

# Teknisk dokumentation af data til lønanalyser

Af Rikke Ibsen og Niels Westergård-Nielsen  
Center for Corporate Performance  
Aarhus School of Business  
Aarhus Universitet

Århus juni 2010

<b>1</b>	<b>Indledning</b>	<b>3</b>
<b>2</b>	<b>Validering Lønstatistikens Serviceregister</b>	<b>3</b>
2.1	<i>Sektorer</i>	3
2.2	<i>Omsorgsdage i staten</i>	3
2.3	<i>Timefortjeneste</i>	4
2.4	<i>Validering af uddannelse</i>	6
<b>3</b>	<b>Sammenkøring af lønstatistikken med FLD-data og Personalestyrelsens data</b>	<b>7</b>
3.1	<i>Merge mellem Lønstatistikken og FLD.</i>	7
3.1.1	<i>1-1 match på alle nøglevariable</i>	7
3.1.2	<i>Match på 5-cifret lønklasse</i>	7
3.1.3	<i>Manglende CPR i FLD.</i>	7
3.1.4	<i>De resterende observationer</i>	7
3.1.5	<i>Ekstra stk-kode</i>	8
3.2	<i>Merge af lønstatistikken og Personalestyrelsens data</i>	8

## 1 Indledning

Datagrundlaget for lønanalyserne i Lønkommissionenes rapport er Lønstatistikens Serviceregister som er sammenkørt med en række andre registre<sup>1</sup> fra Danmarks Statistik, Det Fælleskommunale Lønkontor (FLD) og Personalestyrelsens løndata.

Danmarks Statistiks dokumentation af Lønstatistikens Serviceregister findes på Danmarks Statistiks hjemmeside og i kapitlet om lønbegreber i Lønkommissionens rapport. Der har dog til brug for lønanalyserne været nødvendigt at foretage yderligere validering og forbedringer af data, og disse er beskrevet i dette notat.

Lønanalyserne er baseret på grupper indmeldt af parterne. Disse grupper er afgrænset på stillingskoderne i FLD og Personalestyrelsens data frem for Disco-koder, fordi det er de grupper man forhandler overenskomst for. Lønstatistikens Serviceregister har kun stillingsoplysninger på Disco-koder, så derfor er Lønstatistikken Serviceregister sammenkørt med data fra FLD og Personalestyrelsens løndata. Denne sammenkøring fører til et bortfald af observationer, som også er beskrevet i dette notat.

## 2 Validering Lønstatistikens Serviceregister

### 2.1 Sektorer

Som udgangspunkt, er der ikke problemer med at opdele i de 3 sektorer

- privat
- stat
- amt /region og kommune

men opdelingen af amt/region og kommune skal laves særskilt.

Til dette formål bruges den såkaldte funktionskode. Funktionskoden er dog missing i en del observationer, og derfor bruges en særlig variabel konstrueret af DST (hvor der manuelt er tildelt funktionskoder) til at eftervalidere opdelingen.

Ud over denne validering, bruges månedsdata fra FLD, hvor variabelen "forhomr" (forhandlingsområde), merges på via senr, og dermed yderligere giver mulighed for opdeling på amt/region og kommune.

De manuelle funktionskoder findes tilbage til 2003, så derfor er 2002 korrigeret med 2003 data. Ligeledes findes FLD-data tilbage til 2004, så 2002 og 2003 er korrigeret med 2004-data.

Observationer, der ikke kan henføres til amt/region eller kommune slettes.

### 2.2 Omsorgsdage i staten

Omsorgsdage i staten er ikke registreret, og derfor er den præsterede timefortjeneste i staten for lav i perioden. Fra 2005 til 2007 gives der 2 omsorgsdage pr. år. pr. barn 0-7 år, og derfor kan der vha. antal børn og deres alder tildeles et antal omsorgsdage til en ansættelse. Der

---

<sup>1</sup> Uddannelsesregistret, Ida (den integrerede arbejdsmarkeds database), sygedagpengeregistret, kursusregistret (som kun er brugt i de gruppevis regressioner).

