid and unique generator matrix F_t can be obtained from discrete time transitions is called the *embeddability problem* and is an issue tackled in many areas of science. In theory, there exists a tight link between the stochastic matrix P_t and the generator matrix F_t. More specifically, the transition probabilities associated with the continuous time process over the time interval $n \in [0;\ T]$ have to satisfy the following differential equations (see Norris (1997)):

$$\dot{P}_t(n) = P_t(n)F_t$$

with $P_t(0) = I$ the identity matrix and $P_t(n)$, the discrete time transition matrix for month t over the interval of time $n \in [0;\ T]$. Note that in this paper, the unit of time is a month which implies that month t ends when $n = 1$.[30] Thus $P_t(1)$ is the discrete time transition matrix at the end of month t which is the transition matrix computed from the data, P_t. These differential equations are known as the *Chapman-Kolmogorov* equations and have solution $P_t(n) = e^{F_t n}$ (see Norris (1997)). They imply that at the end of month t (or $n = 1$), P_t is equal to the matrix exponential of F_t. Therefore, the matrix of hazards rates F_t can be obtained by computing the logarithm of the discrete time matrix P_t. Provided that the eigenvalues of P_t are positive, real and distinct, this can be achieved through an eigenvalue decomposition $P_t = V_t D_t V_t^{-1}$ with D_t the diagonal matrix of eigenvalues, taking the log of the eigenvalues and computing the generator matrix using these modified eigenvalues and the eigenvectors V_t. This is the solution pursued by Shimer (2012) and he further proposes to compute the corrected transition probabilities as $\hat{p}_t^{ij} = 1 - e^{(-f_t^{ij})}$. These quantities correspond to the probabilities of moving from state i to j given that no transitions to other states are observed (see Shimer (2012)).

However, I cannot implement the eigenvalue decomposition of the matrix P_t since complex eigenvalues are obtained for specific periods.[31] Israel et al. (2001) show that it is possible to circumvent this issue and approximate the generator matrix through the following infinite sequence:

$$F_t = P_t - I + \frac{(P_t - I)^2}{2} + \frac{(P_t - I)^3}{3} + \dots \tag{15}$$

This sequence does not guarantee convergence to a valid generator matrix for the continuous time Markov chain as it will, in some instances, lead to negative off diagonal hazard rates.[32] Israel et al.

[30] Furthermore, given that F_t is the month t transition matrix, the length of the time interval also has to be a month implying $T = 1$. Otherwise, $n > 1$ would imply a transition to month $t + 1$ for which the transition matrix should be F_{t+1}.

[31] Complex eigenvalues (which appear in conjugate pairs) are obtained for 25 periods out of the 419 in the sample (around 6% of the sample). These complex eigenvalues always have modulus smaller than 1 and do not constitute any issues except for implementing Shimer's proposed correction.

[32] It should be noted that the issue of negative off diagonal hazards also arises when the eigenvalues of the discrete time transition matrix are positive, real and distinct.

(2001) provide conditions under which a valid generator matrix cannot be recovered from discrete time data. One of these conditions is that for some states i and j, j is accessible from state i but directly transitioning from i to j is not possible ($p_t^{ij} = 0$). In the context of this paper, the transition rate from middle skill unemployment to high skill unemployment is bigger than 0 but transitioning from high skill employment directly to middle skill unemployment is seldom observed ($p_t^{E^hU^m} \approx 0$). This explains why the generator matrix obtained from the eigenvalue decomposition or from the infinite sequence (15) often results in negative off-diagonal hazard rates f_t^{ij} and a non-valid generator matrix for the CTMC.

Furthermore, some hazard rates should be equal to 0 as it is impossible to instantaneously transition between some labor market states. This restrictions comes from the fact that the occupation of an unemployed should be the last one she worked in. For instance, it is impossible to directly transition between high skill employment and middle skill unemployment ($f^{E^hU^m} = 0$). This implies that some elements of the generator matrix F_t have to be restricted to 0, which is not possible (as far as I know) through an eigenvalue decomposition. More precisely, the *infinitesimal generator* matrix F_t has the following form:

$$F_t = \begin{bmatrix} f^{E^h} & f^{E^mE^h} & f^{E^lE^h} & f^{U^hE^h} & f^{U^mE^h} & f^{U^lE^h} & f^{IE^h} \\ f^{E^hE^m} & f^{E^m} & f^{E^lE^m} & f^{U^hE^m} & f^{U^mE^m} & f^{U^lE^m} & f^{IE^m} \\ f^{E^hE^l} & f^{E^mE^l} & f^{E^m} & f^{U^hE^l} & f^{U^mE^l} & f^{U^lE^l} & f^{IE^l} \\ f^{E^hU^h} & 0 & 0 & f^{U^h} & 0 & 0 & f^{IU^h} \\ 0 & f^{E^mU^m} & 0 & 0 & f^{U^m} & 0 & f^{IU^m} \\ 0 & 0 & f^{E^lUl} & 0 & 0 & f^{U^l} & f^{IU^l} \\ f^{E^hI} & f^{E^mI} & f^{E^lI} & f^{U^hI} & f^{U^mI} & f^{U^lI} & f^{I} \end{bmatrix}_t \tag{16}$$

Nonetheless, many corrections have been proposed to obtain a generator matrix when the eigenvalue decomposition of the transition matrix is not possible. These range from simple adjustments of the problematic hazard rates to more developed techniques such as the use of the EM algorithm. Inamura (2006) proposes a review of a subset of possible solutions. To tackle this issue, Bladt and Sørensen (2005) suggest a bayesian approach which is the solution pursued in this paper. As argued by Bladt and Sørensen (2005), this solution presents the key advantage of ensuring the existence and estimation of a valid generator matrix F_t since non-negativity constraints can be imposed on the hazard rates through a proper choice of prior distributions. The zero constraints on some of the hazard rates can also be imposed in a straightforward manner. The next section presents the main aspect of the estimation method proposed by Bladt and Sørensen (2005). More technical details on this method and on the issue of existence of the generator matrix can be found in their paper. Lastly, it should be noted that this correction will imply $e^{F_t} \approx P_t$.

4.3.2 A Bayesian Approach

The vector of parameters Θ contains 30 hazard rates that have to be estimated since $f^i = \sum_{i \neq j} f^{ij}$ and 12 hazard rates are restricted to 0. Using Bayes rule, the posterior distribution for the parameters vector Θ conditional on data X, $p(\Theta|X)$, is given by:

$$\begin{aligned} p(\Theta|X) &= \frac{p(X|\Theta)p(\Theta)}{p(X)} \\ &\propto p(X|\Theta)p(\Theta) \end{aligned}$$

meaning that the posterior distribution is proportional to the product of the likelihood function $p(X|\Theta)$ and the prior distribution $p(\Theta)$. Over the interval of time $[0;\ T]$, the likelihood function is given by:

$$L = \prod_{i=1}^{K} \prod_{j \neq i} e^{f^i R^i(T)} f^{ij\, N^{ij}(T)} \tag{17}$$

where K is the total number of states, $R^i(T)$ and $N^{ij}(T)$ are respectively defined as the total amount of time spent in state i and the total number of transitions from state i to state j by time T. The derivation of the likelihood function is presented in Appendix A.3.1.

For the prior distribution, Bladt and Sørensen (2005) argue in favor of the use of a Gamma distribution, since it allows to impose a non-negativity constraint on the hazard rates. This distribution is also a conjugate prior for the likelihood function that provides a closed form expression for the posterior distribution. The prior distribution is:

$$p(\Theta) = \prod_{i=1}^{K} \prod_{j \neq i} f^{ij\, \alpha^{ij} - 1} e^{-f^{ij} \beta^i} \tag{18}$$

where $\alpha^{ij} > 0$ and $\beta^i > 0$ are the shape and rate parameters of the Gamma distribution which do not have to be estimated. The posterior distribution is then equal to:

$$p(\Theta|X) = \prod_{i=1}^{K} \prod_{j \neq i} f^{ij\, N^{ij}(T) + \alpha^{ij} - 1} e^{-f^{ij}(R^i(T) + \beta^i)}. \tag{19}$$

Therefore, the posterior distribution follows a gamma distribution with shape parameter $N^{ij}(T) + \alpha^{ij}$ and rate parameter $R^i(T) + \beta^i$.

In order to simulate draws from the posterior distribution, Bladt and Sørensen (2005) propose to use Gibbs sampling by alternatively drawing from the conditional distributions $p(\Theta|X)$ and $p(X|\Theta)$. More specifically, an initial draw for the parameters $\Theta^{(0)}$ can be obtained from the prior distribution (18). Given these hazard rates, we subsequently have to draw $X^{(1)}$ from $p(X|\Theta^{(0)})$ and a new set of parameters $\Theta^{(1)}$ from $p(\Theta|X^{(1)})$. These 2 steps are then repeated G times creating a sequence $\{\Theta^{(g)}, X^{(g)}\}_{g=1}^{G}$. After discarding part of the initial sequence (burn-in period), the estimated hazard rates can be obtained by computing moments (e.g. the mean) from the series of simulated hazard rates.

The closed form expression for the posterior distribution (19) simplifies the drawing for the sequence of hazard rates $\Theta^{(g)}$. To draw $X^{(g)}$ from $p(X|\Theta^{(g-1)})$, Bladt and Sørensen (2005) propose to use a rejection-sampling algorithm. For a given month t, this algorithm requires simulating individual Continuous Time Markov Chains to reproduce the total number of gross flows from each states observed in the data. When the total number of transitions for a given gross flow (e.g. $E^h E^h$) has been reached, all simulated transitions that would lead to this gross flows are discarded. The quantities R^i and N^{ij} can then be computed from these simulations and used to draw new hazard rates. The simulations of CTMCs also raise questions on how to record the labor market status for individual simulations. In the CPS, an individual has to work in the week before she is interviewed to be recorded as employed. This means that an individual who worked for a few days during the reference week but he's out of work (either inactive or unemployed) at the moment of the interview would still be recorded has employed. This aspect has not been taken into account when simulating CTMCs since I assume that the recorded labor market state is the one at the end of the simulation. This issue is also quickly discuss by Elsby et al. (2015) who points that the Time Aggregation correction they apply assumes a contemporaneous mapping between labor market activities and the recorded status of the worker while in practice, the recorded status also depends on past activities. The framework presented in this paper should be able to account for this dynamic mapping through the simulated CTMCs which provide the labor market status for the entire simulated period. This point is further discussed in Appendix

A.3.2. This appendix also gives more details on the rejection sampling algorithm and the simulation of CTMCs.

The full process can be summarized as follows:

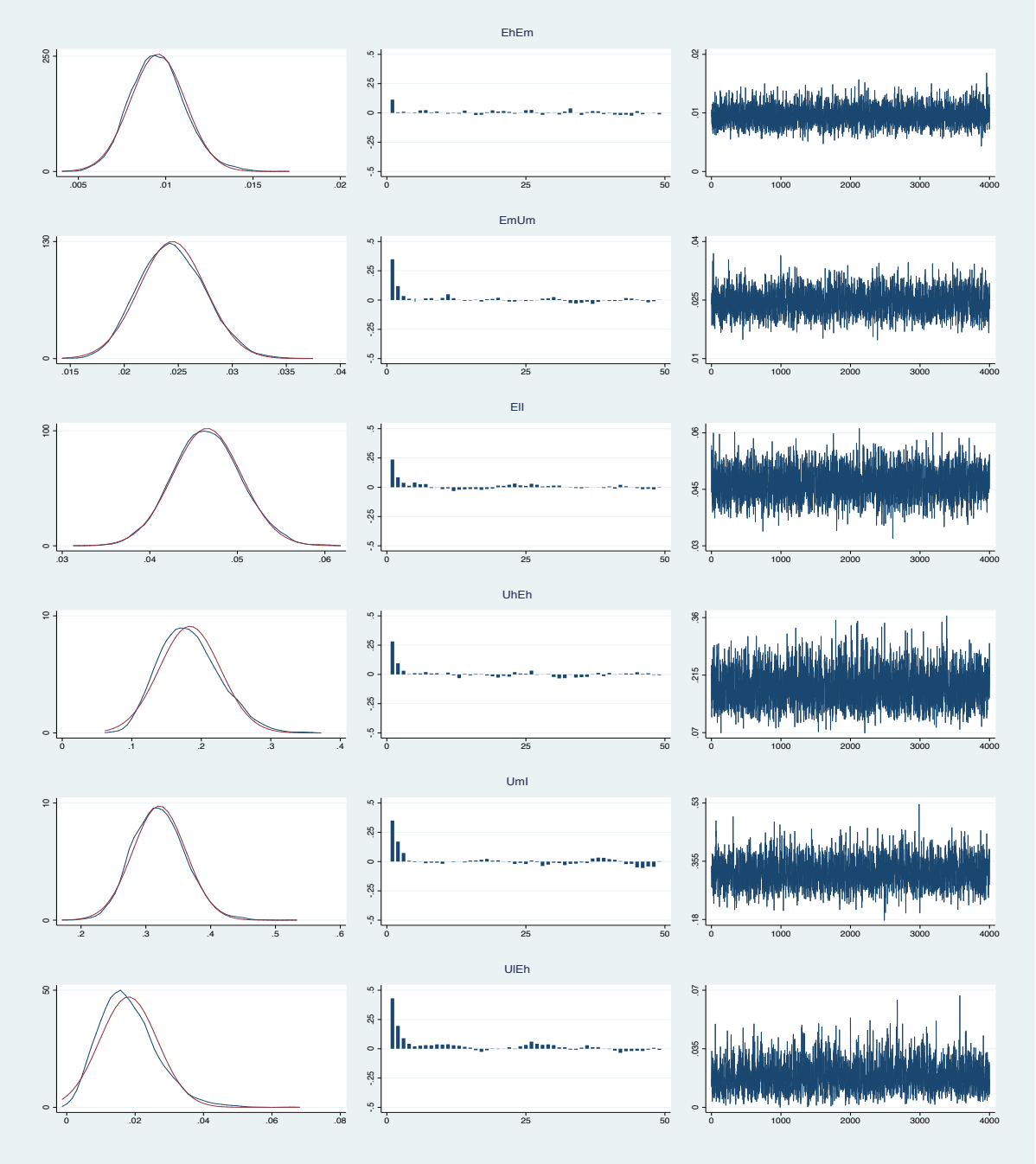
0. Start from the first period t in the sample (i.e. February 1976)
1. Draw initial hazard rates, ${f_t^{ij}}^0$ from the prior distribution (18).
2. Use these hazards to simulate CTMCs until the observed number of gross flows in t is reached for each labor market state.
3. Compute R^i and N^{ij} from the simulated CTMCs.
4. Draw new hazard rates f_t^{ij} from the posterior distribution (19).
5. Iterate on step 2-4 G times and compute statistics of interest (mean or median) from the series of hazard rates.
6. Repeat step 1-5 for the next period in the sample.

The number of replications G is set to 5000 which is twice as low as the number used by Bladt and Sørensen (2005) and Inamura (2006) who perform 10000 replications. Reducing the number of replication is primarily motivated by the fact this procedure has to be repeated for all periods in the sample (419 periods). Therefore, lowering G allows for a significant time gain. The parameters of the prior distribution are set following Bladt and Sørensen (2005) and Inamura (2006). They choose $\alpha^{ij} = \beta^i = 1$. A plot of the prior distribution is displayed at the bottom of Figure 11. Once series of hazard rates have been obtained for a given period t, the first 10% of each series is dropped. The estimated hazards are obtained by computing the median rather than the mean as some posterior distributions appear to be skewed.

Some diagnostic checks are performed in order to assess the convergence of the Gibbs sampler (see for instance Cowles and Carlin (1996)). No actual and fully satisfactory diagnoses are available and it is usually recommended to perform both graphical analysis and tests based on simulations of multiple chains. However, due to the fact that the estimation is repeated 419 times (for all periods), simulating multiple chains for each period would be extremely time consuming. As a result, mostly graphical assessments are performed. These assessments involve visually inspecting the sequence for each hazard rates, checking the autocorrelation function for the simulated series, and a plot of the posterior density. Figures 10 and 11 display these checks for some selected hazard rates and periods. These figures seem to indicate that there are no real complications with regard to the simulations. In particular, simulated series (3rd column) look stationary (constant mean and (co)-variance). There seems to be some significant autocorrelation at lag 1 for some series (2nd column) but this is usually a common feature of Markov Chain Monte Carlo (MCMC) procedures. Note that the autocorrelation cofficients are never bigger than 0.5 and decrease very quickly to values close to 0 after lag 1. The posterior densities (1rst column) are close to normal densities for most series. These plots also show that posterior densities are quite different from the prior density (Figure 11). This suggests that the choice of prior has a limited influence on the simulated posterior apart from imposing non-negativity constraints on the estimated hazards.

I also perform Geweke tests (see Geweke (1992)). These tests are equivalent to a t-test for equality of means and compare the mean for an initial share of the sample (e.g. 10%, 15% ...) with the mean obtained from the remaining 50% of the sample. According to this test, failure to reject the null hypothesis is interpreted as indicating convergence of the chain. This test is performed for each

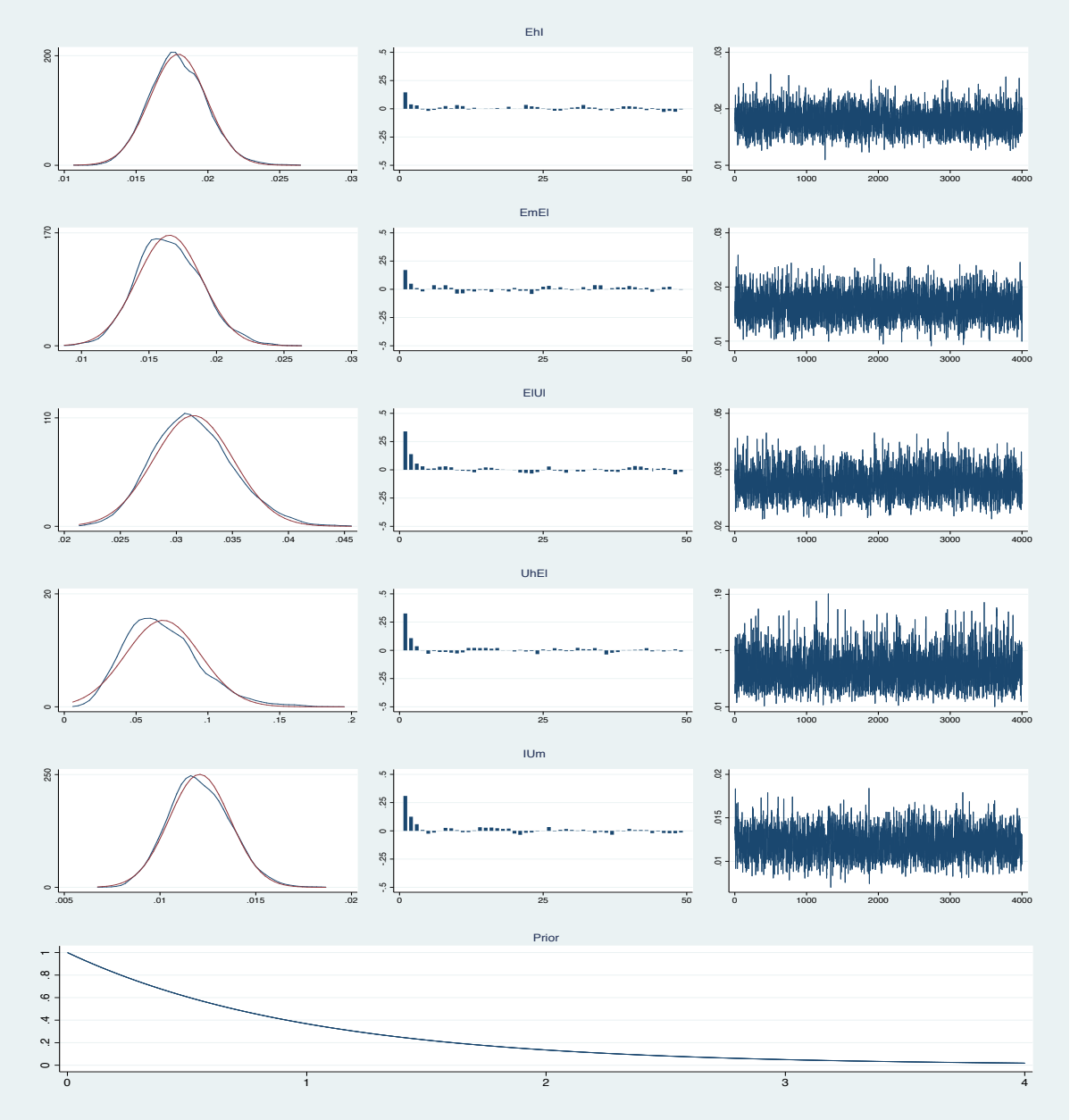
hazard and each period.[33] Over all the test performed, the null is rejected in 5.8% of the tests at a 5% significance level. This rate is slightly higher than the retained significance level but the difference is only marginal (less than 1pp) and could be coming from the presence of autocorrelation and/or the relatively low number of replications G. Thus, these results as well as the graphical evidence in Figures 10 and 11 seem to indicate that estimation results are satisfactory.