## 4 | Teknisk dokumentation af data til lønanalyser

tildeles omsorgsdage uanset, om de er afholdt eller ej, da vi ikke ved, om omsorgsdagene er afholdt, kun om der er ret til omsorgsdage.

I perioden 2002-2004, tildeles omsorgsdage kun for personer, som får børn, mens de er ansat i staten og afholdelsen af omsorgsdage ligger frit indtil barnet er 14 år. Derfor er det ikke muligt at lave en fornuftig korrektion af omsorgsdage før 2005.

Hvis der er flere ansættelser i løbet af året, overlappende ansættelser eller ansættelser, som ikke varer hele året, tildeles omsorgsdagene så omfanget af ansættelsen er afgørende. Derfor kan en person aldrig have flere omsorgsdage end hvad der svarer til årets samlede ansættelser.

Ud over den præsterede timefortjeneste korrigeres fraværet samt alle løndele for den præsterede timefortjeneste.

### 2.3 Timefortjeneste

Der er lavet en validering af timefortjenesten ud over den validering, der er lavet af Danmarks Statistik.

Valideringen laves ud fra fortjeneste pr. præsteret time excl. gene og derefter for fortjeneste pr. præsteret time incl. gene og til sidst for den standardberegnete timefortjeneste.

Der er generelt et problem for den standardberegnete timefortjeneste for de ansættelser, som ikke er helårs. Det skyldes, at perioderegistreringerne ikke er valide i en række tilfælde, og det betyder igen, at timefortjenesterne bliver for lave. Dette problem må antages at være generelt for alle timefortjenester, men problemet er kun muligt at se i data for de lønninger, som er åbenbart for lave.

I den offentlige sektor vil der ikke være nogen, som får en løn under mindstelønnen, men da de lave lønninger, der opstår som følge af fejlagtige perioderegistreringer, findes for alle lønkategorier, fjernes de 5% nederste lønninger fordelt på 6-cifret discokode og sektorer (privat, stat, region, kommune). Denne metode er bedre end blot at fjerne lønninger under mindstelønnen, da de nederste lønninger for f.eks. folkeskolelærere er væsentlig højere end mindstelønnen i den kommunale sektor. Ved at fjerne de nederste 5%, skabes en ensartet validering for alle discokoder og sektorer.

Da der vil være forskel i antallet af observationer fordelt på discokode, vil der også være tilfælde, hvor der selv efter fjernelse af de nederste 5%, er meget lave lønninger. Det drejer sig om 5121 observationer, hvor timefortjenesten er mindre end 80% af industriens mindsteløn eller mindre end den offentlige mindsteløn (2007-data).

De er i 2007 fordelt med

- 154 observationer i amt/region (0,1% af observationer i amt/region)
- 2339 observationer i kommunen (0,4% af observationer i kommunen)
- 33 observationer i staten (0,01% af observationer i staten)
- 2595 observationer privat (0,2% af observationer i den private sektor)

Disse observationer fjernes fra data.

Valideringen består af følgende trin

- Derefter fjernes de nederste 5% for hver discokode og sektor
- Derefter fjernes alle observationer, hvor personen er under uddannelse
- Derefter fjernes observationer med lønninger under 80% af industriens mindsteløn og under den offentlige mindsteløn
- Efter korrektion for omsorgsdage i staten fjernes outliers i toppen af lønningerne.
- Valideringen følger DST's valideringsgrænser for årene 2002-2004.
- Der er forskel på, hvorledes der er valideret i den offentlige sektor i 2007 og årene før. I perioden 2002-2006 er fortjeneste pr. præsteret time uden gene over 1000 kr fjernet fra data. I 2007 er der taget hensyn til disco-koden, da der er stor forskel på fraværet indenfor forskellige discokoder.
- DST har lavet en beskrivelse af valideringen for fortjenesten pr. præsteret time uden gene.
- Ud over valideringen af fortjeneste pr. præsteret time excl. gene, er fortjeneste pr. præsteret incl. gene over 3000 kr fjernet.
- Valideringen afsluttes med at fjerne observationer, hvor den standardberegnete timeløn i den offentlige sektor overstiger den højeste løn, der kan gives i de 3 offentlige sektorer, samt en fastsat grænse i den private sektor. Grænserne er angivet i tabel 1.