Diagnostic checks for hazard rates estimates for some selected series. The left column displays a plot of the posterior density in blue against the normal distribution in red. The middle column shows the autocorrelation function for the simulated series of hazard rates. These series are displayed in the right column.

Figure 10: Diagnostic checks: August 1991

[33] $419 \times 30 = 12570$ tests are performed. I use the *LeSage* toolbox (`https://www.spatial-econometrics.com/`) to obtain the test results.



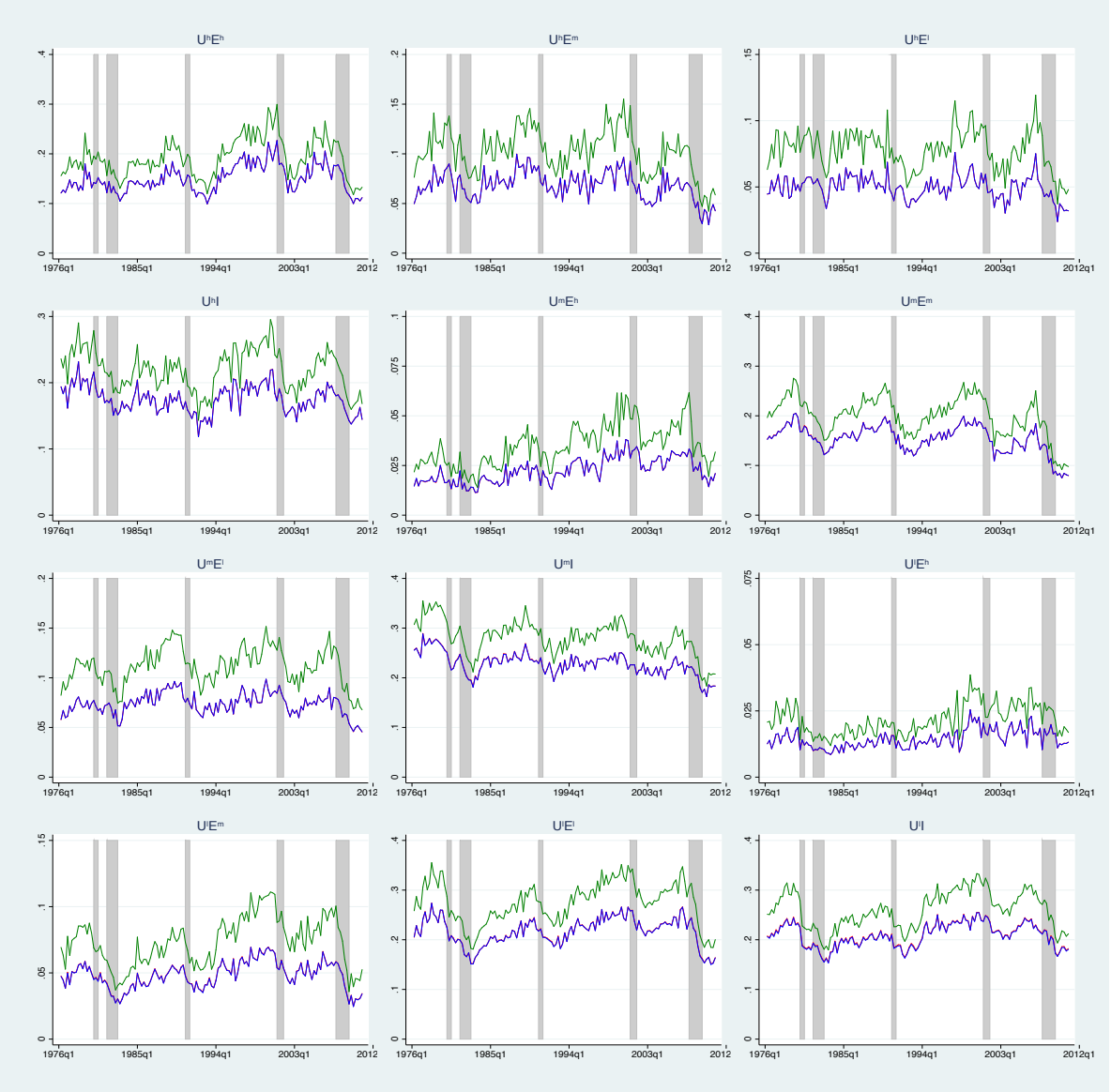
Diagnostic checks for hazard rates estimates for some selected series. The left column displays a plot of the posterior density in blue against the normal distribution in red. The middle column shows the autocorrelation function for the simulated series of hazard rates. These series are displayed in the right column. The bottom graph plots the prior distribution.

Figure 11: Diagnostic checks: March 2004

4.4 Corrected Series: Margin Of Adjustment and Time Aggregation.

To assess the impact of the Margin of Adjustment and the Time Aggregation corrections, Table 8 compares flow rates at various stages of the correction procedure. This table provides average flow rates for the original series, the series corrected for the 1994 redesign, the population and classification changes (Section 3) and the series corrected for the Margin of Adjustment. The last column displays the average transition probabilities corrected for Time Aggregation ($\hat{p}_t^{ij} = 1 - e^{-f_t^{ij}}$). A plot of a subset of transition rates from unemployment can be found in Figure 12 and in Figures 20 and 21 for flow rates from inactivity and employment in Appendix A.3.3.

In agreement with what is reported by Elsby et al. (2015), the Margin of Adjustment correction only leads to minor changes to the series obtained after the corrections of Section 3. The largest adjustments (for $\tilde{p}^{E^lI}$ and $\tilde{p}^{IU^l}$) never exceed 0.2 percentage point on average ($\tilde{p}^{II}$ is obtained as a residual). Moreover, flows rates from unemployment remain almost unchanged (see Figure 12) and the adjustment affects mostly flow rates from employment and inactivity. Flow rates from employment to inactivity decrease for all occupations whereas flow rates to unemployment tend to increase (except for $\tilde{p}^{E^hU^h}$). All flows from inactivity increase, with larger adjustments flow rates series to low skill employment and unemployment. Table 9 allows to compare adjustments in terms of aggregate flow rates to those reported by Elsby et al. (2015) for CPS data. This table shows that the sign of the correction for each flow rate are the same as the one they obtained.



Quarterly averages of monthly flow rates. The corrected flow rates from Section 3 are displayed in red, series corrected for Margin of Adjustment are displayed in blue and the series corrected Time Aggregation are plotted in green.

Figure 12: Flow rates corrected for Margin of Adjustment and Time Aggregation: Unemployment

	Full samp.					Full samp.			
	Orig.	Corr.	Margin	TA		Orig.	Corr.	Margin	TA
High Skill									
$EhEh$	90.70	95.92	96.17	95.84	$UhEh$	15.07	14.98	14.99	19.18
	(5.19)	(0.36)	(0.39)	(0.43)		(3.91)	(3.19)	(3.19)	(4.33)
$EhEm$	4.50	0.99	0.92	0.96	$UhEm$	6.87	6.83	6.82	10.23
	(3.44)	(0.20)	(0.20)	(0.21)		(2.07)	(1.84)	(1.83)	(2.80)
$EhEl$	2.43	0.65	0.63	0.68	$UhEl$	5.31	4.96	4.96	7.79
	(1.76)	(0.12)	(0.13)	(0.18)		(1.66)	(1.38)	(1.38)	(2.13)
$EhUh$	0.52	0.58	0.57	0.84	$UhUh$	49.44	54.55	54.57	40.86
	(0.14)	(0.10)	(0.10)	(0.15)		(7.25)	(5.97)	(5.96)	(9.86)
EhI	1.73	1.82	1.67	1.68	UhI	16.90	17.50	17.48	21.95
	(0.32)	(0.17)	(0.18)	(0.18)		(2.89)	(2.64)	(2.64)	(3.70)
Middle Skill									
$EmEh$	4.29	0.99	1.05	1.08	$UmEh$	2.36	2.24	2.25	3.49
	(2.91)	(0.26)	(0.26)	(0.31)		(0.88)	(0.77)	(0.77)	(1.28)
$EmEm$	87.74	93.08	92.97	92.32	$UmEm$	15.54	15.35	15.33	20.03
	(5.10)	(0.61)	(0.60)	(0.69)		(3.43)	(3.05)	(3.05)	(4.32)
$EmEl$	3.60	1.31	1.35	1.49	$UmEl$	7.42	7.36	7.37	11.11
	(2.20)	(0.21)	(0.21)	(0.24)		(1.67)	(1.39)	(1.39)	(2.27)
$EmUm$	1.23	1.37	1.42	2.14	$UmUm$	48.24	50.92	50.97	37.29
	(0.26)	(0.22)	(0.24)	(0.31)		(6.17)	(5.96)	(5.95)	(9.56)
EmI	2.95	3.16	3.13	2.97	UmI	21.48	22.84	22.79	28.08
	(0.55)	(0.29)	(0.30)	(0.30)		(2.95)	(2.63)	(2.62)	(3.80)
Low Skill									
$ElEh$	2.45	0.79	0.81	0.82	$UlEh$	1.44	1.44	1.44	2.16
	(1.62)	(0.19)	(0.20)	(0.23)		(0.54)	(0.49)	(0.49)	(0.77)
$ElEm$	3.81	1.45	1.41	1.42	$UlEm$	4.95	4.83	4.81	7.33
	(2.42)	(0.18)	(0.19)	(0.23)		(1.37)	(1.17)	(1.16)	(2.03)
$ElEl$	87.15	91.04	91.09	89.94	$UlEl$	22.15	21.73	21.68	27.50
	(4.68)	(0.91)	(0.91)	(1.06)		(4.10)	(3.21)	(3.21)	(4.67)
$ElUl$	2.11	2.28	2.44	3.73	$UlUl$	47.51	50.21	50.42	36.99
	(0.44)	(0.44)	(0.45)	(0.57)		(6.18)	(6.36)	(6.38)	(10.26)
ElI	4.26	4.34	4.16	4.09	UlI	20.51	20.99	20.85	26.02
	(1.03)	(0.51)	(0.48)	(0.53)		(3.08)	(2.58)	(2.59)	(3.94)
Inactivity					**Aggregate**				
IEh	0.96	0.97	1.09	1.10	EE	95.62	95.49	95.55	94.96
	(0.23)	(0.18)	(0.20)	(0.17)		(0.70)	(0.49)	(0.49)	(0.55)
IEm	1.50	1.52	1.54	1.43	EU	1.44	1.45	1.52	2.19
	(0.25)	(0.17)	(0.19)	(0.18)		(0.29)	(0.25)	(0.26)	(0.33)
IEl	2.18	2.23	2.38	2.23	EI	2.94	3.05	2.93	2.86
	(0.48)	(0.25)	(0.31)	(0.30)		(0.59)	(0.30)	(0.29)	(0.28)
IUh	0.30	0.34	0.37	0.56	UE	27.16	26.75	26.71	36.16
	(0.08)	(0.09)	(0.09)	(0.12)		(4.80)	(3.94)	(3.94)	(6.26)
IUm	0.79	0.89	0.96	1.43	UU	52.61	52.21	52.34	37.79
	(0.14)	(0.19)	(0.23)	(0.36)		(6.23)	(5.67)	(5.69)	(9.23)
IUl	0.95	1.03	1.22	1.72	UI	20.23	21.04	20.95	26.04
	(0.16)	(0.18)	(0.20)	(0.22)		(2.56)	(2.12)	(2.13)	(3.29)
II	93.31	93.03	92.44	91.54	IE	4.64	4.72	5.01	4.76
	(0.88)	(0.44)	(0.53)	(0.65)		(0.80)	(0.37)	(0.43)	(0.40)
					IU	2.04	2.26	2.55	3.70
						(0.31)	(0.35)	(0.40)	(0.48)

Average monthly flow rates over the period February 1976 - December 2010. The columns labelled "Orig." displays average flow rates for uncorrected series, the column "Corr." gives results for series corrected for the 1994 redesign, population and classification changes. The columns "Margin" displays averages of series corrected for the Margin of Adjustment and the last column gives the transition probabilities corrected for Time Aggregation ($1 - e^{-f_t^{ij}}$). Standard deviations are given in parenthesis.

Table 8: Average Flow rates Corrected for Margin of Adjustment and Time Aggregation

The correction for Time Aggregation leads to adjustments similar to those reported by Shimer (2012) and Elsby et al. (2015). The EI and IE transitions probabilities decrease while those involving unemployment (EU, UE, UI and IU) increase. In other words, transitions between employment and inactivity often miss a transition to unemployment ($EI \Rightarrow EUI$ and $IE \Rightarrow IUE$).

Table 9 shows that the corrections for aggregate transition probabilities are rather consistent with the ones obtained by Elsby et al. (2015) despite differences in samples and corrections applied to series. $\hat{p}^{UE}$ and $\hat{p}^{UI}$ increase by 35% and 24% (9.45 pp and 5.09 pp), whereas $\hat{p}^{EU}$ and $\hat{p}^{IU}$ increase by 44% and 45% (0.67pp and 1.15pp). On the other hand, Elsby et al. (2015) report increases of respectively, 37%, 40%, 35% and 39% (9.6pp, 8.36pp, .53pp and 1.08pp).[34] These substantial adjustments in transition probabilities from and to unemployment occur with relatively small changes in $\hat{p}^{EI}$ and $\hat{p}^{IE}$ (0.07pp and 0.15pp on average). This suggests that an important share of missed transitions are also related to EE and II flows with an intermediate transition to unemployment (EUE and IUI). On average, $\hat{p}^{EE}$ and $\hat{p}^{II}$ decrease by 0.59pp and 0.9pp (.47pp and .83pp for Elsby et al. (2015)).

	This paper 1976-2010				EHS 1978-2012			
	Orig.	Corr.	Margin	TA	Orig.	Corr.	Margin	TA
Employment								
EE	95.62	95.49	95.55	94.96	95.61	95.64	95.69	95.22
	(0.70)	(0.49)	(0.49)	(0.55)	(0.68)	(0.43)	(0.42)	(0.48)
EU	1.44	1.45	1.52	2.19	1.48	1.47	1.53	2.06
	(0.29)	(0.25)	(0.26)	(0.33)	(0.30)	(0.26)	(0.26)	(0.31)
EI	2.94	3.05	2.93	2.86	2.91	2.89	2.79	2.72
	(0.59)	(0.30)	(0.29)	(0.28)	(0.54)	(0.28)	(0.26)	(0.27)
Unemployment								
UE	27.16	26.75	26.71	36.16	26.15	26.13	25.80	35.40
	(4.80)	(3.94)	(3.94)	(6.26)	(4.68)	(3.92)	(3.84)	(6.73)
UU	52.61	52.21	52.34	37.79	51.83	51.84	53.21	35.26
	(6.23)	(5.67)	(5.69)	(9.23)	(6.11)	(5.62)	(5.43)	(10.32)
UI	20.23	21.04	20.95	26.04	22.02	22.03	20.98	29.34
	(2.56)	(2.12)	(2.13)	(3.29)	(2.50)	(2.35)	(2.22)	(4.28)
Inactivity								
IE	4.64	4.72	5.01	4.76	4.60	4.59	4.73	4.48
	(0.80)	(0.37)	(0.43)	(0.40)	(0.75)	(0.39)	(0.41)	(0.40)
IU	2.04	2.26	2.55	3.70	2.56	2.57	2.74	3.82
	(0.31)	(0.35)	(0.40)	(0.48)	(0.44)	(0.35)	(0.35)	(0.37)
II	93.31	93.03	92.44	91.54	92.84	92.84	92.53	91.70
	(0.88)	(0.44)	(0.53)	(0.65)	(0.95)	(0.35)	(0.36)	(0.41)

Average monthly flow rates over the period February 1976 - December 2010. The columns labelled "Orig." display average flow rates for uncorrected series, the columns "Corr." give results for series corrected for the 1994 redesign, population and classification changes. For Elsby et al. (2015)'s results (EHS), this column shows the average of seasonally adjusted series. The columns "Margin" display averages of series corrected for the Margin of Adjustment and the last column gives the transition probabilities corrected for Time Aggregation ($1 - e^{-f_t^{ij}}$). Standard deviations are given in parenthesis.

Table 9: Average Flow rates Corrected for Margin of Adjustment and Time Aggregation.

[34] In their paper, Elsby et al. (2015) mention increases in unemployment inflows (EU and IU) and outflows (UE and UI) of 30% and 15%. These results are smaller than the ones displayed in Table 9 because they apply to series corrected using the Abowd and Zellner (1985) correction. This correction leads to a decrease in EI and IE gross flows.

The Time Aggregation correction for disaggregated transition rates is actually relatively similar across occupations. For instance, $\hat{p}^{E^hU^h}$, $\hat{p}^{E^mU^m}$ and $\hat{p}^{E^lU^l}$ increase by 47%, 51% and 53% respectively (0.27pp, 0.72pp and 1.29pp) whereas increases in transition rates from inactivity to unemployment range from 41% for low skill occupations to 51% for high skill ones. Therefore, it seems that the Time Aggregation bias affects transition rates by occupations in a fairly identical way (in relative terms). Finally, as Elsby et al. (2015) note, the graphical evidence in Figure 12 suggests that the Time Aggregation correction preserves the cyclical behavior of the transition rates. Furthermore, Shimer (2012) argues that missed transitions are more likely to occur in booms than recessions. The reason being that workers transition much faster through unemployment in good times ($\hat{p}^{UU}$ is small). This aspect can be inferred, for instance, from the U^mE^l and U^lE^m transition probabilities in Figure 12 as the difference between the 2 series appears to decrease in recessions and increase in booms (particularly before the 2001 recession).

Conclusion

Using CPS data over the period 1976-2010 and the occupation classification of Autor and Dorn (2013) to rank occupations between high, medium and low skills, I propose a framework for adjusting various problems and breaks affecting these series.

In a first step, I use an Unobserved Component model to deseasonalize time series of stocks and gross flows and adjust these series for breaks due to the 1994 redesign of the CPS questionnaire, changes in occupational classification and revisions in the size and composition of the US population.

The 1994 redesign is shown to have significant effects on most series of stocks and flows. In particular, it leads to a drop in the number of *New Unemployed Entrants* for which an occupation code is not available. As a result, the pool of unemployed that can be classified between high, middle and low skill occupations significantly increases after 1994. This turns out to have substantial effects on series of unemployment stocks and gross flows. Moreover, the 1994 redesign saw the introduction of *dependent interviewing technique* which reduced the number of spurious transitions between occupations. Thus gross flow series from employment need to be adjusted. The estimation results also confirm that the 1994 redesign has a significant effect on gross flow series between inactivity and (un)-employment (Abraham and Shimer (2001)).

From the 3 classification changes happening throughout the 1976-2010 period, adjustments for the 1976-1982 and the 2003-2010 classification changes are required. The correction for the 1976-182 change leads to a reallocation of employment from middle to low skill series and the correction for the 2003-2010 change reallocate employment from middle/low skill occupations to high skill occupations. Lastly, I adjust the population updates of 1990 and 2003 which, according to the BLS results, are the most important ones that happened over the 1976-2010 period. Moreover, the series retrieved from micro data do not correct for population updates applied prior to 1980. It is therefore required to adjust series for an additional population change over the 1976-1979 period.

In a second step, I adjust series for the Margin of Adjustment problem and the Time Aggregation bias. I follow Elsby et al. (2015) to correct the Margin of Adjustment problem which only leads to a minor adjustments in flow rate series. Due to some issues and specific constraints related to flow rate series by occupations, I cannot apply the Time Aggregation correction used by Shimer (2012) and Elsby et al. (2015). As a result, I use the bayesian estimation method proposed by Bladt and Sørensen (2005) to correct for this bias. I further show that this method implies adjustments similar to those reported by Shimer (2012) and Elsby et al. (2015) with increases in transition rates from and to unemployment and decreases in flow rates between employment and inactivity. The adjustments obtained for disaggregated flow rate series are relatively similar across occupations.

The adjustments presented in this paper focused particularly on the occupational dimension. From the work of Abowd and Zellner (1985) and Poterba and Summers (1986), it is known that the CPS also suffers from misclassification errors between unemployment and inactivity. The adjustment

proposed by Abowd and Zellner (1985) leads to important decreases in flows from and to inactivity. Since the 1994 redesign has significant effects on these flows too, it would be interesting to apply an adjustment for these misclassification errors before running the framework presented in this paper. Furthermore, spurious transitions between occupations could still be problematic after the redesign of 1994. Moscarini and Thomsson (2007) propose a treatment of the data that could be worth applying to series before adjusting them. The Time Aggregation correction could also be extended to account for the dynamic mapping between the recorded labor market status and the labor market activities of workers. Keeping in mind these limitations, the adjustments applied in this paper allow nonetheless, to obtain time series of stocks and gross flows that are more consistent throughout the 1976-2010 period, especially when compared with the large breaks affecting original series.

References

John M. Abowd and Arnold Zellner. Estimating gross labor-force flows. *Journal of Business & Economic Statistics*, 3(3):254–283, 1985.

Katharine G. Abraham and Robert Shimer. Changes in unemployment duration and labor force attachment. Working Paper 8513, National Bureau of Economic Research, October 2001.

David H. Autor. The "task approach" to labor markets: an overview. Working Paper 18711, National Bureau of Economic Research, January 2013.

David H. Autor and David Dorn. The growth of low-skill service jobs and the polarization of the US labor market. *American Economic Review*, 103(5):1553–97, August 2013.

Mogens Bladt and Michael Sørensen. Statistical inference for discretely observed markov jump processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(3):395–410, 2005.

Olivier Blanchard and Peter Diamond. The cyclical behavior of the gross flows of US workers. *Brookings Papers on Economic Activity*, 21(2):85–156, 1990.

Chung Chen and Lon-Mu Liu. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297, 1993.

Guido Matias Cortes, Nir Jaimovich, Christopher J. Nekarda, and Henry E. Siu. The micro and macro of disappearing routine jobs: A flows approach. Technical report, 2016.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

Michael R. Darby, John C. Haltiwanger, and Mark W. Plant. The ins and outs of unemployment: The ins win. Working Paper 1997, National Bureau of Economic Research, August 1986.

Piet De Jong. The diffuse kalman filter. *The Annals of Statistics*, 19(2):1073–1083, 1991.

Piet De Jong and Jeremy Penzer. Diagnosing shocks in time series. *Journal of the American Statistical Association*, 93(442):796–806, 1998.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2nd edition, 2012.

Michael W.L. Elsby, Bart Hobijn, and Ayşegül Şahin. On the importance of the participation margin for labor market fluctuations. *Journal of Monetary Economics*, 72:64–82, 2015.

Bruce C. Fallick and Charles A. Fleischman. Employer-to-employer flows in the US labor market: the complete picture of gross worker flows. Finance and Economics Discussion Series 2004-34, Board of Governors of the Federal Reserve System (US), 2004.

Martin S. Feldstein. The importance of temporary layoffs: an empirical analysis. *Brookings Papers on Economic Activity*, 1975(3):725–745, 1975.

Christopher L. Foote and Richard W. Ryan. Labor market polarization over the business cycle. Working Paper 21030, National Bureau of Economic Research, March 2015.

Anthony J. Fox. Outliers in time series. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 350–363, 1972.

Shigeru Fujita and Garey Ramey. The cyclicality of separation and job finding rates. *International Economic Review*, 50(2):415–430, 2009.

John Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statisitics 4*, pages 169–193. Oxford University Press, 1992.

James D. Hamilton. *Time series analysis*, volume 2. Princeton University Press, Princeton, NJ, 1994.

Andrew C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

Yasunari Inamura. Estimating continuous time transition matrices from discretely observed data. Bank Of Japan Working Papers Series 06 - E07-, Bank Of Japan, April 2006.

Robert B. Israel, Jeffrey S. Rosenthal, and Jason Z. Wei. Finding generators for markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, 11(2): 245–265, 2001.

Nir Jaimovich and Henry E. Siu. The trend is the cycle: Job polarization and jobless recoveries. Working Paper 18334, National Bureau of Economic Research, August 2012.

Hyman B. Kaitz. Analyzing the length of spells of unemployment. *Monthly Labor Review*, pages 11–20, 1970.

Gueorgui Kambourov and Iourii Manovskii. Rising occupational and industry mobility in the united states: 1968-97. *International Economic Review*, 49(1):41–79, 2008.

Gueorgui Kambourov and Iourii Manovskii. Occupational mobility and wage inequality. *The Review of Economic Studies*, 76(2):731–759, 2009.

Gueorgui Kambourov and Iourii Manovskii. A cautionary note on using (march) current population survey and panel study of income dynamics data to study worker mobility. *Macroeconomic Dynamics*, 17(01):172–194, 2013.

Brigitte C. Madrian and Lars John Lefgren. A note on longitudinally matching Current Population Survey (CPS) respondents. Working Paper 247, National Bureau of Economic Research, November 1999.

Giuseppe Moscarini and Kaj Thomsson. Occupational and job mobility in the us*. *The Scandinavian Journal of Economics*, 109(4):807–836, 2007.

James R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

Jeremy Penzer. Diagnosing seasonal shifts in time series using state space models. *Statistical Methodology*, 3(3):193–210, 2006.

George L. Perry, Robert E. Hall, Charles Holt, and Hyman B. Kaitz. Unemployment flows in the us labor market. *Brookings Papers on Economic Activity*, 1972(2):245–292, 1972.

Anne E. Polivka and Stephen M. Miller. The CPS after the redesign: Refocusing the economic lens. In *Labor Statistics Measurement Issues*, pages 249–289. University of Chicago Press, January 1998.

Anne E. Polivka and Jennifer M. Rothgeb. Redesigning the CPS questionnaire. *Monthly Labor Review*, 116(9):10–28, 1993.

James M. Poterba and Lawrence H. Summers. Reporting errors and labor market dynamics. *Econometrica: Journal of the Econometric Society*, pages 1319–1338, 1986.

Paul Ryscavage. *Measuring spells of unemployment and their outcomes*. US Census Bureau, 1989.

Robert Shimer. Reassessing the ins and outs of unemployment. *Review of Economic Dynamics*, 15(2): 127–148, 2012.

Ruey S. Tsay. Outliers, level shifts, and variance changes in time series. *Journal of forecasting*, 7(1): 1–20, 1988.

A Appendix

A.1 Data

A.1.1 Complementary descriptive statistics

- average flow rates from employment and inactivity.

	FS.	Age			Marit. st		Race				Educ.		
		<25	**25-50**	**>50**	**M.**	**NM.**	**W.**	**H.**	**B.**	**O.**	**<HS**	**HS**	**>HS**
High Skill													
$p^{E^h E^h}$	96.16	91.02	96.74	96.23	96.81	94.86	96.53	94.28	94.20	95.70	89.47	94.44	96.62
$p^{E^h E^m}$	0.99	2.01	0.92	0.83	0.82	1.34	0.89	1.53	1.59	1.06	1.68	1.62	0.86
$p^{E^h E^l}$	0.66	1.53	0.61	0.50	0.52	0.93	0.56	1.31	1.20	0.61	2.67	1.37	0.49
$p^{E^h U^h}$	0.59	1.27	0.54	0.49	0.45	0.86	0.54	0.78	0.89	0.61	1.21	0.67	0.56
$p^{E^h U^m}$	0.02	0.06	0.02	0.02	0.02	0.03	0.02	0.05	0.05	0.02	0.08	0.05	0.02
$p^{E^h U^l}$	0.03	0.06	0.02	0.02	0.02	0.04	0.02	0.06	0.05	0.02	0.17	0.06	0.02
$p^{E^h I}$	1.56	4.06	1.14	1.91	1.36	1.93	1.45	1.98	2.03	1.99	4.72	1.79	1.44
Middle Skill													
$p^{E^m E^h}$	1.42	1.31	1.51	1.24	1.40	1.44	1.43	1.10	1.45	1.73	0.46	0.90	2.06
$p^{E^m E^m}$	92.92	87.72	94.24	94.47	94.45	91.13	93.59	91.50	91.05	91.75	88.70	93.80	93.22
$p^{E^m E^l}$	1.39	2.38	1.22	0.90	1.04	1.81	1.18	2.13	1.86	1.42	2.68	1.52	1.00
$p^{E^m U^h}$	0.03	0.02	0.02	0.03	0.02	0.03	0.03	0.02	0.03	0.03	0.01	0.02	0.03
$p^{E^m U^m}$	1.27	2.28	1.08	0.79	0.85	1.77	1.11	1.59	1.92	1.26	2.19	1.26	1.06
$p^{E^m U^l}$	0.07	0.11	0.06	0.04	0.05	0.10	0.05	0.12	0.10	0.07	0.18	0.08	0.04
$p^{E^m I}$	2.90	6.17	1.87	2.52	2.19	3.73	2.61	3.53	3.59	3.75	5.78	2.42	2.59
Low Skill													
$p^{E^l E^h}$	0.86	0.81	0.91	0.77	0.88	0.85	0.92	0.53	0.93	1.03	0.32	0.62	1.55
$p^{E^l E^m}$	1.36	2.17	1.16	0.93	1.06	1.67	1.26	1.45	1.62	1.58	1.35	1.34	1.39
$p^{E^l E^l}$	91.97	86.52	93.66	93.52	94.06	89.80	92.48	91.51	90.35	91.07	89.31	93.15	92.33
$p^{E^l U^h}$	0.02	0.03	0.03	0.02	0.02	0.03	0.03	0.02	0.03	0.02	0.02	0.02	0.04
$p^{E^l U^m}$	0.06	0.10	0.06	0.05	0.05	0.08	0.05	0.08	0.10	0.08	0.09	0.07	0.05
$p^{E^l U^l}$	2.06	3.10	1.90	1.28	1.48	2.68	1.82	2.59	2.63	2.00	2.98	1.99	1.53
p^{El^I}	3.65	7.28	2.28	3.44	2.45	4.89	3.44	3.82	4.33	4.23	5.95	2.81	3.12
Inactivity													
p^{IE^h}	1.57	1.25	2.20	1.16	1.82	1.33	1.76	0.89	1.15	1.95	0.32	0.83	3.26
p^{IE^m}	2.24	3.59	2.26	1.02	1.84	2.62	2.20	2.47	2.17	2.32	1.77	2.28	2.63
p^{IE^l}	3.11	5.12	3.02	1.43	2.11	4.05	2.93	4.18	3.10	2.69	3.86	3.13	2.42
p^{IU^h}	0.53	0.33	0.81	0.40	0.56	0.49	0.56	0.32	0.51	0.63	0.11	0.31	1.06
p^{IU^m}	1.20	1.73	1.44	0.46	0.84	1.53	1.00	1.38	1.94	1.16	0.98	1.42	1.21
p^{IU^l}	1.53	2.37	1.79	0.50	0.86	2.16	1.25	2.00	2.44	1.33	1.92	1.77	0.99
p^{II}	89.83	85.62	88.49	95.03	91.96	87.80	90.30	88.76	88.69	89.92	91.04	90.25	88.43

Flow rates from employment $p^{E \cdot X}$ and inactivity $p^{I \cdot X}$ expressed in percentage, computed from the raw matched CPS files and averaged over the period Feb. 1994 to Dec. 2010. All observations are weighted using weights provided by the CPS. FS stands for full sample, M for Married and NM for not married. The races W., H., B. and O. stands for white, hispanic, black and others.

Table 10: Average Flow Rates from Employment and Inactivity (1)

A.2 Unobserved Component Model

A.2.1 State Space Matrices

This section presents the matrices for the state space representation (10)-(11). Each series is corrected individually implying that y_t is a scalar. Assuming that all components of model (2-9) are introduced into the model specification, the length of the state vector is $m = k+k_\chi+15+nx$ where $k = \max\{p, q+1\}$, $k_\chi = \max\{p_\chi, q_\chi+1\}$ are related to the number of parameters in the $ARMA$ specifications for the irregular component (6) and the 1994 redesign irregular component (9). I follow Hamilton (1994) to write the $ARMA$ in state space form. The number 15 corresponds to the mean and trend components (3), (4) and the mean and trend for the 1994 component in (8) as well as the 11 seasonal components (5). The number of exogenous variables (the classification and population effects) is nx. The state vector is ordered in the following way:

$$\alpha_t = \begin{bmatrix} \varepsilon_t & \dots & \varepsilon_{t-p+1} & \varepsilon_{\chi,t} & \dots & \varepsilon_{\chi,t-p_\chi+1} & \mu_t & \nu_t & \mu_{\chi,t} & \nu_\chi & \gamma_t & \dots & \gamma_{t-10} & \beta \end{bmatrix}'$$

where β is the $nx \times 1$ vector of coefficients for the exogenous variables capturing the classification/population changes.
The vector Z_t in the measurement equation (10) is the only time dependent vector in the state space representation (10)-(11). For t smaller than January 1994, we have

$$Z_t = \begin{bmatrix} 1 & \theta_1 & \dots & \theta_q & 0_{p-q-1} & 1 & \theta_{\chi,1} & \dots & \theta_{\chi,q_\chi} & 0_{p_\chi-q_\chi-1} & 1 & 0 & 1 & 0 & 1 & 0_{10} & x_t \end{bmatrix}$$

x_t is the tth row of the matrix of exogenous variables X_t (a vector of 1 and 0 depending on which classification and population changes affect the series).
For t equal or greater than January 1994, we have to remove the 1994 component and we obtain:

$$Z_t = \begin{bmatrix} 1 & \theta_1 & \dots & \theta_q & 0_{p-q-1} & 0_{q_\chi+1} & 0_{p_\chi-q_\chi-1} & 1 & 0 & 0 & 0 & 1 & 0_{10} & X_t \end{bmatrix}$$

The matrix T, from the state equation (11), can be written as:

$$T = \begin{bmatrix} T_\varepsilon & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & T_{\varepsilon_\chi} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & T_\mu & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & T_{\mu_\chi} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & T_\gamma & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} & T_\beta \end{bmatrix}$$

where the respective block matrices are given by:

$$T_\varepsilon = \begin{bmatrix} \rho_1 & \dots & 0 & \rho_p & 0_{q+1-p} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix};\ T_{\varepsilon_\chi} = \begin{bmatrix} \rho_{\chi,1} & \dots & 0 & \rho_{\chi,p_\chi} & 0_{q_\chi+1-p_\chi} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix};$$

$$T_\mu = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix};\ T_{\mu_\chi} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix};\ T_\gamma = \begin{bmatrix} -1 & \dots & -1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & \dots & 1 & 0 \end{bmatrix};\ T_\beta = I_{nx}$$

The matrix of variances Q is the following:

$$Q = \begin{bmatrix} \sigma^2_{\varepsilon} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{\varepsilon_\chi} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{\mu} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{\mu_\chi} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{\nu} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{\gamma} \end{bmatrix}$$

and the $m \times 6$ selection matrix R with columns of 0 and 1 assigns the disturbances to the respective state components. See Durbin and Koopman (2012) for more details.

A.2.2 Kalman Filter and Smoother: Recursions

The Kalman Filter is a set of recursions for sequentially computing the conditional expectation of the state vector $a_{t+1} = E[\alpha_{t+1}|Y_t]$ and its variance $P_{t+1} = Var[\alpha_{t+1}|Y_t]$ for $t = 1, ..., T$ with $Y_t = (y_1, y_2, ..., y_t)'$. Assuming that the initial state vector α_1 is gaussian with known mean a_1 and variance P_1, the recursions starts with a forecast of the observed variable y_t using (10). In a second step, the state vector is updated using the realization of y_t. The updated state vector can then be used to forecast the next period state vector using (11) These steps are usually performed in one recursion through the following set of equations:

$$\begin{aligned} \hat{y}_t &= Z_t a_t \\ v_t &= y_t - \hat{y}_t \\ F_t &= Z_t P_t Z_t' + H_t \\ K_t &= T_t P_t Z_t' F_t^{-1} \\ L_t &= T_t - K_t Z_t \\ a_{t+1} &= T a_t + K_t v_t \\ P_{t+1} &= T_t P_t L_t + R_t Q_t R_t' \end{aligned} \tag{20}$$

where $\hat{y}_t$ is the one step ahead forecast for y_t given a_t, v_t is the one step ahead forecast error (or innovations) with variance F_t and K_t is the Kalman gain. These recursions can be derived using a result for bivariate normal distributions.[35] See Durbin and Koopman (2012) chapter 4 or Hamilton (1994) chapter 13 for a detailed derivation of the above recursions.