Tabel 1: Øvre valideringsgrænser for standardberegnet timefortjeneste

	Region/ Stat kommune		Privat
2002	900 kr	750 kr	2440 kr
2003	900 kr	800 kr	2500 kr
2004	950 kr	850 kr	2550 kr
2005	1000 kr	850 kr	2600 kr
2006	1000 kr	940 kr	2650 kr
2007	1200 kr	1000 kr	2700 kr

Kilde: Kilde: Egne beregninger på Lønstatistikens Serviceregister

### Yderligere validering

Yderligere validering omfatter

- begrænsning af populationen til arbejdsstyrken (16-64 år)
- observationer i det offentlige, som kan henføres til Grønland og Færøerne eller offentlig arbejdsplads udenfor Danmark slettes
- præster optræder i den private sektor i stedet for staten, så de flyttes til staten
- undervisning på erhvervsskoler (8346 observationer) og arbejdsmarkedsuddannelser (905 observationer) har ingen sektorkode. De hører til i staten, så de får statens sektorkode.

Fra de oprindelige rådata til de fuldt validerede er der følgende andel af observationerne tilbage, som det fremgår af tabel 2.

Tabel 2: Andel af data efter validering

	Privat %	Stat %	Region %
2002	81,53	87,20	86,53
2003	83,09	87,26	87,59
2004	82,79	87,91	87,33
2005	82,57	87,44	86,37
2006	81,80	86,96	84,97
2007	82,96	86,62	85,03

Kilde: Egne beregninger på Lønstatistikens Serviceregister

## 2.4 Validering af uddannelse

Lønstatistikens Serviceregister har et ubegrænset antal ansættelser på år pr person, men kun én oplysning om uddannelse pr. år. Derfor vil der være personer, som færdiggør en uddannelse i løbet af året, men som i data ikke vil være registreret som havende den pågældende uddannelse i de job, som ligger efter færdiggørelsestidspunktet. Derfor er der lavet en korrektion af uddannelsesvariablen ved hjælp af oplysninger om uddannelse og færdiggørelsestidspunkter fra uddannelsesregistret året før, i året og året efter.

Uddannelse er korrigeret i flere trin ved hjælp af uddannelsesregistret

- Når en uddannelse er afsluttet i løbet af året og ansættelsen ligger efter afslutningstidspunktet får observationen den nye uddannelse
- Hvis uddannelse er missing i lønstatistikken, men ikke missing i uddannelsesregistret, får observationen uddannelsen fra uddannelsesregistret. Afhængigt af afslutningstidspunktet og ansættelsesperioden giver uddannelse fra 2007 eller 2008.
- Hvis ansættelsesperioden efter uddannelsestidspunktet er mere end dobbelt så lang som før, og uddannelsen fra uddannelsesregistret er højere end lønstatistikken, gives uddannelsen fra uddannelsesregistret.
- Uddannelser tilrettes via discokoden, da f.eks. en discokode 2, og en person, som afslutter en LVU i ansættelsesperioden får uddannelsen fra uddannelsesregistret. Derimod vil en discokode 9 og en person, som afslutter i ansættelsesperioden *ikke* få uddannelsen fra uddannelsesregistret, men den uddannelse, den har i lønstatistikken. Og så fremdeles detaljeret på discokoder og ansættelser.

### 3 Sammenkøring af lønstatistikken med FLD-data og Personalestyrelsens data

#### 3.1 Merge mellem Lønstatistikken og FLD.

De 2 datasæt matches med:

- CPR
- Brugernummer
- Medarbejdernummer
- Discokode
- Lønklasse
- Leverance

I lønstatistikken er der 821.434 observationer med sektorkode=3 (Amt/kommune)

Det færdige match er på over 99%. Beskrivelse af matchet er gennemgået i det følgende.