The quantities computed by the filter (the innovations v_t and their variance F_t) serve to evaluate the likelihood function of the model and therefore estimate parameters. One other advantage of the Kalman Filter is the ease with which it allows to deal with missing observations. As pointed by Durbin and Koopman (2012), when an observation for y_t is missing, one can set $v_t = 0$ and $K_t = 0$. This implies that the updating step is not performed since the information provided by y_t on the current state vector is not available.

Together, with these forward (filtering) recursions, backward (smoothing) recursions can also be computed. These recursions allow to obtain the conditional mean $\hat{a}_t = E[\alpha_t|Y_T]$ and conditional variance $\hat{P}_t = Var[\alpha_t|Y_T]$ where $Y_T = (y_1, y_2, ...y_T)'$ is the vector of all observations. The smoothing

[35] This result states that if x and y are jointly normally distributed with

$$E[(x\ y)'] = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and } Var[(x\ y)'] = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix},$$

then the conditional distribution of x given y is also normal with mean $E[x|y] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ and $Var[x|y] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}'$. See Durbin and Koopman (2012).

recursions provide an estimate of the state vector and its variance conditional on all the information available in the sample. As for the filter equations, the derivations of the smoothing recursions are based on the results for bivariate normal distribution mentioned in footnote 35 and a detail derivation can be found in chapter 4 of Durbin and Koopman (2012). Starting from the last period in the sample ($t = T$), the Kalman smoothing recursions are given by:

$$\begin{aligned} r_{t-1} &= Z_t' F_t^{-1} v_t + L_t' r_t \\ N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t \\ \hat{a}_t &= a_t + P_t r_{t-1} \\ \hat{P}_t &= P_t - P_t N_{t-1} P_t \\ u_t &= F_t^{-1} v_t - K_t' r_t \end{aligned} \tag{21}$$

which are initialized with $r_T = 0$ and $N_T = 0$. The last recursion for u_t gives the smoothed residuals for the measurement equation (10). This recursion is not required to obtain $\hat{a}_t$ and $\hat{P}_t$ but it is used in the outlier detection procedure.

As pointed by Hamilton (1994), these smoothing estimates are of particular interest when the state vector is given a structural interpretation. The quantity obtained through these recursions will be used to replace missing values and to remove effects related to seasonality, classification/population changes and from the 1994 redesign.

A.2.3 Initialization and the Augmented Kalman Filter and Smoother

Provided that the unobserved state vector is stationary,[36] the forward recursions can be initialized using the unconditional mean and variance for α_t computed from (11). However, in UC model, some components like the mean, trend and seasonal ones (equations (3), (4) and (5)), are not stationary which implies that some elements in a_1 and P_1 cannot be set to their respective unconditional means and variances.

One solution is to resort to a *diffuse initialization* of the Kalman Filter.[37] The main idea is to separate the state vector between stationary and non stationary components as follows:

$$\alpha_1 = a + A\delta + R_0 q_0, \qquad q_0 \sim \mathcal{N}(0, Q_0)$$

where a is a $(m \times 1)$ vector of known constant, A is a $(m \times q)$ selection matrix with q ones in positions corresponding to the non stationary component in the state vector. Since m is the size of the state vector, we must have $q \leq m$. δ is a $(q \times 1)$ vector of unknown quantities, R_0 is also a $(m \times (m-q))$ selection matrix with $m - q$ ones in positions associated to the stationary components of the state vector.[38] Because a given component cannot be stationary and non stationary at the same time, we must have $A' R_0 = \mathbf{0}$. The vector of unknown quantities δ can be treated either as a vector of fixed parameters to be estimated (not considered here) or as a vector of random normal variables:

$$\delta \sim \mathcal{N}(0, \kappa I_q)$$

[36] That is, the eigenvalues of the matrix T lie within the unit circle. See Hamilton (1994).

[37] This quick presentation is based on chapter 5 in Durbin and Koopman (2012) and much more details regarding the topic of initialization can be found in this chapter.

[38] Assume, for instance, that the state vector is made of 4 elements ($m = 4$) with the first 2 being stationary and the last 2 non stationary ($q = 2$) then we have

$$R_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}' ; \; A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}'$$

where $\kappa \to \infty$ at a suitable moment. The term *diffuse* being related to this infinite variance assumption. The unconditional mean and variance are then given by:

$$\begin{aligned} a_1 &= E[\alpha_1] \\ &= a \\ P_1 &= Var[\alpha_1] \\ &= AA'\kappa + R_0 Q_0 R_0' \end{aligned}$$

These quantities can be used to initialize the filter by setting a large arbitrary value for κ and use the resulting values for a_1 and P_1 to start the filter. Durbin and Koopman (2012) points to the fact that this way to proceed leads to numerical inaccuracies when running the filter and it should therefore be avoided.

2 other solutions are available. A first one is called *The exact initial Kalman Filter* by Durbin and Koopman (2012) and is the one that they advocate for. While this procedure is the most efficient in terms of the number of computations (and hence speed) required to run the Kalman Filter, it is not suited to the specification presented in Section 3.[39] The second solution, due to De Jong (1991), is called *The augmented Kalman Filter* and is the one pursued in this work. It makes use of the linearity of the Kalman Filter and requires augmenting the observation and state vectors. More precisely, for a given δ, the unconditional mean and variance of α_1 are given by $a_{\delta,1} = a + A\delta$ and $P_1 = R_0 Q_0 R_0'$. The linearity of the filter implies that we can write the one step ahead forecast $\hat{y}_t$, the innovations v_t and the state vector forecast a_{t+1} in (20) as:

$$\begin{aligned} \hat{y}_{\delta,t} &= \hat{y}_{a,t} + \hat{Y}_{A,t}\delta \\ v_{\delta,t} &= v_{a,t} + V_{A,t}\delta \\ a_{\delta,t+1} &= a_{a,t} + A_{A,t}\delta \end{aligned}$$

In order to see this more explicitly, let us the derive the first recursion (t=1) of the Kalman Filter:

$$\begin{aligned} \hat{y}_{\delta,1} &= Z_1 a_{\delta,1} \\ &= \underbrace{Z_1 a}_{\hat{y}_{a,1}} + \underbrace{Z_1 A}_{\hat{Y}_{A,1}}\delta \\ v_{\delta,1} &= y_1 - \hat{y}_1 \\ &= \underbrace{y_1 - Z_1 a}_{v_{a,1}} + \underbrace{-Z_1 A}_{V_{A,1}}\delta \\ a_{\delta,2} &= T_1 a_{\delta,1} + K_1 v_{\delta,1} \\ &= \underbrace{T_1 a + K_1 v_{a,1}}_{a_{a,1}} + \underbrace{(T_1 A + K_1 V_{A,1})}_{A_{A,1}}\delta \end{aligned}$$

and the expression for F_1, K_1 and P_2 are not modified as they do not depend directly on a_1.
As pointed by Durbin and Koopman (2012) for a given t, the quantities $\hat{y}_{a,t}, \hat{Y}_{A,t}, v_{a,t}, V_{A,t}, a_{a,t}$ and $A_{A,t}$ can be computed in one recursion by augmenting the observation vector y_t with q zeros (the

[39] The problems comes from the exogenous variables in the matrix X_t and the mean component μ_t. To better understand this point, one would have to go into the details of *The exact initial Kalman Filter* but because of these exogenous variables there is an identification problem between these variables and the mean component. For instance, the mean and the 1994 redesign components cannot be disentangled before January 1994 when the 1994 component is no longer affecting series. The number of periods required to initialize the filter (d in Durbin and Koopman (2012)) is large which implies that a substantial number of observations have to be dropped.

number of diffuse elements):

$$\begin{aligned}
\left(\hat{y}_{a,t}, \hat{Y}_{A,t}\right) &= Z_t\left(a_{a,t}, A_{A,t}\right) \\
\left(v_{a,t}, V_{A,t}\right) &= \left(y_t, 0\right) - \left(\hat{y}_{a,t}, \hat{Y}_{A,t}\right) \\
F_t &= Z_t P_t Z_t' + H_t \\
K_t &= T_t P_t Z_t' F_t^{-1} \\
\left(a_{a,t+1}, A_{A,t+1}\right) &= T_t\left(a_{a,t}, A_{A,t}\right) + K_t\left(v_{a,t}, V_{A,t}\right) \\
P_{t+1} &= T_t P_t\left(T_t - K_t Z_t\right) + R_t Q_t R_t',
\end{aligned} \tag{22}$$

which constitutes the *The augmented Kalman Filter*.

Furthermore, Durbin and Koopman (2012) show that the vector δ can be estimated. Defining the conditional expectation $\bar{\delta} = E[\delta|Y_t]$, we have:

$$\begin{aligned}
b_T &= \sum_{i=1}^{T} V_{A,i}' F_i^{-1} v_{a,i} \\
S_{A,T} &= \sum_{i=1}^{T} V_{A,i}' F_i^{-1} V_{A,i} \\
\bar{\delta} &= -\left(S_{A,T} + \frac{1}{\kappa} I_q\right)^{-1} b_T \\
Var[\delta|Y_t] &= \left(S_{A,T} + \frac{1}{\kappa} I_q\right)^{-1},
\end{aligned}$$

which, on letting $\kappa \to \infty$, become:

$$\begin{aligned}
\bar{\delta} &= -S_{A,T}^{-1} b_T \\
Var[\delta|Y_t] &= S_{A,T}^{-1}.
\end{aligned}$$

In practice, *The augmented Kalman Filter* requires running the recursions (22) for a given δ (a vector of zeros for instance) and computing b_T and $S_{A,T}$. This allows then to obtain an estimate for δ that can be use to compute $\hat{y}_{\delta,t}, v_{\delta,t}$ and $a_{\delta,t+1}$.

Before turning to the outlier detection procedure, it is worth mentioning that the smoothing recursions also need to be adjusted in a similar way to the filter's ones by introducing the following quantities:

$$\begin{aligned}
r_{\delta,t} &= r_{a,t} + R_{A,t}\delta \\
\hat{a}_{\delta,t} &= \hat{a}_{a,t} + \hat{A}_{A,t}\delta \\
u_{\delta,t} &= u_{a,t} + U_{A,t}\delta
\end{aligned}$$

and then modifying the recursions in (21) accordingly. However, the most straightforward way to proceed is to run *The augmented Kalman Filter* twice to obtain $\bar{\delta}, \hat{y}_{\delta,t}, v_{\delta,t}$ and $a_{\delta,t+1}$. The Kalman Smoother can then be run using the recursions in (21):

$$\begin{aligned}
r_{\delta,t-1} &= Z_t' F_t^{-1} v_{\delta,t} + L_t' r_{\delta,t} \\
N_{t-1} &= Z_t' F_t^{-1} Z_t + L_t' N_t L_t \\
\hat{a}_{\delta,t} &= a_{\delta,t} + P_t r_{\delta,t-1} \\
V_t &= P_t - P_t N_{t-1} P_t \\
u_{\delta,t} &= F_t^{-1} v_{\delta,t} - K_t' r_{\delta,t}
\end{aligned} \tag{23}$$

which is started from $t = T$ with $r_{\delta,T} = 0$ and $N_T = 0$.

Finally, the quantities b_T and $S_{A,T}$ are used to evaluate the diffuse log-likelihood function. Durbin and Koopman (2012) show that this function is given by:

$$\log L_d = -\frac{T}{2}\log 2\pi - \frac{1}{2}\log|S_{A,T}| - \frac{1}{2}\sum_{t=1}^{T}\log|F_t| - \frac{1}{2}\left(S_{A,T} - b_T' S_{A,T}^{-1} b_T\right) \tag{24}$$

This function is used for the estimation of the model's parameters in Section 3.

A.2.4 Outliers

The effect of an outlier is usually thought in terms of the product of an impact effect and a dynamic response. The effect in period t of an outlier that occurred in period i is given by:

$$\psi_t(i) = \omega D_t(i).$$

The impact effect ω is estimated, whereas $D_t(i)$ captures how the outlier affects the series through time following the initial impact effect ω. Both *X12-ARIMA* and *TRAMO-SEATS* implement an outlier detection procedure inspired by the work of Chen and Liu (1993). Their proposed procedure is developed within the $ARIMA$ framework and it does not fit directly into the UC framework.

De Jong and Penzer (1998) demonstrate how outliers can be detected through a simple modification of the state space form (10)-(11) by introducing shocks to the measurement and state equations:

$$\begin{aligned} y_t &= Z_t\alpha_t + \tilde{X}_t\omega \\ \alpha_{t+1} &= T_t\alpha_t + W_t\omega + R_t\eta_t \end{aligned}$$

where ω, is the impact effect of the potential outliers, $\tilde{X}_t$ is a scalar (y_t is univariate) taking the value 1 for the period in which the outlier occurs and 0 otherwise. W_t is a $(m \times 1)$ vector which takes the value 1 for the component and the period in which the outlier happens. These simple shocks allow to generate variables D_t that correspond to usual dynamic responses found in the literature on outliers. De Jong and Penzer also show how the dynamic response at time t of an outlier happening in period i can be obtained through:

$$D_t(i) = \begin{cases} 0, & \text{if } t < i \\ \tilde{X}_t, & \text{if } t = i \\ Z_t T_{t-1,i+1} W_i, & \text{if } t > i \end{cases} \tag{25}$$

where $T_{j,t} \equiv T_j \dots T_t$ for $j \geq t$ and $T_{t-1,t} \equiv I$. It can be verified that this expressions allows to generate standard dynamic responses. For instance a one time shock to the mean component in the state vector generates a dynamic response $D_t(i)$ which takes the value 0 until the shock is realised and 1 thereafter (for $t \geq i$). This dynamic response is the same as the one that would be assumed for a level shift outlier (Chen and Liu (1993)).

We then have to estimate the impact effect ω. The attractive feature of the framework developed by De Jong and Penzer (1998) is that the impact effect can be estimated directly from the output of the Kalman Filter and Smoother. For an outlier in period i, we have:

$$\hat{\omega}_i = S_i^{-1} s_i, \qquad Var[\hat{\omega}_i|X] = S_i^{-1} \tag{26}$$

and

$$\begin{aligned} s_i &= \tilde{X}_i' u_i + W_i' r_i, \\ S_i &= \tilde{X}_i' F_i^{-1} \tilde{X}_i + (W_i - K_i\tilde{X}_i)' N_i^{-1} (W_i - K_i\tilde{X}_i) \end{aligned}$$

where F_i, K_i are quantities computed through the Kalman Filter and r_i, N_i, u_i are obtained from the output of the Kalman Smoother. Furthermore, as De Jong and Penzer (1998) demonstrate, it is possible to perform several tests to check for the statistical significance of the estimated impact effect $\hat{\omega}_i$. In particular, an analogue to the standard t-statistics is obtained as:

$$\tau_i = S_i^{-\frac{1}{2}} s_i \tag{27}$$

Finally, I need a procedure to identify the type and the location of outliers. Chen and Liu (1993) proposed a method, within the ARIMA framework for jointly estimating parameters and outliers. I follow their proposed procedure and slightly adapt it for UC model. The procedure goes through the following 3 steps:

1. Estimate parameters assuming that no outliers are present. Using these estimates and starting from $t = 1$, compute $\tau_{AO,1}, \tau_{LS,1}, ...$ and obtain the maximum of the absolute value of these t-statistics. If the maximum is bigger than C where C is a pre-defined critical value, remove the outlier effect according to its type. This is done by augmenting the state vector with an additional component and by adding the dynamic response of this outlier to the matrix of exogenous variables X_t. Repeat this process for all periods in the sample. If some outliers are found during this first loop, re-estimate parameters. Repeat this process until no outliers are found within a loop.
2. In the second stage, estimate jointly all the outliers found in stage 1 and compute their t-statistics. If the minimum of these statistics is smaller than C, remove the outlier and reestimate all of the remaining outliers jointly. This is done until all t-statistics of the remaining outliers are greater than C.
3. Keeping outliers found at the end of stage 2, stage 1 and 2 are repeated (without reestimating parameters at the end of each loop in the first stage) until no outliers are found within a loop in stage 1.

Results on outliers detection and the critical value used can be found in Tables 22 and 23 in Appendix A.2.6.

A.2.5 Population and Classification changes selection process

As explained in Section 3.3.2, population changes estimates can be obtained from aggregate series which should not be affected by classification changes. I start by estimating these population changes from unadjusted official series published by the BLS for (un)employment and inactivity. These results are then used as benchmark for aggregate series obtained from micro data.

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000	2004	2005	2006	2007	2008	2009	2010
E_{bls}	−0.28	−0.34	−0.4	1.51^{***}	-	1.14^{***}	0.28	−0.12	0.41	2^{***}	0.34	0.28	0.47	0.25	0.49	-0.81^{*}	0.94^{**}
				(0.58)		(0.88)	(0.29)	(−0.26)	(0.06)	(2.1)	(−0.41)	(−0.05)	(−0.12)	(0.15)	(−0.60)	(−0.41)	(−0.24)
U_{bls}	-0.36^{**}	0.05	−0.27	0.01	-	−0.01	0.11	0.12	0.02	0.09	−0.03	−0.17	-0.42^{**}	0.05	-0.56^{**}	0.24	−0.66
				(0.04)		(0.18)	(0.03)	(0.03)	(0)	(0.12)	(−0.03)	(−0.00)	(−0.01)	(0.01)	(−0.04)	(−0.04)	(−0.01)
I_{bls}	0.44^{**}	0.84^{***}	0.92^{***}	−0.21	-	0.35	0.14	0.22	−0.05	0.62^{**}	-0.59^{*}	−0.03	−0.03	0.29	-0.73^{**}	−0.42	-0.96^{**}
				(0.33)		(0.05)	(0.15)	(0.23)	(0.25)	(1.00)	(−0.12)	(0.04)	(0.06)	(0.16)	(−0.11)	(−0.03)	−(0.01)

Estimation results for not seasonally adjusted series from the BLS. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. The results reported by the BLS are given in parenthesis. For most population change, these results are obtained by applying new weights in the month (December) prior to their introduction and comparing the estimates obtained with old weights. The 1990 and 2000 population change effects are computed in a different way by the BLS. See `https://www.bls.gov/cps/eetech_methods.pdf`. All these results are expressed in millions.

Table 11: Population estimates : Officially Released Series (1)

Table 11 presents results where all population changes mentioned by the BLS are corrected. Variables for classification changes and the 1994 redesign are also included into the specification to check whether these effects are statistically significant. Since there are no signs of changes in the variance of these stocks before and after 1994, the 1994 redesign is simply assumed to be a constant as are the classification/population changes. Furthermore all the stock series are estimated in log. The results in the following tables are therefore transformed back, expressed in millions and averaged over the period in which they affect series (e.g. for the 1990 pop. change, the effect is averaged over the period 1990-2010, for 1997, the effect is averaged over the period 1997-2010...). The "*" symbol indicates statistical significance at 10%, 5% and 1%. Table 11 also displays in parenthesis, the effects reported by the BLS for population revisions. These effects are not always computed in the same way by the BLS so they only serves as indicating the sign and likely size of population changes.

Table 11 shows that classification effects are never significant for E and U. There seems to be significant effects for inactivity but this series should not be affected by these changes. This could suggest the possibility for a change in the mean of this series over these periods. Note also that, the dummy variables meant to capture the effect of the redesign is statistically significant for U and I. Polivka and Miller (1998) estimate a non significant increase in the unemployment rate (see discussion in Section 3.4). The results displayed in Table 11 show that employment also increase which implies an average percentage increase of 1.5% in the unemployment rate consistent with the results of 1% reported by Polivka and Miller (1998).

Next, I remove classification effects (except 2003-2010 which also captures a population change) and reestimate all the population effects. The results are displayed in Table 12

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000	2004	2005	2006	2007	2008	2009	2010
E_{bls}	-0.23	-	-	1.51^{***}	-	1.08^{***}	0.26	-0.07	0.5	2.08^{***}	0.26	0.19	0.41	0.17	0.39	-0.88^{**}	0.83^{*}
				(0.58)		(0.88)	(0.29)	(-0.26)	(0.06)	(2.1)	(-0.41)	(-0.05)	(-0.12)	(0.15)	(-0.60)	(-0.41)	(-0.24)
U_{bls}	-0.36	-	-	0.01	-	-0.01	0.11	0.12	0.02	0.09	-0.02	-0.17	-0.42^{**}	0.05	-0.56^{**}	0.24	-0.65
				(0.04)		(0.18)	(0.03)	(0.03)	(0)	(0.12)	(-0.03)	(-0.00)	(-0.01)	(0.01)	(-0.04)	(-0.04)	(-0.01)
I_{bls}	0.44	-	-	-0.21	-	0.35	0.14	0.22	-0.05	0.62^{**}	-0.59^{*}	-0.03	-0.03	0.29	-0.73^{**}	-0.42	-0.95^{**}
				(0.33)		(0.05)	(0.15)	(0.23)	(0.25)	(1.00)	(-0.12)	(0.04)	(0.06)	(0.16)	(-0.11)	(-0.03)	$-(0.01)$

Estimation results for not seasonally adjusted series from the BLS. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. The results reported by the BLS are given in parenthesis. For most population change, these results are obtained by applying new weights in the month (December) prior to their introduction and comparing the estimates obtained with old weights. The 1990 and 2000 population change effects are computed in a different way by the BLS. See `https://www.bls.gov/cps/eetech_methods.pdf`. All these results are expressed in millions.

Table 12: Population estimates : Officially Released Series (2)

From the results in Table 12, the 1990, 1997, 1998, 2000, 2003 and 2007 population changes have signs that corresponds to what the BLS reports for employment, unemployment and inactivity. Exceptions are the 1990 pop. change for unemployment and 2003 one for inactivity. In terms of magnitude, the results appears to be quite consistent with the BLS results for the 1990 (E), 1997 (E and I) and 2007 (E) pop.changes. The estimated effects for the 2000 and 2003 pop. changes are larger than what the BLS reports for employment. For these 2 effects, the BLS studied in more details the adjustment implied by these 2 pop. changes. I discuss these 2 population revisions in the following paragraph. In the next table, I keep effects that appear to have a sign consistent with the BLS and reestimate the population effects.

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000	2004	2005	2006	2007	2008	2009	2010
E_{bls}	-0.2	-	-	1.32^{***}	-	1.06^{***}	0.13	-0.31	-	1.89^{***}	-	-	-	-0.2	-	-	-
				(0.58)		(0.88)	(0.29)	(-0.26)		(2.1)				(0.15)			
U_{bls}	-0.39^{**}	-	-	0.03	-	0.03	0.14	0.14	-	0.13	-	-	-	0.41	-	-	-
				(0.04)		(0.18)	(0.03)	(0.03)		(0.12)				(0.01)			
I_{bls}	0.45^{***}	-	-	0.08	-	0.37^{*}	0.17	0.29	-	0.7^{***}	-	-	-	0.67^{**}	-	-	-
				(0.33)		(0.05)	(0.15)	(0.23)		(1.00)				(0.16)			

Estimation results for not seasonally adjusted series from the BLS. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. The results reported by the BLS are given in parenthesis. For most population change, these results are obtained by applying new weights in the month (December) prior to their introduction and comparing the estimates obtained with old weights. The 1990 and 2000 population change effects are computed in a different way by the BLS. See `https://www.bls.gov/cps/eetech_methods.pdf`. All these results are expressed in millions.

Table 13: Population estimates : Officially Released Series (3)

The 2007 pop. change is no longer of the right sign for employment and the estimate for unemployment and inactivity are much higher compared to the BLS results. The 1990 pop. change is now of the right sign for unemployment but the estimated effect for inactivity is much larger than the BLS results. The new estimate for the 2003 pop. change for inactivity also has now a sign consistent with the BLS evidence.

I then remove the 2007 population change and try to add the 2008 and 2009 population change effects as these 2 pop. changes seem to have substantial effects according to the BLS results. When these are introduced, the estimates are not consistent with the BLS results (they're much larger) in particular for unemployment.[40] Therefore it seems that the 1990, 1997, 1998, 2000 and 2003 population changes are the ones that have signs consistent with what the BLS reports. The 1997 and 1998 pop. changes for unemployment are however larger than the BLS estimates (around 5 time bigger) and since these are not statistically significant for any labor market state, I decide to remove them. It should also be noted that according to the BLS results, these 2 population change seem to have small effects on all series. The 1990 pop. change for inactivity is also very large compared to the BLS results which reports an effect close to 0. Therefore, I also remove this effect for inactivity. The new estimates for the officially released series are displayed in Table 14.

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000
E_{bls}	-0.2	-	-	1.34^{***}	-	1.09^{***}	-	-	-	1.92^{***}
				(0.58)		(0.88)				(2.1)
U_{bls}	-0.38^{**}	-	-	0.02	-	0.02	-	-	-	0.14
				(0.04)		(0.18)				(0.12)
I_{bls}	0.49^{***}	-	-	0.09	-	-	-	-	-	0.73^{***}
				(0.33)						(1.00)

Estimation results for not seasonally adjusted series from the BLS. I average the value of a given effect over the time period in which it affects the series. For the 1994 redesign and the 76-82 and 83-91 classification changes, effects are average over the entire period in which they could affect series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. The results reported by the BLS are given in parenthesis. For most population change, these results are obtained by applying new weights in the month (December) prior to their introduction and comparing the estimates obtained with old weights. The 1990 and 2000 population change effects are computed in a different way by the BLS. See `https://www.bls.gov/cps/eetech_methods.pdf`. All these results are expressed in millions.

Table 14: Population estimates : Officially Released Series (4)

[40] A possible explanation for this observation comes from the Great Recession which also affected the level of these series over this period. I tried including a recession dummy that would capture level changes originating from recessions but this did not improve the results.

All the estimated effects have now signs that are in line with the BLS evidence. For the 1990 pop. change, the size of the estimated effects are quite consistent with the BLS results for employment but much smaller for unemployment. For the 2000 pop. change, the effect displayed in Table 14 is similar to what the BLS reports for employment and unemployment and smaller for inactivity.

For these time series, the BLS further suggests a potential effect of this population change on their trends. In January 2000, the BLS reports initial increases of 1.6 millions and 1 million in the labor force and inactivity. By may 2002 (results displayed in the above table), labor force and inactivity had increased by 2.2 (2.1 for E and 0.12 for U) and 1.3 millions. Results in terms employment and unemployment are not available for January 2000 but assuming that 95% of the labor force was employed, we can assume that the initial increase was around 1.5 and 0.1 million in employment and unemployment. I therefore try to adjust the specification for this population revision to see if a change in trend can be detected.

First, it should be noted that even though population changes are modelled as constant, the fact that a multiplicative model is specified for all these series implies that estimated effects are constant only in relative terms (they are constant percentages). In other words, the presence of a trend in the series would mechanically raise the estimated population effects in level. There are increasing trends in the series of E, U and I over the period 2000-2003 but these trends are not enough to obtain a population effect that match the evidence reported by the BLS.

I have also tried to include a change in trend from January 2000 onwards but the estimated trend change was of the wrong sign. I further check if adding a random component to the 2000 pop. change to allow for a random walk specification could capture this change in trend but the estimated results were not improved. Therefore, it seems that the set-up developed in this paper is unable to detect any trend change coming from this population change. However, It should be noted that series from the BLS are only used as benchmark and the correction for the 2000 pop. change will be implemented on micro data series by using old unadjusted weights (see Section 3.3.2). Moreover, the results reported by the BLS are consistent with an average increase over the 2000-02 period of 1.8, 1.9 million for employment (assuming a linear trend) which corresponds to the value reported in Table 14.

Finally, the fact that the specification does not capture this trend change can explain why the 2003 estimates are higher than reported by the BLS for employment. The sum of the 2000 and 2003 pop. changes in January 2003, according to the BLS results, should lie around 2.7, 2.8 millions (2.1 in May 2002 and we can assume that this difference kept increasing + 0.6 in January 2003. The estimation results in Table 14 imply an increase of around 3.2 millions (1.9+1.3) on average after January 2003 for employment.

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000	2004	2005	2006	2007	2008	2009	2010
E_{micro}																	
	-0.06	-0.33	-0.4	3.44^{***}	-1.59^{***}	0.94^{***}	0.4	0.24	0.4	0.49	0.57	0.37	0.69	0.21	0.05	-0.65	1.09^{**}
	-0.04	-	-	3.43^{***}	-1.59^{***}	0.93^{***}	0.38	0.22	0.38	0.49	0.55	0.35	0.67	0.18	0.02	-0.67	1.07^{**}
	-0.01	-	-	3.19^{***}	-1.62^{***}	0.89^{***}	0.26	0	-	-	-	-	-	-0.14	-	-	-
	-	-	-	3.19^{***}	-1.62^{***}	0.98^{***}	-	-	-	-	-	-	-	-	-	-	-
U_{micro}																	
	-0.62^{***}	0.17	-0.21	0.04	-0.34^{**}	0.09	0.17	0.09	-0.05	0.32	-0.12	-0.22	-0.47^{**}	-0.07	-0.51^{**}	0.15	-0.72^{*}
	-0.61^{***}	-	-	0.05	-0.37^{**}	0.1	0.17	0.08	-0.06	0.31	-0.11	-0.22	-0.46^{**}	-0.06	-0.5^{*}	0.16	-0.71^{*}
	-0.61^{***}	-	-	0.12	-0.41^{***}	0.11	0.18	0.12	-	-	-	-	-	0.09	-	-	-
	-0.59^{***}	-	-	0.11	-0.41^{***}	0.1	-	-	-	-	-	-	-	-	-	-	-
I_{micro}																	
	0.41	0.64^{*}	0.73^{***}	1.75^{***}	-1.42^{***}	0.28	0.08	0.08	0.3	-0.41	-0.34	0.55	0.5	1.02^{***}	0.37	0.29	-0.49
	0.35	-	-	1.84^{***}	-1.43^{***}	0.34	0.11	0.11	0.34	-0.37	-0.27	0.59^{*}	0.53	1.05^{***}	0.44	0.37	-0.41
	0.37	-	-	1.79^{***}	-1.42^{***}	0.33	0.05	0.07	-	-	-	-	-	0.92^{***}	-	-	-
	0.4	-	-	1.75^{***}	-1.42^{***}	-	-	-	-	-	-	-	-	-	-	-	-

Estimation results for not seasonally adjusted Micro Data series from the CPS. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. All these results are expressed in millions.

Table 15: Population estimates: Micro Data

I then perform the exact same steps on micro data obtained from CPS data. Series for employment and unemployment are constructed by aggregating series by skill ($E_{micro} = E^h + E^m + E^l$). For unemployment, this implies that the aggregated series does not contain *New Unemployed Entrants* which have been dropped because no information were available on their occupations. So while $E_{agg} \approx E_{bls}$ and $I_{micro} \approx I_{bls}$, we have $U_{micro} < U_{bls}$. The micro data from the NBER are also not corrected for population changes prior to 1980 so an additional dummy variable is included for this period (see Figure 5 in Section 2.2). As mentioned above, the 2000 population adjustment can be corrected by using non revised weights for this period which are the ones available by default when downloading CPS files from the NBER website. This further implies that the 2003 pop. change will now also capture the effect of the 2000 pop. change. Classification changes should also not affect these series.

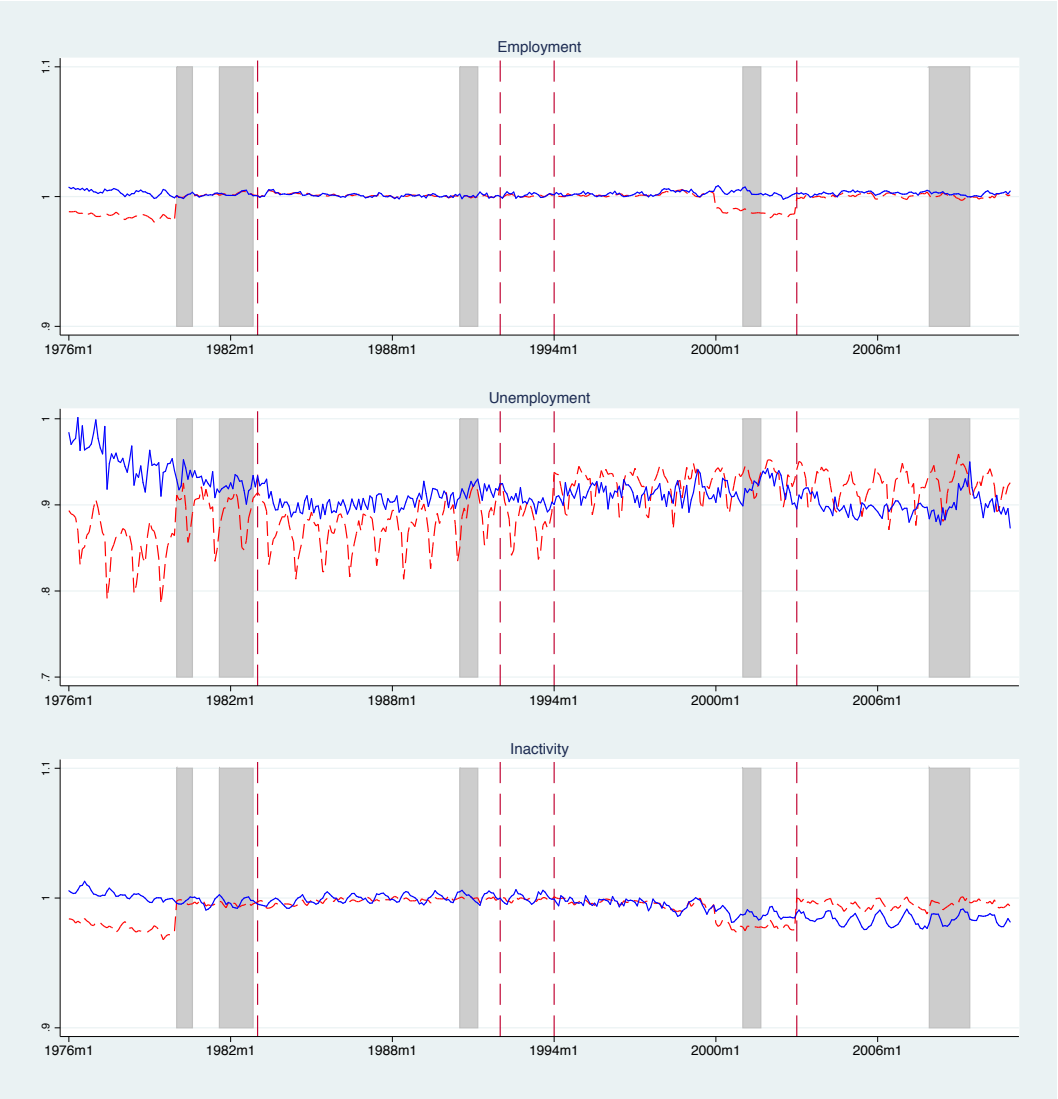
The estimated effects are displayed in Table 15 and are obtained by applying the same step as described above for official series (with each row giving the results of one step). For employment, the results are quite similar between the BLS and micro series. In particular the 2003 pop. change is estimated to be 3.19 millions on average for the micro series while the sum of the 2000 and 2003 pop. changes is equal to 3.26 millions (1.92+1.34) for the BLS series. The estimates for the 1990 pop. change is also consistent between the 2 series. There is a difference in estimates for the 1994 redesign which is negative for the BLS series and close to 0 for the micro series. This effect being not significative for both series, I decide to remove it.

For unemployment, the effect of the 1990 and 2003 pop. changes are also similar for BLS and micro series. The effect of the 1990 pop change is actually bigger for micro series but this is more consistent with the results reported by the BLS in Table 14 (0.18 millions). The main difference between the estimation results regards the 1994 component which is much larger for the micro data series. This is however consistent with the drop in *New Unemployed Entrants* following the redesign of the CPS questionnaire (see Section 2.2 and the discussion in Section 3.4).

For inactivity, the effect of the 1994 redesign is similar for both series (but no longer significative for the micro data series). The effect of the 2003 pop. change is quite different (1.75 for Micro vs 0.73+0.09 for BLS). As for the 1990 pop. change for unemployment, the estimation results for micro series is closer to what the BLS reports (around 1.33 millions). Note also that the variable for the 1976-79 pop. change is significative for all series.

In order to have clearer idea of these differences, Figure 13 plots the ratio of the corrected micro series to the corrected BLS series (in blue) against the same ratio obtained for the original (uncorrected) series (dotted red line). This figure shows that the effects for employment are indeed quite consistent and that the correction for the 1976-1979 period applied to the micro series allows to bridge the gap with the BLS series. Likewise for inactivity but there is a difference of around 2% between the micro and BLS series after 2000. This is explained by the fact that the estimated results for the 2000 pop. change is smaller for the BLS series (see Table 14). The estimated effect for the Micro series being more in line with the BLS results, this discrepancy does not appear to be very relevant. For unemployment, there seem to be slight problems over the period 1976-1982. One solution is therefore to add the classification effect for this period and reestimate the effects to see if this allows for an improvement.

The results are presented in Table 16 and Figure 14 shows that the inclusion of this classification allows to correct the difference displayed in Figure 13. This fact could be explained by a potential interaction between the 1976-82 classification change and *New Unemployed Entrants*. Note that this correction allows to keep the cyclical property of the micro data series. In particular, the difference between the 2 unemployment series decreases during recession (the ratio increases) when less new entrants enter the labor market.