##### 3.1.1 1-1 match på alle nøglevariable

FLD-data er leveret i 2 omgange:

Første levering gav et match på alle nøglevariable for 684.226 observationer, svarende til 83%.

De resterende 17% (137.208) observationer kunne ikke genfindes i de leverede FLD-data på CPR-nummer, hvilket betød, at personerne ikke eksisterede i FLD-data. Der blev efterfølgende leveret et nyt datasæt fra FLD, som blev brugt til yderligere match.

Match mellem de resterende 17% af observationerne i og de nye FLD-data på alle nøglevariable gav yderligere match for 93.235 observationer. Det bragte 1-1 matches på alle nøglevariable op på 777.461 observationer, svarende til 95%.

##### 3.1.2 Match på 5-cifret lønklasse

De sidste 43.973 observationer kan ikke matches 1-1 på alle nøglevariable, så efter at have testet flere muligheder laves et match, hvor lønklassen matches på de første 5 cifre og de andre nøglevariable matches som før.

Det giver et match for 14.754 observationer. Det samlede match er herefter oppe på 792.215 observationer.

Samme laves for samtlige data fra FLD (gammel og ny leverance), i stedet for bare den nye leverance, og der kan matches 6339 observationer. Matchet er så 798.554, 97%.

##### 3.1.3 Manglende CPR i FLD.

Efter den anden leverance fra FLD, er der stadig 14.157 observationer i lønstatistikken, som ikke kan genfindes i FLD via CPR. Næsten alle disse kan via discokoden identificeres som gymnasielærere mm. som i forbindelse med kommunalreformen er flyttet fra amterne til staten, og derfor ikke burde have sektorkode 3. Disse kan derfor uden problemer fjernes fra populationen.

Det betyder, at det samlede antal observationer i lønstatistikken bliver 807.277, og det samlede match med FLD er oppe på 99%.

##### 3.1.4 De resterende observationer

Der er nu 6723 observationer i lønstatistikken, som ikke har et match.

Der er en overvægt af discokoder, som ligger inden for sundhed og omsorg og pædagogiske områder, så derfor laves et manuelt match baseret på lønklasser. Som udgangspunkt merges på alle nøglevariable undtagen lønklassen, og derefter laves en manuel kategorisering af lønklasser i sundhedsområdet og det pædagogiske område samt socialrådgivere.

Dette manuelle match giver et match for 2780 observationer.

De resterende data droppes fra data.

### 3.1.5 Ekstra stk-kode

Der er under matchet opstået redundans, da samme ansættelse i FLD kan have flere stk-koder. Dette er løst ved, at den ekstra stk-kode har fået sin egen variabel, så der ikke er gået information tabt. På denne måde kan stk-koder for grupper med sikkerhed findes i den første eller ekstra stk-kode.

## 3.2 Merge af lønstatistikken og Personalestyrelsens data

For at få personalekategorierne fra personalestyrelsens data merges disse variable på lønstatistikken.

Personalestyrelsens data omfatter kun ansættelser i 4. Kvartal, og der merges derfor på ansættelser i lønstatistikken i 4. kvartal.

Der er 129.856 statslige ansættelser i lønstatistikken i 2007.

Der merges på følgende variable:

- Pnr (anonymiseret cpr-nummer)
- Medarbejdernummer
- Leverance
- Pkat\_stat (personalekategori)
- Stiko\_stat (stillingskategori)

For at få yderligere match inddrages discokoden, så der merges på:

- pnr
- medarbejdernummer
- leverance
- discokode
- og ENTEN
  - o pkat\_stat ELLER
  - o stiko\_stat

Af disse får 126.414 obs et merge. 2214 har samme observationer fra personalestyrelsens data på flere ansættelsesforhold.

Merget er på 97,3%. 3.442 observationer får ikke et match.