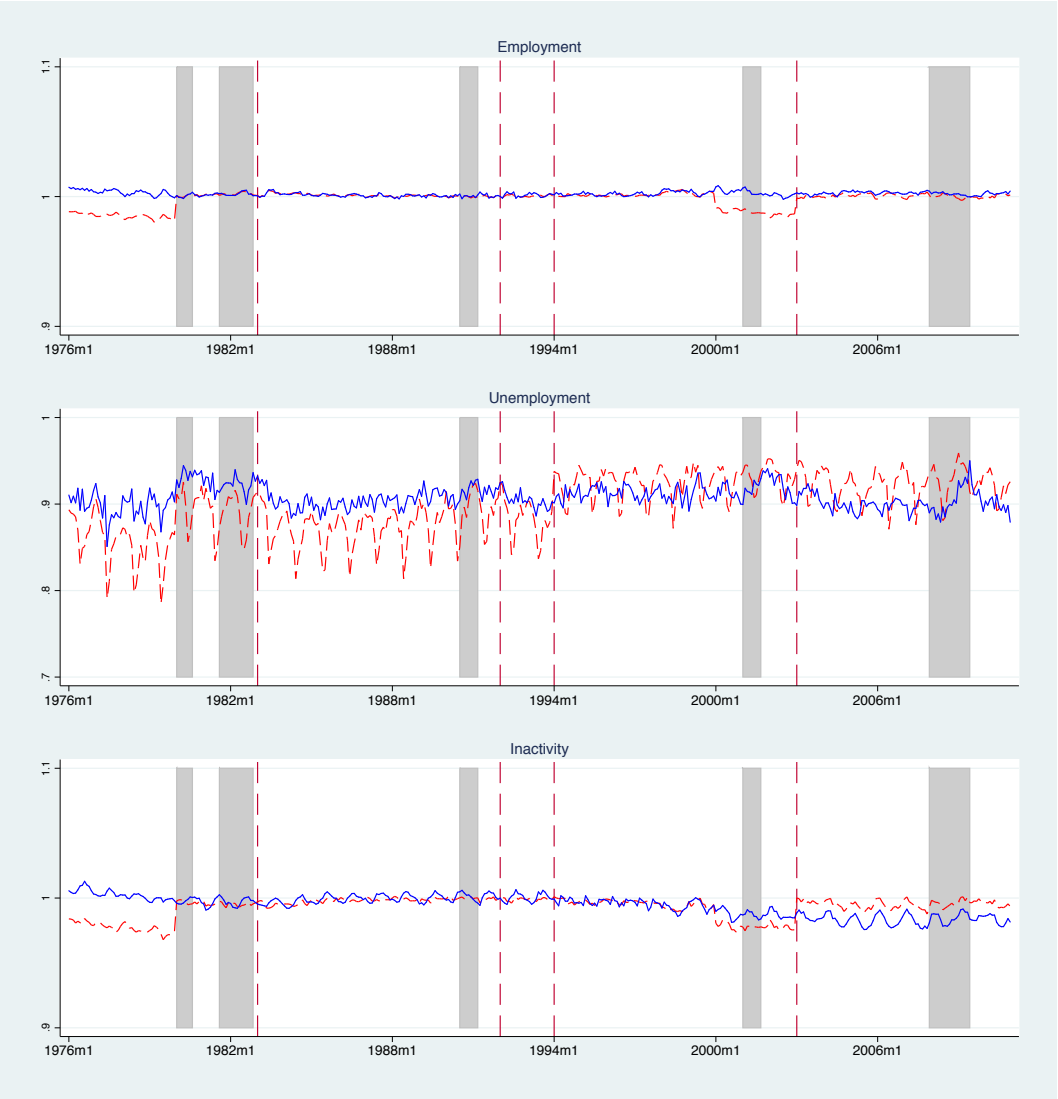
Ratio of Micro data series to the BLS series. The ratio of corrected series is plotted in blue while the ratio of original series is plotted in red.

Figure 13: Ratio of (un)corrected Micro to BLS Series: Intermediate Step

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990	1997	1998	1999	2000
E_{bls}	-	-	-	1.34^{***}	-	1.08^{***}	-	-	-	1.92^{***}
E_{micro}	-	-	-	3.19^{***}	-1.62^{***}	0.98^{***}	-	-	-	-
U_{bls}	-0.38^{**}	-	-	0.02	-	0.02	-	-	-	0.14
U_{micro}	-0.59^{***}	0.33^{**}	-	0.09	-0.37^{**}	0.09	-	-	-	-
I_{bls}	0.49^{***}	-	-	0.09	-	-	-	-	-	0.73^{***}
I_{micro}	0.4	-	-	1.75^{***}	-1.42^{***}	-	-	-	-	-

Estimation results for not seasonally adjusted Micro Data series from the CPS and for BLS series. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. All these results are expressed in millions

Table 16: BLS vs Micro : Final Estimated Effects



Ratio of Micro data series to the BLS series. The ratio of corrected series is plotted in blue while the ratio of original series is plotted in red.

Figure 14: Ratio of (un)corrected Micro to BLS Series: Final Results

Having an idea of which population changes appears to matter for aggregate series, I focus on disaggregated series by occupation. I am unsure of the effect of classification changes on employment and unemployment. Therefore, I start by correcting the labor force series by skill (e.g. $L^h = E^h + U^h$) which will allow to get an idea on which classification effects should be considered for each occupations. I will then correct each individual series using the results for the labor force series as benchmark.

From Table 17, results for the 1976-79, 1990 and 2003 population changes are consistent in magnitudes and signs with what is reported until now for official and micro series. For instance the 1976-79 pop. change is estimated to increase aggregate employment and unemployment by around 2 millions in Table 16 (1.62+0.37) while looking at the results for the labor force series we obtain an increase of 1.7 millions (0.61+0.58+0.49). Similar computations can be done for the 1990 and 2003 pop. changes (as well as for the 1994 redesign) which confirm that results are quite consistent with those obtained for Micro series.

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990
L^h	-0.27^{*}	-0.08	-0.13	-1.86^{***}	-0.61^{***}	-0.1
L^m	-0.24	0.92^{***}	-0.71^{***}	3.45^{***}	-0.58^{***}	0.28
L^l	-0.13	-0.99^{***}	0.17	1.14^{***}	-0.49^{***}	0.53^{**}

Results for Labor Force series by skills. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1%. All these results are expressed in millions.

Table 17: Labor Force Series (1)

With regards to classification changes, it was mentioned in Section 3.3.2, that these effects should generate a reallocation between occupations without changing the aggregate level of stocks (at least for employment and inactivity). From the results in Table 17, this condition is not met for the 1976-82 and 1983-91 classification changes which results in a net increase of aggregate labor force (particularly the 1983-91 classification change). As explained previously (see also Figures 13 and 14), including the 1976-82 classification change helped in making the unemployment series from Micro data more consistent with the BLS one. Therefore, I reestimate the classification effects by only including a variable for the 1976-82 classification:

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990
L^h	-0.26^{*}	0.01	-	-1.86^{***}	-0.61^{***}	-0.11
L^m	-0.3	1.44^{***}	-	3.5^{***}	-0.59^{***}	0.36
L^l	-0.15	-1.15^{***}	-	1.14^{***}	-0.49^{***}	0.56^{**}

Results for Labor Force series by skills. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. All these results are expressed in millions.

Table 18: Labor Force Series (2)

The results displayed in Table 18 seem to indicate a reallocation between middle and low skill labor force series. Moreover, accounting for the fact that the 1976-1982 classification estimate for unemployment is on average equal to 0.30 million (Table 16), we obtain that the increase in low skill labor force is almost perfectly compensated by a decrease in middle skill labor force (respectively -1.15 and 1.14+0.30 from middle skill unemployment). This further suggests that the 1976-1982 has an effect on middle skill unemployment.

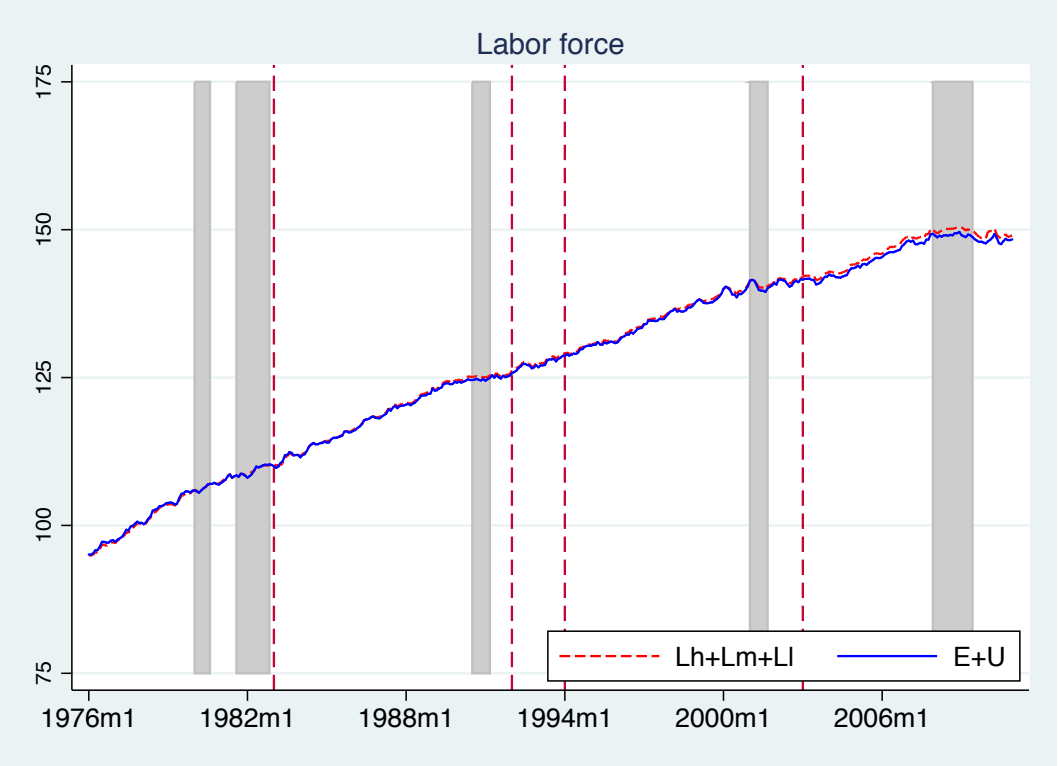
The graphical evidence presented in Figure 15 seems support this choice as the corrected labor force series obtained by summing E_{micro} and U_{micro} is very similar to the series obtained by summing L^h, L^m and L^l. Finally, the 1990 pop. change for high skill labor force has the wrong sign. Since this effect is not statistically significant, I remove it from the specification of this series. This allows to match more closely the sum of the estimates obtained for E_{micro} and U_{micro}.

The final results are displayed in Table 19:

	$76-94$	$76-82$	$83-91$	$03-10$	$76-79$	1990
L^h	-0.26^{*}	-	-	-1.86^{***}	-0.61^{***}	-
L^m	-0.3	1.44^{***}	-	3.5^{***}	-0.59^{***}	0.36
L^l	-0.15	-1.15^{***}	-	1.14^{***}	-0.49^{***}	0.56^{**}

Results for Labor Force series by skills. I average the value of a given effect over the time period in which it affects the series. The "*" symbol indicates statistical significance at 10%, 5% and 1% and "-" implies that a variable for this effect was not included into the specification. All these results are expressed in millions.

Table 19: Labor Force Series: Final Results



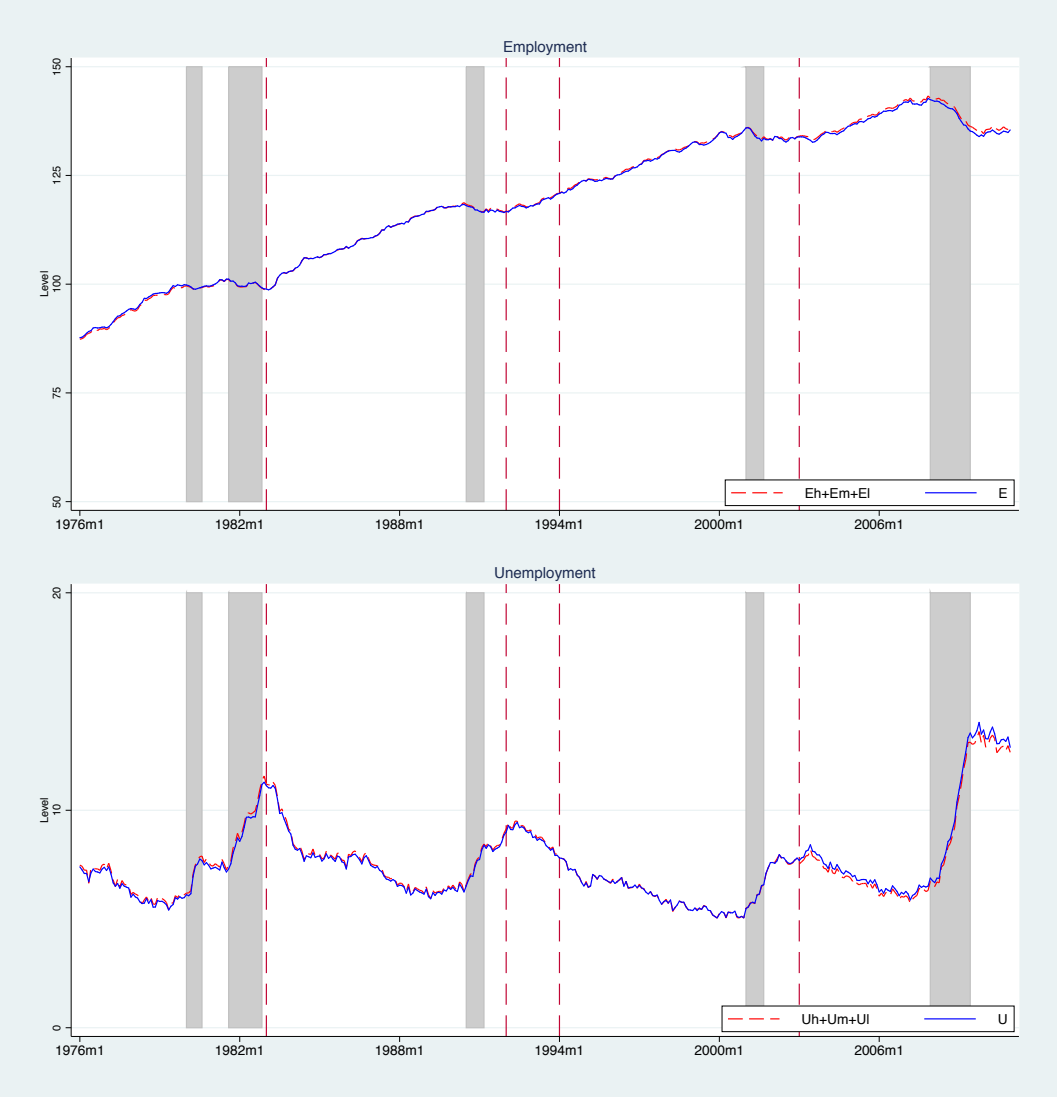
Aggregate labor force series in level expressed in millions and obtained from correcting micro data series for employment and unemployment (in blue) against aggregate labor force computed from labor force series by occupations (dotted red line). Shaded areas display recession periods as defined by the NBER.

Figure 15: Labor Force Series

From this selection process, I therefore decide to correct the 1976-79, the 1990 and the 2003 population changes. These effects being the one consistent with BLS results and having, usually, a significant effect on aggregate series. Note that these effects also correspond to the largest adjustments operated by the BLS across the 1976-2010 period. For classification changes, I correct the 1976-1982

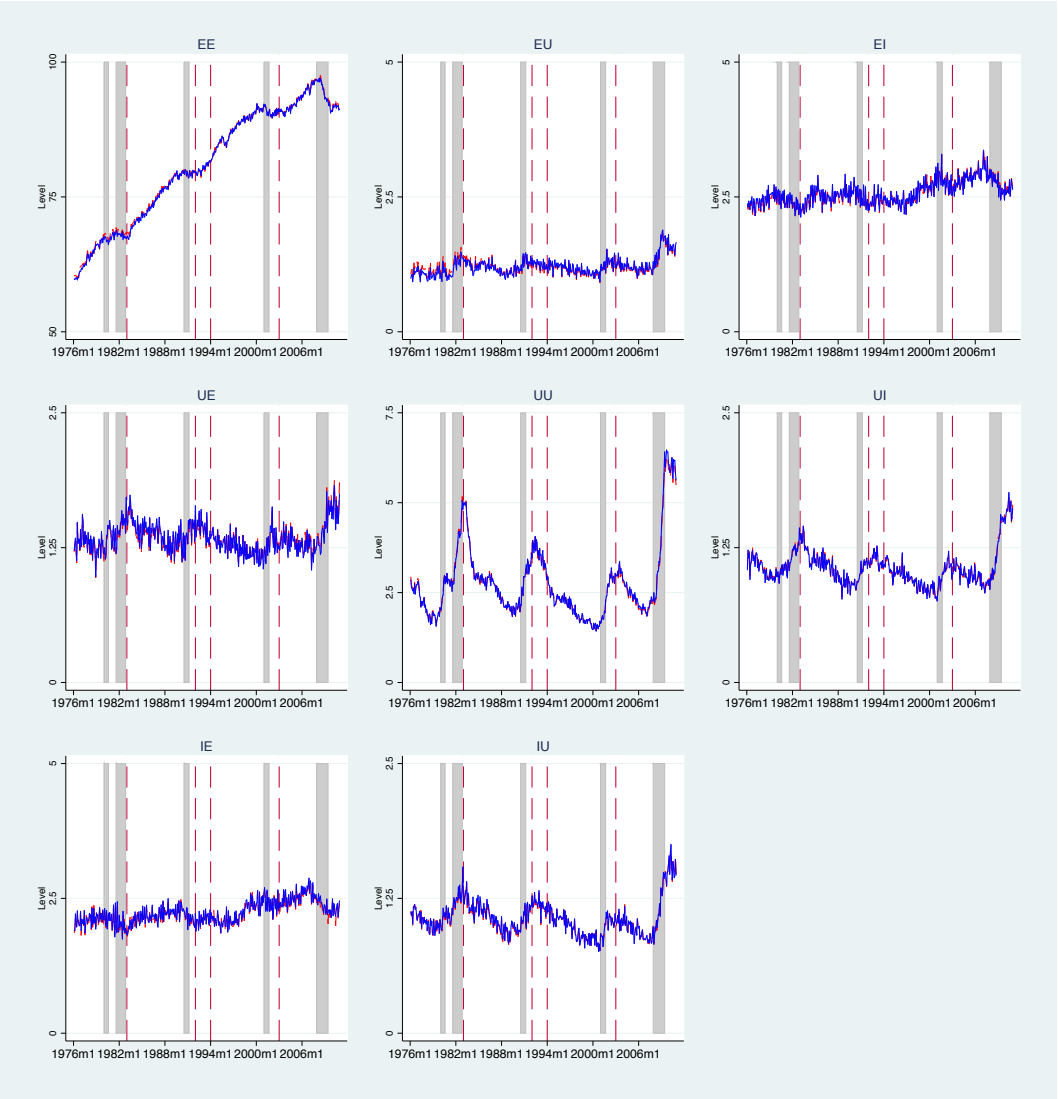
and 2003-10 changes. This choice matches evidences from Cortes et al. (2016) on the fact that the 1976-1982 classification change was a major modification compared to the one introduced in 1983 and the classification of 1992 only introduced minor changes compared to the 1983-1991 one. Since the 1992-2002 classification is used as reference, only correcting for the 1976-1982 is consistent with this evidence.

Having an idea of which population and classification changes to correct for, I can move to employment and unemployment series by occupations. The results for stocks and flows are presented in the main text in Section 3.4. In Figures 16 and 17, I plot series obtained from correcting aggregate series against series obtained by aggregating corrected series by occupations. These 2 figures show that the estimated effects for stock and gross flow series by occupations are consistent with those estimated at the aggregate level.



Corrected aggregate stocks series in level and expressed in millions. The blue line displays the series obtained by correcting directly the aggregate series while the red dashed line plots the sum of series by occupations. Shaded areas display recession periods as Defined by the NBER.

Figure 16: Aggregate vs Disaggregated Corrected Series: Stocks



Corrected aggregate flow series in level and expressed in millions. The blue line displays the series obtained by correcting directly the aggregate series while the red dashed line plots the sum of series by occupations. Shaded areas display recession periods as Defined by the NBER.

Figure 17: Aggregate vs Disaggregated Corrected Series: Flows

A.2.6 Estimation Results: Specifications, Tests and Outliers

Tables 20 and 21 give the specification retained to correct all series. Tables 22 and 23 display the various test results that are presented in Section 3.3.1. 2 normality tests, an F-test for equal variance and Ljung-Box Q-test (LBQ) for the presence of autocorrelations are performed on standardized innovations ($\tilde{v}_{\delta,t} = \frac{v_{\delta,t}}{F_t}$. See Appendix A.2.3). The last four columns of these 2 tables display results on the outlier detection.

Specifications

Pop/Flows	**Log**	**Baseline Components**										**1994 Components**								*Cov.*
		μ	σ_μ	ν	σ_ν	γ	σ_γ	ε	p	q	σ_ε	μ_χ	$\sigma_{\chi\mu}$	ν_χ	$\sigma_{\chi\nu}$	ε_χ	p	q	$\sigma_{\chi\varepsilon}$	
Stocks																				
Employment																				
E_{BLS}	1	1	0	1	1	1	1	1	12	1	1	0	0	0	0	0	0	0	0	0
E_{micro}	1	1	0	1	1	1	1	1	14	1	1	0	0	0	0	0	0	0	0	0
E^h	1	1	1	1	1	1	1	1	13	2	1	1	0	0	0	1	0	0	1	0
E^m	1	1	1	1	1	1	1	1	12	1	1	1	0	0	0	0	0	0	0	0
E^l	1	1	1	1	1	1	1	1	14	1	1	1	0	0	0	0	0	0	0	0
Unemployment																				
U_{BLS}	1	1	0	1	1	1	0	1	12	0	1	1	0	0	0	0	0	0	0	0
U_{micro}	1	1	0	1	1	1	1	1	9	1	1	1	0	0	0	0	0	0	0	0
U^h	1	1	0	1	1	1	0	1	8	1	1	1	0	0	0	0	0	0	0	0
U^m	1	1	0	1	1	1	1	1	12	1	1	1	0	0	0	0	0	0	0	0
U^l	1	1	0	1	1	1	0	1	12	0	1	1	0	0	0	0	0	0	0	0
Inactivity																				
I_{BLS}	1	1	0	1	1	1	1	1	12	1	1	1	0	0	0	0	0	0	0	0
I	1	1	0	1	0	1	1	1	12	0	1	1	0	0	0	0	0	0	0	0
Labor Force																				
L^h	1	1	0	1	1	1	1	1	12	2	1	1	0	0	0	1	0	0	1	0
L^m	1	1	0	1	1	1	1	1	7	1	1	1	0	0	0	0	0	0	0	0
L^l	1	1	0	1	1	1	1	1	12	1	1	1	0	0	0	0	0	0	0	0
Flows																				
Employment																				
EE	1	1	0	1	1	1	1	1	10	1	1	1	0	0	0	0	0	0	0	0
EU	1	1	0	1	1	1	1	1	12	1	1	1	0	0	0	0	0	0	0	0
EI	1	1	1	0	0	1	1	1	9	1	1	1	0	0	0	0	0	0	0	0
E^hE^h	1	1	1	1	1	1	1	1	12	0	1	1	0	0	0	1	0	0	1	0
E^hE^m	0	1	0	1	0	1	1	1	8	0	1	1	0	1	0	1	2	0	1	1
E^hE^l	0	1	0	1	0	1	1	1	3	1	1	1	0	1	0	1	2	0	1	0
E^hU^h	1	1	1	1	0	1	1	1	4	0	1	1	0	0	0	1	0	0	1	0
E^hU^m	0	1	1	1	0	1	1	1	2	0	1	1	1	1	0	1	0	0	1	0
E^hU^l	0	1	1	1	0	1	1	1	12	0	1	1	1	0	0	1	0	0	1	0
E^hI	1	1	0	1	0	1	1	1	6	0	1	1	0	0	0	1	1	0	1	0
E^mE^h	0	1	0	1	0	1	1	1	5	1	1	1	0	1	0	1	2	0	1	0
E^mE^m	0	1	1	1	0	1	1	1	7	0	1	1	0	1	0	1	2	0	1	0
E^mE^l	1	1	1	1	1	1	1	1	12	0	1	1	0	0	0	1	1	0	1	0
E^mU^h	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	1	2	0	1	0
E^mU^m	0	1	1	0	0	1	1	1	11	1	1	1	1	0	0	1	2	0	1	1
E^mU^l	1	1	1	1	0	1	1	1	8	0	1	1	0	0	0	0	0	0	0	0
E^mI	1	1	1	0	0	1	1	1	9	0	1	1	0	0	0	0	0	0	0	0
E^lE^h	0	1	1	0	0	1	0	1	4	1	1	1	1	0	0	1	2	0	1	0
E^lE^m	1	1	0	1	1	1	1	1	12	2	1	1	0	0	0	0	0	0	0	0
E^lE^l	0	1	1	1	0	1	1	1	7	1	1	1	1	1	0	1	0	0	1	0
E^lU^h	0	1	1	0	0	1	0	1	6	0	1	1	0	0	0	1	2	0	1	1
E^lU^m	1	1	1	1	0	1	1	1	9	0	1	1	0	1	0	1	0	0	1	0
E^lU^l	0	1	1	1	0	1	0	1	7	0	1	1	1	1	0	1	2	0	1	1
E^lI	1	1	1	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0

A 1 indicates that the component is included into the specification. For AR and MA parameters, this table gives the maximum lag for the irregular specifications. The *Cov.* column specifies whether a covariance parameter between the irregular component and the irregular component of the 1994 redesign is included.

Table 20: Unobserved Component Model Specification (1)

Flows	Log	Baseline Components										1994 Components								*Cov.*
		μ	σ_μ	ν	σ_ν	γ	σ_γ	ε	p	q	σ_ε	μ_χ	$\sigma_{\chi\mu}$	ν_χ	$\sigma_{\chi\nu}$	ε_χ	p	q	$\sigma_{\chi\varepsilon}$	
Flows																				
Unemployment																				
UE	1	1	0	0	0	1	0	1	12	5	1	1	0	0	0	0	0	0	0	0
UU	1	1	1	1	1	1	1	1	10	1	1	1	0	0	0	0	0	0	0	0
UI	1	1	0	1	1	1	1	1	4	1	1	1	0	0	0	0	0	0	0	0
U^hE^h	1	1	1	1	0	1	0	1	1	0	1	1	0	0	0	0	1	0	0	0
U^hE^l	1	1	1	1	0	1	0	1	9	0	1	1	0	0	0	0	0	0	0	0
U^hE^m	1	1	1	0	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0
U^hI	1	1	1	0	0	1	0	1	9	1	1	1	0	0	0	1	0	0	1	0
U^hU^h	1	1	0	1	1	1	1	1	6	1	1	1	0	0	0	1	0	0	1	0
U^hU^l	0	1	1	0	0	1	0	1	8	0	1	1	1	0	0	1	0	0	1	0
U^hU^m	0	1	1	0	0	1	0	1	6	1	1	1	1	0	0	1	1	0	1	0
U^mE^h	1	1	0	0	0	1	0	1	6	0	1	1	0	0	0	0	0	0	0	0
U^mE^l	1	1	1	0	0	1	0	1	7	0	1	1	0	0	0	0	0	0	0	0
U^mE^m	1	1	0	0	0	1	0	1	9	1	1	1	0	0	0	0	0	0	0	0
U^mI	1	1	1	0	0	1	1	1	8	0	1	1	0	0	0	0	0	0	0	0
U^mU^h	0	1	0	0	0	1	1	1	10	0	1	1	0	1	0	1	2	0	1	1
U^mU^l	1	1	0	1	1	1	1	1	14	0	1	1	0	0	0	0	0	0	0	0
U^mU^m	0	1	0	0	0	1	1	1	3	0	1	1	0	0	0	1	2	0	1	1
U^lE^h	1	1	0	1	0	1	1	1	13	0	1	1	0	0	0	0	0	0	0	0
U^lE^l	1	1	1	0	0	1	0	1	13	1	1	1	0	0	0	0	0	0	0	0
U^lE^m	1	1	1	0	0	1	1	1	10	0	1	1	0	0	0	0	0	0	0	0
U^lI	1	1	1	0	0	1	0	1	8	0	1	1	0	0	0	0	0	0	0	0
U^lU^h	0	1	1	1	0	1	0	1	11	0	1	1	1	1	0	1	2	0	1	0
U^lU^h	0	1	1	0	0	1	0	1	10	0	1	1	0	0	0	1	2	0	1	1
U^lU^h	1	1	1	1	0	1	0	1	10	0	1	1	0	0	0	1	0	0	1	0
Inactivity																				
IE	1	1	1	0	0	1	1	1	9	1	1	1	0	0	0	0	0	0	0	0
IU	1	1	1	0	0	1	1	1	3	1	1	1	0	0	0	0	0	0	0	0
II	1	1	1	1	1	1	1	1	9	0	1	1	0	0	0	0	0	0	0	0
IE^h	1	1	1	0	0	1	1	1	10	1	1	1	0	0	0	0	0	0	0	0
IE^m	1	1	1	0	0	1	1	1	4	0	1	1	0	0	0	0	0	0	0	0
IE^l	1	1	1	0	0	1	1	1	5	1	1	1	0	0	0	0	0	0	0	0
IU^h	1	1	1	1	0	1	1	1	8	0	1	1	0	0	0	0	0	0	0	0
IU^m	1	1	1	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	0
IU^l	1	1	1	0	0	1	0	1	4	0	1	1	0	0	0	0	0	0	0	0

A 1 indicates that the component is included into the specification. For AR and MA parameters, this table gives the maximum lag for the irregular specifications. The *Cov.* column specifies whether a covariance parameter between the irregular component and the irregular component of the 1994 redesign is included.

Table 21: Unobserved Component Model Specification (2)

Tests and Outliers

	Norm. tests		Het. test		LBQ. test			Outliers			
	JB	KS	F stat	Reject	lag1	lag6	lag12	Num.	Type	Location	C
Stocks											
Employment											
E_{BLS}	0.50	0.21	0.79	1	0.98	0.56	0.66	3	8; 4; 4	221; 396; 89	3.4
E_{micro}	0.39	0.71	0.81	0	0.73	0.89	0.58	1	4	89	3.5
E^h	0.16	0.10	0.93	0	0.84	0.81	0.60	0	-1	-1	4
E^m	0.50	0.09	0.95	0	0.95	0.79	0.88	0	-1	-1	3.5
E^l	0.50	0.35	0.94	0	0.72	0.79	0.50	1	4	89	4
Unemployment											
U_{BLS}	0.27	0.42	1.01	0	.99	0.83	0.74	1	4	51	3.5
U_{micro}	0.50	0.85	0.92	0	0.88	0.86	0.38	2	4; 4	51; 231	3.5
U^h	0.47	0.85	0.90	0	0.89	0.80	0.81	1	13	307	3.5
U^m	0.50	0.85	1.10	0	0.85	0.36	0.37	1	4	51	3.25
U^l	0.46	0.72	1.15	0	0.48	0.78	0.70	1	4	51	3.75
Inactivity											
I_{BLS}	0.31	0.26	0.94	0	0.92	0.81	0.74	2	4; 14	192; 269	3.75
I	0.50	0.48	0.83	0	0.79	.85	0.26	3	11; 9; 14	229; 257; 218	3.75
Labor force											
L^h	0.13	0.50	0.91	0	0.83	0.89	0.59	0	-1	-1	4
L^m	0.50	0.22	1.08	0	0.94	0.96	0.01	0	-1	-1	4
L^l	0.28	0.46	0.95	0	0.53	0.34	0.51	0	-1	-1	4
Flows											
Employment											
EE	0.03	0.69	0.72	1	0.46	0.42	0.31	1	12	241	4
EU	0.35	0.18	1.01	0	0.90	0.94	0.79	0	-1	-1	3.75
EI	0.50	0.76	0.93	0	0.93	0.51	0.57	2	9; 10	149; 288	3.5
E^hE^h	0.5	0.2	0.97	0	1	0.32	0.55	0	-1	-1	3.5
E^hE^m	0.01	0.36	1	0	0.63	0.72	0.17	1	1	150	3
E^hE^l	0.06	0.3	0.96	0	0.96	0.86	0.76	1	12	327	3.5
E^hU^h	0.5	0.87	1.02	0	0.79	0.78	0.97	1	1	40	3.5
E^hU^m	0	0.03	0.97	0	0.69	0.88	0.74	2	1; 13	170; 377	3
E^hU^l	0.01	0.03	0.96	0	0.6	0.64	0.55	1	1	118	3
E^hI	0.5	0.64	1.04	0	0.74	0.74	0.82	0	-1	-1	3.5
E^mE^h	0.5	0.1	0.95	0	0.78	0.96	0.99	0	-1	-1	3.5
E^mE^m	0.30	0.69	0.99	0	0.98	0.90	0.67	1	4	408	3.5
E^mE^l	0.06	0.14	0.98	0	0.87	0.97	0.99	2	1; 12	190; 204	3.5
E^mU^h	0.00	0.20	0.91	0	0.93	0.97	0.85	1	1	324	3.5
E^mU^m	0.49	0.37	0.89	0	0.98	0.96	0.96	0	-1	-1	3.5
E^mU^l	0.50	0.39	0.93	0	0.79	0.98	0.80	1	1	132	3.5
E^mI	0.30	0.96	1.05	0	0.87	0.78	0.70	1	14	234	3.5
E^lE^h	0.03	0.51	0.96	0	0.74	0.94	0.97	1	10	203	3.5
E^lE^m	0.5	0.53	0.92	0	0.83	1	0.96	0	-1	-1	3.5
E^lE^l	0.37	0.6	1.06	0	0.92	0.85	0.4	1	12	217	3.5
E^lU^h	0.01	0.03	0.86	0	0.96	0.93	0.96	4	1; 7; 4; 9	149; 371; 373; 400	3
E^lU^m	0.01	0.8	0.95	0	0.88	0.98	0.98	1	1	324	3.5
E^lU^l	0.5	0.55	1.19	0	0.85	0.97	0.62	0	-1	-1	3.5
E^lI	0.48	0.78	1.05	0	0.96	0.98	0.85	2	9; 9	95; 239	3.5

The first 2 columns give p-values for a Jarque-Bera and a Kolmogorov-Smirnov normality tests. The next column display the F-statistics of the test for equal variance before and after 1994. The column "reject" is equal to one when the null hypothesis of homoskedasticity is rejected. The next 3 column give the p-values of an LBQ-test at lag 1, 6 and 12. Finally, the last four columns display the number of outliers, their type (1 for AO, 4 for LS and 6 to 16 for Seasonal outliers), location (1 corresponds to January 1976) and the critical value used for the detection.

Table 22: Tests and Outliers (1)

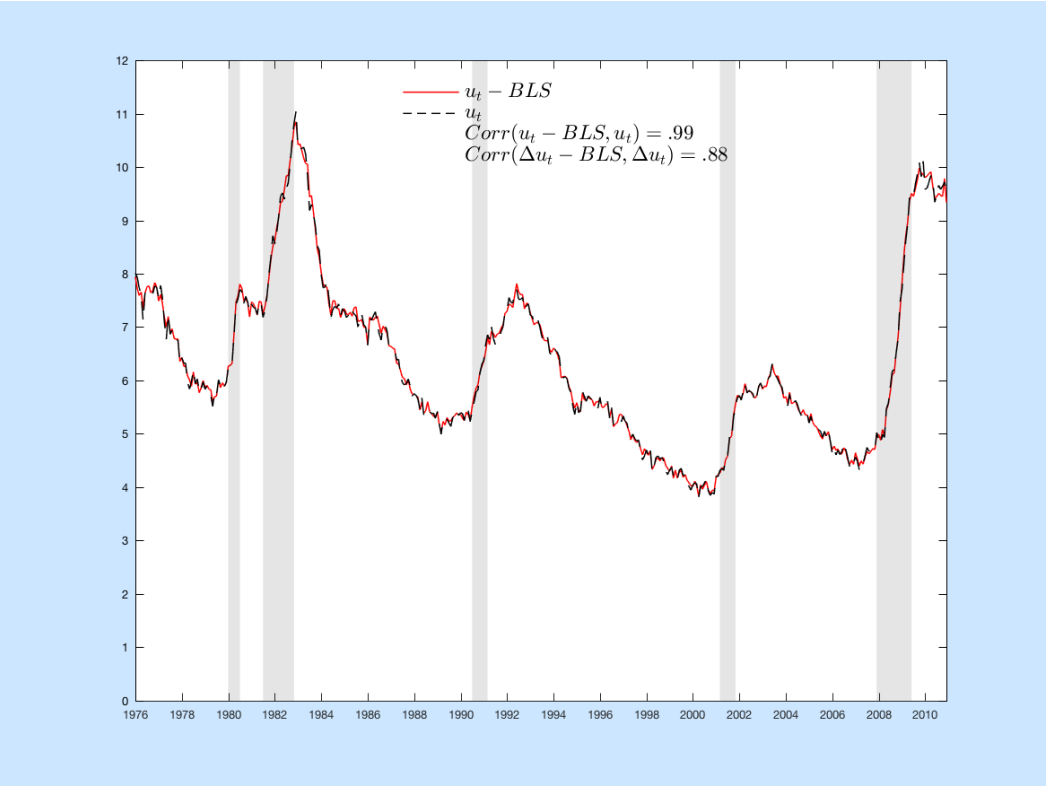
	Norm. tests		Het. test		LBQ. test			Outliers			
	JB	KS	F stat	Reject	lag1	lag6	lag12	Num.	Type	Location	C
Unemployment											
UE	0.16	0.78	0.92	0	0.61	0.30	0.40	0	-1	-1	4
UU	0.10	0.99	0.84	0	0.80	0.36	0.61	0	-1	-1	3.5
UI	0.34	0.24	1.07	0	0.78	0.42	0.38	0	-1	-1	3.75
U^hE^h	0.50	0.18	0.76	1	0.96	0.96	0.98	2	1; 11	39; 288	3.75
U^hE^m	0.44	0.10	0.96	0	0.79	0.86	0.84	1	4	401	3
U^hE^l	0.00	0.39	1.00	0	0.87	0.80	0.49	1	1	151	3
U^hU^h	0.50	0.49	0.98	0	0.77	0.96	0.55	0	-1	-1	3.5
U^hU^m	0.00	0.15	0.88	0	0.97	0.65	0.72	1	10	411	3.5
U^hU^l	0.00	0.44	0.90	0	0.76	0.99	0.48	1	1	381	3.5
U^hI	0.08	0.04	0.99	0	0.93	0.96	0.93	1	1	209	3.25
U^mE^h	0.00	0.33	1.10	0	0.83	0.77	0.93	1	1	179	3.5
U^mE^m	0.30	0.43	1.23	0	0.99	0.99	0.99	1	1	299	3.5
U^mE^l	0.03	0.54	1.18	0	0.85	0.99	0.96	1	14	381	3.5
U^mU^h	0.00	0.46	0.92	0	0.86	0.97	0.58	2	1; 8	183; 391	3
U^mU^m	0.16	0.47	0.92	0	0.68	0.38	0.40	0	-1	-1	3
U^mU^l	0.03	0.79	0.86	0	0.87	0.99	0.99	2	4; 12	395; 406	3.25
U^mI	0.04	0.26	1.08	0	0.90	0.86	0.69	2	1; 4	153; 391	3.25
U^lE^h	0.00	0.44	1.25	0	0.87	0.95	0.70	2	1; 4	176; 388	3.25
U^lE^m	0.02	0.27	1.01	0	0.98	.98	0.92	2	1; 9 3	12; 406	3.25
U^lE^l	0.18	0.16	0.84	0	0.74	0.86	0.64	2	11; 13	346; 393	3.5
U^lU^h	0.00	0.30	0.94	0	0.55	0.90	0.88	0	-1	-1	3.5
U^lU^m	0.07	0.11	0.97	0	0.99	0.99	0.98	2	1; 4	381; 403	3.5
U^lU^l	0.22	0.68	0.96	0	0.58	0.74	0.79	0	-1	-1	3.5
U^lI	0.50	0.50	1.14	0	0.74	0.94	0.75	2	4; 10	393; 285	3.5
Inactivity											
IE	0.27	0.64	0.88	0	0.96	0.75	0.72	4	11; 10; 11; 12	106; 255; 126; 383	3.5
IU	0.50	0.28	1.15	0	0.90	0.99	0.84	0	-1	-1	3.5
II	0.34	0.05	0.82	0	0.58	0.78	0.41	3	4; 4; 4	312; 313; 315	4
IE^h	0.50	0.01	0.79	1	0.74	0.71	0.96	2	11; 4	261; 265	3.5
IE^m	0.23	0.57	1.02	0	0.82	0.97	0.55	2	10; 11	137; 208	3.75
IE^l	0.24	1.00	0.97	0	0.96	0.98	0.92	2	8; 11	195; 256	3.75
IU^h	0.05	0.18	1.10	0	0.74	0.74	0.86	2	1; 11	154; 409	3.5
IU^m	0.50	0.38	1.17	0	0.93	0.84	0.96	2	1; 4	353; 390	3.5
IU^l	0.50	0.62	1.07	0	0.84	0.95	0.97	0	-1	-1	4

The first 2 columns give p-values for a Jarque-Bera and a Kolmogorov-Smirnov normality tests. The next column display the F-statistics of the test for equal variance before and after 1994. The column "reject" is equal to one when the null hypothesis of homoskedasticity is rejected. The next 3 column give the p-values of an LBQ-test at lag 1, 6 and 12. Finally, the last four columns display the number of outliers, their type their type (1 for AO, 4 for LS and 6 to 16 for Seasonal outliers), location (1 corresponds to January 1976) and the critical value used for the detection.

Table 23: Tests and Outliers (2)

A.2.7 Estimation Results: Comparison of seasonally adjusted unemployment rates

Figure 18 compares the seasonally adjusted unemployment rate series obtained using the framework presented in Section 3 with the official series released by the BLS. Given that the unemployment rate series used in this paper does not feature *New unemployed entrants*, I download the seasonally unadjusted unemployment rate series from the BLS website. Note that to deseasonalize time series, the BLS proceeds by deseasonalizing series by gender and age and then aggregates these seasonally adjusted series to obtain the aggregate unemployment rate series. In spite of these different methods, Figure 18 reveals that adjusted series are quite similar.



Seasonally adjusted series for the unemployment over the period 1976-2010 expressed in percentages. The official series released by the BLS,$u_t - BLS$, is displayed in red while the seasonally adjusted series obtained from the framework presented in Section 3 is shown as the dashed black line. This figure also gives the correlation between both series (both in level and first differenced).

Figure 18: Seasonal adjustment of the unemployment rate

A.3 Margin of Adjustment and Time Aggregation Corrections

A.3.1 Likelihood Function.

To derive the likelihood function, we should recall that a CTMC can be thought in terms of holding time and jump chain (see section 2.6 in Norris (1997)). The holding time τ in state i follows an exponential distribution with parameter $f_i \equiv -f_{ii}$. Once the holding time is over, an individual transition from state i to j with probability $\frac{f_{ij}}{f_i}$. Defining the time spent in the starting state i as τ_0, an observation is a sequence of state (the jump chain) $\{S_0, S_1, S_2, ...\}$ with $S_l \in \{E^h, E^m, E^l, ...\}$ and holding time $\{\tau_0, \tau_1, ...\}$. The likelihood contribution of this observation (using a subscript 1 to indicate the observation number) over the interval of time $[0, T]$ is given by :

$$\begin{aligned} L_1 &= f_i e^{-f_i \tau_{1,0}} \frac{f_{ij}}{f_i} f_j e^{-f_j \tau_{1,1}} \frac{f_{jk}}{f_j} ... \\ &= e^{-f_i \tau_{1,0}} f_{ij} e^{-f_j \tau_{1,1}} f_{jk} ... \\ &= \prod_{i=1}^{K} \prod_{j \neq i} e^{f_i R_{1,i}(T)} f_{ij}^{N_{1,ij}(T)} \end{aligned}$$

where K is the total number of states, $R_{1,i}(T) = \int_0^T \mathbf{1}(S_l = i)$ is a total time spent in state i by time T and $N_{1,ij}(T)$ is total number of transitions observed from state i to state j by time T. The joint

likelihood function is then obtained by taking the product of all individual contributions. Defining $R_i(T)$ and $N_{ij}(T)$ as the total amount of time spent in state i and the total number of transitions from state i to state j for all observations, we obtain expression (17) displayed in the main text.

A.3.2 CTMC Simulation.

Bladt and Sørensen (2005) propose to use Gibbs sampling to simulate the posterior distribution (19). This requires sampling from $P(X|\Theta)$ (the likelihood function) and Bladt and Sørensen (2005) propose to simulate continuous time Markov chain to reproduce the total transitions (gross flows) observed in the data from each labor market state. The following paragraph gives more details on this process.

The simulation of CTMC is done by first drawing a holding time from an exponential distribution with parameter f_i.[41] If this holding time is smaller than the length of a period, a transition from state i to j happens with probability $\frac{f_{ij}}{f_i}$.[42] This process is repeated until the length of the period is reached and the holding times and transitions are recorded.

In order to make this process clearer, let's assume that there are 100 individuals starting the month $t-1$ in high skill employment E^h. Among these 100 individuals, 95 are recorded to be in high skill employment in month t which corresponds to the E^hE^h gross flow. Bladt and Sørensen's procedure would then require simulating 95 individual CTMC where the starting and ending state would be E^h. This could happen in different ways. For instance a simulation in which the holding time drawn would be bigger than 1 would result in no simulated transition and therefore an E^hE^h flow. On the other hand, a simulated path $E^hU^hE^h$ would also be consistent with an E^hE^h gross flow. Once the 95 E^hE^h transitions have been obtained, all other simulations leading to this transition are rejected. The simulations of individual CTMC stops when all gross flows from all states have been reproduced. These simulation results allow to compute the total time spent in state i, $R_i(T)$ and the total number of transitions from state i to j, $N_{ij}(T)$ which are then used to draw new hazard rates from the posterior distribution (19).

The Time Aggregation correction is performed after the Margin of Adjustment one. The Margin of Adjustment corrections is applied to flow rates which implies that I do not have corrected gross flows required for CTMC simulations. I therefore recreate a labor market with 20000 individuals. I can then use the population stocks normalized by total population to obtain the starting states for each individual and the corrected flow rates to obtain the gross flows to be reproduced. Let's assume for instance that in January 1976, 20% of the population was in high skill employment ($e^h = 0.2$). This means that 4000 individuals start in high skill employment. If the E^hE^h flow rate corrected for Margin of adjustment is 0.95 ($\tilde{p}^{E^hE^h} = 0.95$), I would have to simulate 3800 CTMCs starting E^h and ending up E^h. Therefore, the requirement to compute gross flows is to assume a certain number of individuals. I do not have much evidence to pick this number. As a result, I ran the Bayesian estimation procedure for different number of individuals (10000, 20000 and 30000) and for February 1976. It turns out that this number has a limited effect on the results and 20000 was retained over 10000 as it insures that all non zero flow rates lead to at least 1 observed transition.[43]

Finally, note that this Bayesian estimation technique offers interesting aspects that were not

[41] This can be done using the *inverse sampling method*. We have to generate a random numbers from a uniform distribution in the interval $[0, 1]$, invert the exponential cumulative distribution function and evaluate the inverse using the randomly generated number from the uniform distribution.

[42] Note that the unit of time is assumed to be a month given that I work with monthly transitions and estimate monthly hazard rates. Therefore a transition happens if the holding time drawn is smaller than 1. See also Figure 19.

[43] High skill unemployment represents less 0.5% of total population on average. With 10000 individuals, there will only be around 50 individuals high skill unemployed and a transition rate smaller than 1% would then imply 0 observed transitions. This is often the case for $\tilde{p}^{U^hU^m}$ and $\tilde{p}^{U^hU^l}$. Increasing the number of individuals to 20000 ensure that at least 1 of these monthly transitions are reproduced.

exploited in this work. In particular, this estimation method can be used to reproduce how the CPS assign labor market status to respondent. Assume that 1 CTMC simulation for a given month t leads to the following transition:

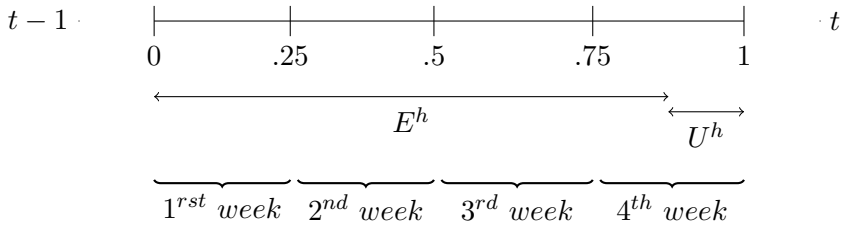
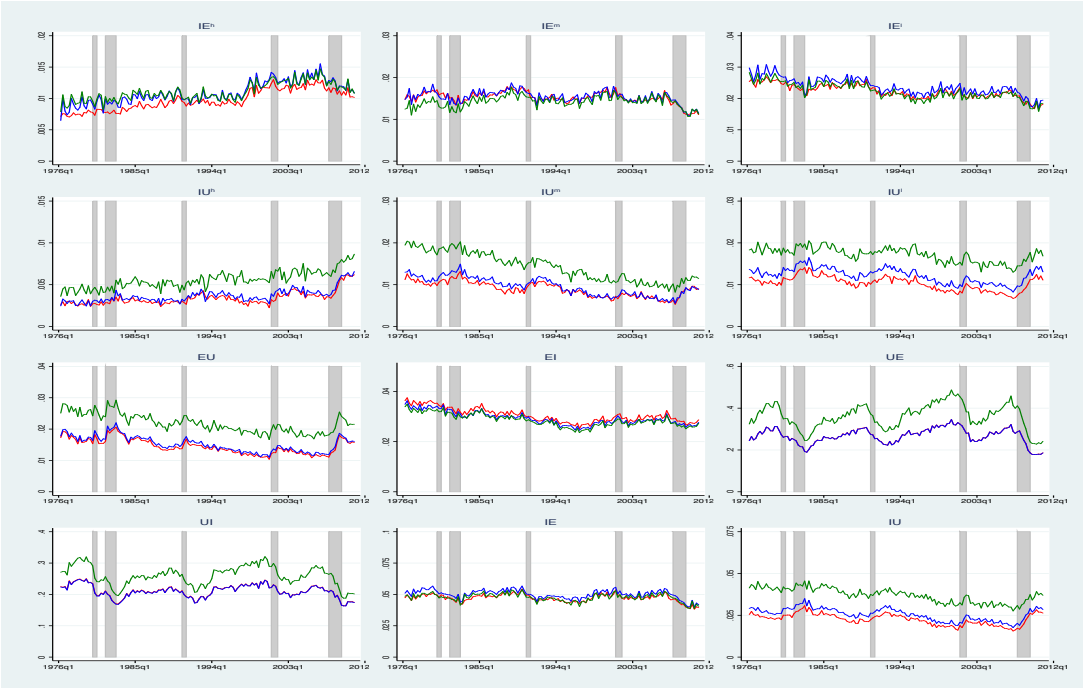


Figure 19: E^hU^h Simulated Transition

As displayed in the above figure, the transition to unemployment occurs during the last week of the month.[44] According to how the CPS measure stocks, the ending state recorded should be E^h and not U^h because some time in employment is observed in the last week. A similar example can be applied to a transition to inactivity in the last week of the month.

The Bayesian estimation procedure, through the simulations of CTMCs, offers the possibility to reproduce the labor market state assignment used by the CPS. As mentioned by Elsby et al. (2015), accounting for this dynamic assignment could matter for the estimation of hazard rates.

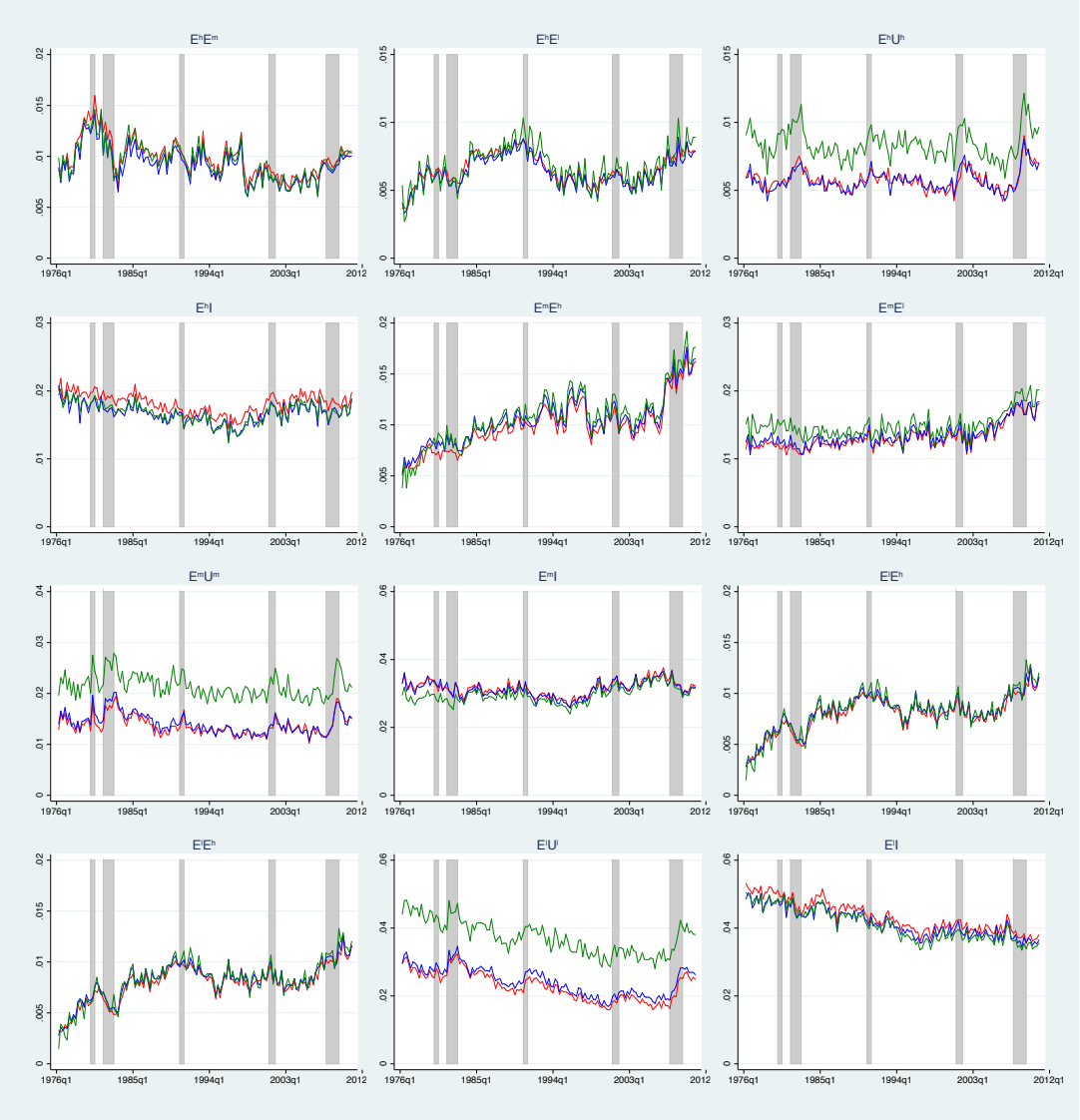
A.3.3 Corrected Flow Rates: Aggregates, from Employment and from Inactivity



Quarterly flow rates. The corrected flow rates from Section 3 are displayed in blue, the series corrected for Margin of adjustment is displayed in red and the Time Aggregation corrected series is plotted in green.

Figure 20: Flow rates corrected for Margin of Adjustment and Time Aggregation (2).

[44] Note that actually, the interview usually takes place the week of the 12th such that the end of a period should be when the interviews have taken place rather than the end of the month.



Quarterly flow rates. The corrected flow rates from Section 3 are displayed in blue, the series corrected for Margin of adjustment is displayed in red and the Time Aggregation corrected series is plotted in green.

Figure 21: Flow rates corrected for Margin of Adjustment and Time Aggregation (3).

INSTITUT DE RECHERCHE
ÉCONOMIQUES ET SOCIALES

Place Montesquieu 3
1348 Louvain-la-Neuve

ISSN 1379-244X D/2020/3082/08

UCLouvain

IRES | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics