Statistics Norway
Statistisk sentralbyrå

Ole Villund

Yrke i Arbeidstakerregisteret

Notater

Seksjon for arbeidsmarkedsstatistikk
Emnegruppe: 06.01

Innhold

1 Innledning

1.1 Målgruppe

Dette notatet er i hovedsak teknisk dokumentasjon til internt bruk og er sammensatt av ulike deler:

- dokumentasjon av systemet for automatisk yrkeskoding i Statistisk sentralbyrå.
- kvalitetskontroll av yrkesdata i Arbeidstakerregisteret.
- imputerings- og estimeringsmetoder.

Målgruppen er brukere av kodesystemet og yrkesdata i Statistisk sentralbyrå, samt andre interesserte. Det kan også ha interesse i forbindelse med utvikling av automatisk klassifikasjon av tekstdata.

1.2 Historisk bakgrunn

Siden folketellingen i 1801 har yrke vært et sentralt spørsmål. Inndelingen i FoB1990 er basert på Standard for yrkesklassifisering (NYK) som ble utarbeidet i 1965, men med en revisjon og utvidelse av yrkeskatalogen i forbindelse med Folke- og boligtellingen i 1980. Yrkesstandarden fra 1965 er bransjeorientert. Det er publisert flere typer tabeller om yrke i FoB90, se NOS om Folke- og boligtelling 1990 - Dokumentasjon og hovedtall, der en både bruker 2 og 3 siffer inndeling i NYK.

I motsetning til i 1990 er yrkesklassifiseringen for de yrkesaktive i FoB2001 basert på registerinformasjon. Den betydeligste innsatsen for å nå målet om yrkesopplysninger gjennom bruk av registrene i FoB2001 er skjedd gjennom at Arbeidstakerregistret også skal inneholde opplysning om yrke. Opprettelsen av variabelen yrke er gjennomført ved å hente inn opplysningene i forbindelse med årskontrollen for 2001 der bedriftene skulle påføre yrke. Den løpende oppdateringen av yrke i Arbeidstakerregistret vil skje gjennom innmeldings- og utmeldingsblanketten for nye arbeidstakere der yrke er et eget felt som skal fylles ut. Arbeidsgivere som leverer papirskjema skal sende løpende meldinger når arbeidstakere bytter yrke i samme bedrift, men det er usikkert i hvilket utstrekning dette skjer. Det betyr at også årskontrollen vil være et viktig verktøy for å holde en årlig oppdatert yrkesfordeling fra registret. Arbeidsgivere med elektronisk innrapportering skal melde endringer hver måned.

I tillegg må en også beregne yrke for sysselsatte utenfor arbeidstakerregistret, dette gjelder først og fremst fra Lønns- og trekkoppgaveregistret og selvstendig næringsdrivende. For disse to gruppene vil det ikke foreligge yrkestittel som tekst eller kode, med unntak av de som kan kobles mot AKU og diverse registre. Vi må i mange tilfeller ta utgangspunkt i egenskaper ved foretaket og personen som gir en mulighet til å beregne yrket. De to viktigste egenskapene vil være næring til foretak/bedrift og utdanningen til den yrkesaktive.

I 1998 ble en ny Standard for yrkesklassifisering (STYRK) tatt i bruk. STYRK er avledet av International Standard Classification of Occupations (ISCO-88). Denne standarden bygger på to prinsipper: kompetansenivå og spesialisering

Siden kompetansenivå formelt sett er knyttet til lengden på utdanning, se avsnitt 2, er STYRK nærmere bundet til utdanning enn NYK som vi tidligere nevnte var nært koblet til næring. Dette kan også formuleres slik at NYK i større grad var forbundet med bedriften, mens STYRK mer er forbundet med arbeidstakeren. Men det er uansett arbeidsoppgavene som skal være avgjørende for yrkesklassifiseringen. Altså er den kompetansen som normalt kreves for å utføre arbeidsoppgavene som skal legges til grunn for klassifiseringen, uavhengig av den enkeltes formelle kompetansenivå.

1.3 Yrkesklassifisering

1.3.1 Yrkesstandarden

Fra forordet til standarden : "Standard for yrkesklassifisering STYRK 98 (NOS C521) Standarden er basert på den internasjonale standarden for yrkesklassifisering som ILO har utarbeidet (International Standard Classification of Occupations - ISCO 88). EU har utarbeidet en versjon av denne (ISCO-88(COM)) som danner grunnlaget for den norske standarden. Standarden skal erstatte Nordisk Yrkesklassifisering (NYK). Standard for yrkesklassifisering er beregnet for bruk i offisiell norsk statistikk og vil også være et viktig verktøy innen arbeidsformidling og yrkesrettledning. Det første hovedbruksområdet er alle typer statistikk. På dette området er det spesielt viktig å ha klassifiseringsverktøy som er internasjonalt sammenlignbart og sikrer nasjonal sammenlignbarhet mellom ulike statistikker. Det andre hovedbruksområdet ligger innenfor arbeidsformidling. I arbeidsmarkedsetatens arbeid med å kople ledige stillinger mot arbeidssøkere er et godt yrkesklassifiseringssystem et helt nødvendig verktøy. For arbeidsmarkedsetaten er det spesielt viktig at standarden inneholder oppdaterte yrkesinndelinger og beskrivelser. Standarden bør være mest mulig knyttet opp til forholdene i dagens arbeidsliv, og den bør være så fleksibel at den lett kan tilpasses endringer. Under arbeidet med den norske standarden er begge typer brukerbehov tatt hensyn til. Ettersom ISCO-88(COM) i utgangspunktet er utviklet som en statistisk standard, vil nok fortsatt de statistiske behovene være best ivaretatt. Arbeidsmarkedsetaten vil arbeide videre med en inndeling av yrker for å dekke sine behov knyttet til formidlingsarbeidet. Det betyr at noen yrker vil kunne deles inn finere, på 5-6 siffernivå, andre yrker grovere, på 2-3 siffernivå. Inndelingen vil imidlertid følge ISCO-88(COM). Arbeidet med den norske standarden har vært organisert gjennom ei arbeidsgruppe med representanter fra Arbeidsdirektoratet og Statistisk sentralbyrå. Arbeidsgruppa la fram et utkast til standard som ble sendt ut på høring høsten 1997. Høringsinstanser var en rekke sentrale brukere og bransjeorganisasjoner."

STYRK er altså utarbeidet for å klassifisere informasjon om yrke i forbindelse med statistikk. Yrkeskoden består av 4 siffer og er hierarkisk oppbygd. Standarden inneholder 353 yrker på 4-siffer nivå. Statistisk sentralbyrås standard bygger på Eurostats versjon av ILOs yrkesstandard ISCO-88, og samsvarer i hovedsak med denne i de 3 første siffer (yrkesgruppe). I standarden er yrke definert *utfra likheter i arbeidsoppgaver*, som hovedregel uavhengig av egenskaper ved arbeidstaker, f.eks. utdanning, eller egenskaper ved bedriften, f.eks. næring.

1.3.2 Yrkeskatalogen

Statistisk sentralbyrå utgir Yrkeskatalogen (Håndbok 2001/72) som kan brukes av arbeidsgivere og andre for å finne riktig yrkeskode. Yrkeskatalogen inneholder yrkestitler, som er en mer detaljert inndeling enn yrke. Yrkestitler i samme yrke omfatter både ulike skrivemåter og mindre variasjoner i arbeidsoppgaver. Hver yrkestittel har den 4-sifrede yrkeskoden fra Yrkesstandarden og et 3-sifret løpenummer. Nye titler får et løpenummeret etter hvert som de opprettes, og inneholder ingen strukturell informasjon. Hvis en tittel slettes, blir ikke dette løpenummeret benyttet igjen. Løpenummeret er 100 for yrkeskoden og 101-999 for de enkelte yrkestitler. Katalogen inneholder for tiden over 5000 yrkestitler. I tillegg til rene synonymtitler er det også andre forhold som gjør seg gjeldene ved inndelingen i ulike titler innen hvert yrke:

- En del ord som brukes i yrkestitler angir ikke yrkets faglige innhold, men er en betegnelse på stilling eller status, f.eks. lærling, formann. Tømrerlærling og tømrermester har lik yrkeskode, men ulikt løpenummer.
- Samme hovedtittel brukes om yrker som kodes forskjellig,
  f.eks. sjåfør: taxisjåfør, renovasjonssjåfør skal ha ulike yrkeskoder.
- En del yrkestitler i Yrkeskatalogen har med andre kjennemerker som næring, sektor, bedriftsstørrelse eller utdanning, fordi de antas å være indikatorer på arbeidsoppgaver.
- I spesielle tilfeller skal yrker med ulike arbeidsoppgaver ha samme yrkeskode, f.eks. militære yrker: flyver og feltprest kan begge få yrkeskode 0112102.

Næring og andre opplysninger brukes for å skjelne ulike yrkestitler innen hvert yrke, men løpenummer er ikke strukturert utfra disse opplysningene. Inndelingen som brukes for næring korresponderer i noen tilfeller ikke med Standard for næringsgruppering (NOS C182), dels fordi man har lagt vekt på å holde seg til internasjonale yrkesinndelinger og dels fordi andre forhold som utdanning, sektortilhørighet og bedriftens størrelse er avgjørende. En del titler formuleres derfor i katalogen på en måte som man ikke kan forventes brukt ved rapportering. Mange av de vanligste titlene som brukes av arbeidsgivere er problematiske på den måten at de ikke er entydige i forhold til yrkesstandarden og yrkeskatalogen. Disse forhold fører til at man for programmeringsformål utarbeider spesielle lister utfra katalogen, og lager egne programrutiner for visse yrker. Yrkeskoden og løpenummeret brukes som regel sammen som en 7-sifret kode, både ved innrapportering til RTV og koding i Statistisk sentralbyrå. I dette notatet brukes "yrkeskode" om 7-sifret kode.

Litt nærmere om STYRK. I forhold til NYK (1965). La oss begynne med den groveste inndelingen som kan sammenliknes med den groveste inndelingen i NYK. Utdanning er mer synlig i flere av gruppene, enn tilfellet var i NYK:

1. Administrative ledere og politikere (lederyrker)

2. Akademiske yrker
3. Yrker med kortere høyskole- og universitetsutdanning og teknikere (høgskoleyrker)
4. Kontor og kundeserviceyrker (kontoryrker)
5. Salgs, service- og omsorgsyrker (salgs- og serviceyrker)
6. Yrker innen jordbruk, skogbruk og fiske (bønder, fiskere o.l.)
7. Håndverkere og liknende (håndverkere)
8. Prosess- og maskinoperatører, transportarbeidere mv. (operatører, sjåfører o.l.)
9. Yrker uten krav til utdanning (andre yrker)
0. Militære yrker og uoppgitt (andre yrker)

Oppbyggingen av standarden er hierarkisk gjennom fire nivåer:

1. siffer i koden: 10 yrkesfelt
2. siffer i koden: 31 yrkesområder
3. siffer i koden: 108 yrkesgrupper
4. siffer i koden: 353 yrker

Et eksempel på et yrke der det er svært nær sammenheng mellom utdanning og yrke kan da være:

Yrkesfelt	2	Akademiske yrker
Yrkesområde	22	Biologiske og medisinske yrker
Yrkesgruppe	222	Medisinske yrker
Yrke	2222	Tannleger

For å utføre yrket tannlege må en ha offentlig godkjenning, noe som krever utdanning innen faget, men ikke alle med utdanningen tannlege vil yrkesklassifiseres som tannlege. Det krever at man utfører arbeidsoppgaver som tilhører faget, altså jobber som tannlege. For mange andre yrker vil det ikke være en så enkel relasjon mellom utdanning og yrke.

Standarden deler inn i fire kompetanse nivåer:

- Yrker som ikke krever mer enn maksimum 9-årig grunnskole (felt 5-9, og felt 1 !)
- Yrker som normalt krever 1-3 års utdanning på videregående skoles nivå (felt 4)
- Yrker som normalt krever 1-3 års utdanning utover videregående skole (felt 3)
- Yrker som normalt krever universitets- eller høyskoleutdanning med varighet 4 år eller mer (felt 2)

Spesialisering gitt ved:

- Kunnskaper/ferdigheter
- Verktøy/maskiner
- Materialer
- Varer/tjenester
- Selvstendighet
- Rutinepreg
- Direkte fysisk arbeid (kroppsarbeid)

For å kunne gjøre STYRK operativ er det utarbeidet en yrkeskatalog (håndbøker 72/2001) der det for hvert yrke er opprettet et varierende antall yrkestitler. Arbeidsgiverne bruker denne når de rapporterer inn yrke til Arbeidstakerregistret og aller helst med den tilhørende sjusifrede koden der de fire første sifrene er yrket, mens tittelen har et tilhørende løpenummer. Det er imidlertid i instruksen fra Rikstrygdeverket gitt mulighet til å gi en mer beskrivende tekst som forklarer yrket til arbeidstakeren. Det er opplagt slik at dersom alle arbeidsgiverne fulgte yrkeskatalogen og leverte kun de yrkestitler som står der eller den sjusifrete koden ville det forenklet arbeidet med å få på plass yrke for arbeidstakerne. I praksis har det imidlertid vist seg at svært mange opplysninger må tolkes gjennom automatisk tekstbasert koding, og manuell koding. Nå er det ikke slik at en umiddelbart kan fastslå at det indirekte svaret fra arbeidsgiveren, dvs. en tekststreng som må tolkes og kodes til et yrke, gir noe dårligere kvalitet enn et direkte svar, dvs. en levert yrkeskode. Grunnen er rett og slett at når arbeidsgiveren svarer med en beskrivende tekst kan dette gi et korrekt yrke gjennom de kodemetoder som brukes. Det direkte svaret (levert yrkeskode) kan være feil uten mulighet til å se dette siden koden er gyldig.

2 Automatisk koding av yrke i Arbeidstakerregisteret

2.1.1 Yrkestittel i Arbeidstakerregisteret, AA

Statistisk sentralbyrå mottar jevnlig data fra Arbeidstakerregisteret til flere statistikkformål. Arbeidstakerregisteret administreres av Rikstrygdeverket (RTV) og skal inneholde yrkesdata for alle arbeidstakere fra og med 2001. Statistisk sentralbyrå koder yrke i henhold til Standard for yrkesklassifisering, for statistikkformål og tilbakemelding til arbeidsgiver for kontroll. I 2000 var det frivillig rapportering av yrke, og dette har vært grunnlag for utprøving av rutiner for yrkeskoding. Fra og med årskontrollen 2001 er det obligatorisk med yrke for alle arbeidstakerforhold. Bedrifter med elektronisk innlevering skal levere yrkeskode for hvert arbeidstakerforhold, andre kan føre inn yrkestittel som tekst eller skrive yrkeskode. Arbeidsgivere kan finne riktig yrkeskode i Yrkeskatalogen som papirversjon eller på Internet, og få hjelp ved henvendelse til Statistisk sentralbyrå. De som benytter stillingskoder som brukes i offentlig sektor kan oppgi disse istedenfor yrkeskode.

2.1.2 Arbeidskraftundersøkelsen, AKU

AKU har i flere år registrert yrkestittel og arbeidsoppgaver for intervjuobjektene. I AKU kodes yrke med 4 siffer utfra disse to opplysninger, og til dels andre opplysninger om intervjuobjektet og bedriften. Datamateriale fra AKU med yrkeskoder er brukt i arbeidet med å lage et system for automatisk koding. Analyser av AKU-materiale er benyttet ved prioritering i utarbeidelse av egne programdeler for særskilte yrker, og til utprøving av det automatiske kodesystemet. Det er også laget samsvarsanalyser av yrkeskodingen i AKU og AA, og andre metoder for å kontrollere kvalitet i systemet for yrkeskoding.

2.2 Tekstbasert koding

Arbeidstakerregisteret inneholder for tiden rundt 2.3 millioner arbeidstakerforhold. Av hensyn til en effektiv og konsekvent koding skal flest mulig arbeidstakerforhold kodes maskinelt. De resterende arbeidstakerforhold vil bli kodet av personell ved Seksjon for Datafangst (s450) med ekspertise og erfaring fra yrkeskoding i AKU. Det er utarbeidet en rutine for automatisk tekstbasert koding. Systemet består av programmer skrevet i SAS/BASE som tilordner en yrkeskode til hvert arbeidstakerforhold, og som skal kjøres hver uke. Innkomne data inneholder endringsmeldinger og årskontroll i Arbeidstakerregisteret, samt støtteinformasjon fra andre registre.

2.2.1.1 Inndata

Det foreligger filer fra RTV hver uke, og disse er tilgjengelig for yrkeskoding kort tid etter. Filene inneholder alle arbeidstakerforhold for de personer som det er innrapportert endringer på, pluss de arbeidstakerforhold som det ikke er endringer i. Filene inneholder en rekke arbeidstakerforhold som ikke har nye yrkesdata, og som derfor ikke skal yrkeskodes. For yrkeskoding skal ha mening må det undersøkes om:

1. Arbeidstakerforholdet er nytt, eller har endret yrke.
2. Arbeidstakerforholdet er aktivt i perioden.
3. Det er tilstrekkelig grunnlag for koding:
    a. Gyldig yrkeskode er foreslått av arbeidsgiver.
    b. Stillingskoder for offentlige eller maritime er benyttet.
    c. Det er skrevet tekst:
        i. Teksten er tilstrekkelig beskrivende.
        ii. Det foreligger tilstrekkelig støtteopplysninger.

Enhet i filene er løpende endringsmeldinger og årskontrollmeldinger. Fødselsnummer, dato for ansettelse og organisasjonsnummer brukes for å identifisere et arbeidstakerforhold, men dette identifiserer ikke en record unikt. Samme arbeidstakerforhold vil kunne forekomme uten sluttdato (aktivt) og med oppgitt sluttdato i en senere (opphørsmelding). Kortvarige arbeidstakerforhold kan ofte innmeldes etter at det er avsluttet, inn- og utmelding kommer samtidig. En annen form for dubletter oppstår rundt årskontrollen når årskontrollskjema og endringsmeldinger kan komme samtidig.
For personer som endrer yrke innen en bedrift hvor endringsmeldingen inneholder samme ansettelsesdato, kan yrkeskode og oppgitt yrkestekst kontrolleres mot tidligere oppgitte data. Hvis det er endringer i tekst, tildeles yrkeskode utfra denne, og arbeidstakerforholdet kan eventuelt få en annen yrkeskode.
Innkomne yrkesdata undersøkes og grupperes utfra muligheter for koding. Tildelt yrkeskode lagres i en egen variabel separat fra av innrapportert kode og tittel, og sendes tilbake til RTV. I tillegg opprettes en kildevariabel som viser med hvilken metode yrkeskoden er tildelt. Denne fungerer også som en kvalitetsrangering av forslag til yrkeskode. Tildelt yrkeskode/lnr lagres i en egen variabel uavhengig av innrapportert tittel, og sendes tilbake til RTV. Ved årskontrollen vil RTV melde tilbake koden og tilhørende tekst til oppgavegiver for kontroll, og eventuell retting.

Tabell 2.2-1: Trygdekontorene som har ansvar for Arbeidstakerregisteret i hvert fylke

Trygdekontor	Sted	Adresse	Poststed	Telefon	Antall AF[1]
Østfold	Halden	Postboks 56	1751 Halden	69 21 64 50	120 000
Akershus	Ullensaker	Postboks 163	2051 Jessheim	63 97 73 50	238 000
Oslo	Gamle Oslo	Postboks 308 Alnabru	0614 Oslo	23 40 33 50	472 000
Hedmark	Stor-Elvdal		2480 Koppang	62 46 00 44	90 000
Oppland	Gran		2770 Jaren	61 32 78 45	83 000
Buskerud	Ringerike	Postboks 98 Sentrum	3502 Hønefoss	32 17 11 20	116 000
Vestfold	Tønsberg	Postboks 2360	3103 Tønsberg	33 00 32 40	108 000
Telemark	Vinje		3890 Vinje	35 06 30 20	78 000
Aust- Agder	Tvedestrand	Postboks 188	4902 Tvedestrand	37 19 97 50	46 000
Vest- Agder	Lindesnes	Postboks 163	4524 Sør- Audnedal	38 25 53 80	75 000
Rogaland	Eigersund	Postboks 23	4379 Eigersund	51 46 41 07	210 000
Hordaland	Kvam	Postboks 93	5601 Norheimsund	56 55 68 00	239 000
Sogn og Fjordane	Sogndal	Postboks 143	6851 Sogndal	57 67 20 11	55 000
Møre og Romsdal	Molde	Storgata 12-14	6406 Molde	71 25 11 55	116 000
Sør- Trøndelag	Holtålen		7380 Ålen	72 40 50 10	146 000
Nord- Trøndelag	Namsos	Serviceboks 1019	7809 Namsos	74 21 83 50	56 000
Nordland	Dønna		8820 Dønna	75 05 20 00	109 000
Troms	Harstad		9480 Harstad	77 05 95 50	80 000
Finnmark	Gamvik		9770 Mehamn	78 49 67 50	35 000
Maritimt reg.	Rana	Postboks 423	8601 Mo i Rana	75 12 72 00	120 000

Arbeidsgivere som leverer papirskjema, får tilsendt årskontroll medio januar. Denne skal returneres medio mars, og alle skjemadata punches manuelt. Arbeidsgiver som har maskinell innrapportering leverer diskett hver måned, og trenger ikke kontrollere papirskjema.

2.2.1.2 Katalogdata

Yrkeskatalogen og andre lister over yrkesdata brukes ved ulike metoder for tekstbasert koding. Skrivemåten i Yrkeskatalogen er forsøkt standardisert og i mange tilfeller mer omstendelig enn det som kan forventes skrevet av arbeidsgiver. Det lages en derfor flere andre lister/kataloger, f.eks. med ulike skrivemåter og rekkefølge på ordene, samt med andre variabler enn i Yrkeskatalogen.

2.2.1.3 Kodemetoder

2.2.1.3.1 Koding basert på kun tekst

Innrapportert yrkestittel i fritekst bearbeides før videre behandling, og den bearbeidde teksten lagres separat. Det foretas konvertering til store bokstaver og fjerning av spesielle tegn og mellomrom. Mange vanlige forkortelser skrives ut til et helt ord. Første trinn tilordner yrkeskode ved sammenlikning av den oppgitte teksten med et utvalg av yrkeskatalogens titler som er entydige, og det er omlag en tredel som får en endelig yrkeskode ved denne metoden.

2.2.1.3.2 Koding støttet av andre kjennemerker i tillegg til tekst

For endel vanlige yrker er det laget egne programdeler basert på en rekke valg av yrkeskode utfra tekst og annen informasjon. Trinn 2 velger yrkeskode utfra oppgitt tekst i kombinasjon med et eller flere andre kjennemerker. I dette trinn bearbeides ikke teksten ytterligere, men det brukes tekst- og søkefunksjoner som man kan si er mer grovmasket enn trinn 1. omlag en firedel kodes i trinn 2.

2.2.1.3.3 Koding basert på andre kjennemerker enn tekst

Utfra analyser av AKU-data, er det laget tabeller over hvilke yrker som er mest sannsynlige innen ulike næringer og utdanninger, og kombinasjoner av dette. Disse tabellene brukes for å forslå en yrkeskode. Alle records fra dette trinn sendes til manuell kontroll og eventuelt omkoding. Det utgjør omlag en tredel av alle records med tekst.

[1] Antall arbeidstakerforhold pr. uke 52, 2002

2.2.2 Støtte av andre kjennemerker

2.2.2.1 Antall ansatte

Erfaringer i AKU viser at for mange blir kodet som administrative ledere hvis man bare tar utgangspunkt i tittelen. Yrkesstandarden skiller mellom ledere av små (færre enn 10 ansatte) og store bedrifter. For å yrkeskode ledere, må det tas hensyn til bedriftens størrelse i kodesystemet. For svært små bedrifter (færre enn 5 ansatte) benyttes ikke koder i yrkesfelt 1. Data om bedriftsstørrelse beregnes ved en kobling mellom data fra Enhetsregisteret og Arbeidstakerregisteret.

2.2.2.2 Institusjonell sektor

Arbeidstakerforhold i offentlige lønnsregistre som benytter stillingskoder skal kodes utfra disse stillingskodene. Allikevel er det for mange kommunale stillinger levert yrkestekst og istedenfor kode. Det tekstbaserte kodesystemet er laget slik at de vanligste yrkestitler kan kodes utfra tekst, men arbeidstakerforhold fra offentlig sektor prioriteres ikke i den manuelle kodingen.

2.2.2.3 Bedriftens næringskode

Næringskode benyttes som grunnlag for yrkeskoding da bedriftens produksjon har mye å si for arbeidstakerens faktiske arbeidsoppgaver. I yrkeskatalogen er næring/bransje det hyppigst brukte tilleggskjennemerke, men inndelingen tilsvarer ulike nivåer av Standard for næringsgruppering. Det benyttes ulike nivåer av næringsinndeling for forskjellige yrker, og noen tilfeller på tvers av aggregeringsnivåene.

2.2.2.4 Arbeidstakers utdanning

Yrke skal defineres utfra faktiske arbeidsoppgaver, og det er ikke alltid klart hvilken sammenheng det er mellom utdanning og faktiske arbeidsoppgaver. En antar at den vil være forskjellig for ulike yrker, og særlig lav for generelle utdanninger. Arbeidstakerens høyeste utdanning kan i visse tilfeller bidra til å tildele yrkeskode ut fra faglig spesialisering, og i andre tilfeller ut fra utdanningsnivå. Dette gjelder særlig skille mellom akademikere og høyskoleutdannede. Faginndelingen i yrkeskatalogen følger ikke alltid inndelingen i utdanningsstandarden, og det er kommet ny utdanningsstandard som også kan ha betydning for de valgene som er gjort.

Et problem ved nivåfordeling er arbeidstakere uten den formelle utdanningen som normalt kreves i yrket men som utfører arbeidsoppgaver på dette nivå. Dette kan være på bakgrunn av lang erfaring, intern-opplæring, eller mangel på arbeidskraft med formell utdanning. Slike forhold kan det automatiske systemet ikke oppfatte. I kodeprogrammet brukes personens utdanningskode som tilleggsopplysning for yrkeskoding av sivilingeniører og akademikere med realfagsutdanning, samt for lærlinger. For lærlinger er også igangværende utdanning viktig. Det er tilordnet en yrkeskode til en del utdanninger etter en vurdering av beskrivelsene av henholdsvis utdanning og yrkestitler. For mange utdanninger er beskrivelsen ikke egnet til å yrkesklassifisere utfra.

2.2.3 Videreutvikling

Som nevnt kan svært mange arbeidstakerforhold med helt vanlige og korrekt skrevet yrkestittel ikke kodes direkte utfra tekstfeltet. Det skyldes såkalte ikke-unike hovedtitler, altså titler som ser like ut men som brukes om helt ulike arbeidsoppgaver. Videre spiller forskjellig skrivemåter, skrivefeil, forkortelser og ikke minst uventet bruk av de ulike titlene en stor rolle.

Det vil i all klassifisering være en mengde arbeidstakerforhold som utfra de foreliggende data ikke kan kodes på det mest detaljerte nivå. Det er diskutert forskjellige måter å behandle disse på:

- Kode på et aggregert nivå. 1-3 siffer av yrkeskoden.
- Fordele til en valgt "normaltittel" innen gruppen eller den vanligste innen gruppen.
- Sende videre til ekspertkoding. Det vil da kunne vurderes utfra skjønn på bakgrunn av flere opplysninger, men dette vil ikke nødvendigvis føre til en mer konsekvent koding.
- For noen yrker finnes restgrupper, "andre innen...", som kan brukes til å kode de som ikke blir fordelt på spesifikke yrkestitler. I enkelte yrker er denne restgruppen tenkt som en samlegruppe for spesielle, sjeldne yrkestitler som ikke har egne koder. Å legge vanlige, men uspesifikke titler til en slik gruppe er derfor misvisende. Å opprette nye restgrupper "andre innen" i tillegg til de eksisterende er problematisk på grunn av oppbygningen av Standard for yrkesklassifisering

STYRK er ikke konsekvent med hensyn til bruk av siste siffer. Som en hovedregel slutter 4-siffer yrkeskode på 1-8 for vanlige yrker. 28 normale yrkeskoder slutter på 0, ikke på 1. For kun 2 yrkeskoder brukes 0 som siste siffer på en uspesifisert gruppe. For de øvrige 106 yrkesgruppene finnes ikke en slik kode.
Siste siffer=9 brukes i mange standarder for "andre innen". Kun for 26 grupper eksisterer slike yrkeskoder, og for 8 av disse er det ikke en restgruppe, men en separat gruppe.

Man har derfor valgt å bruke en 7-sifret kode i alle deler av systemet. En gradvis tilnærming hvor man først kunne fordele yrkestitlene på en mer overordnet kode, og på et finere nivå senere i rutinen, er i mange ikke tilfeller ikke mulig.

Når en yrkestittel skrives ulikt i forhold til katalogen, er det mange andre muligheter enn at man angir yrkeskategori på et mer overordnet nivå. Når den mest nøyaktige koden gis med en gang, så blir det en skarpere skille mellom korrekt koding og feil/ikke tildelt. Derfor vil tvilstilfeller bli kodet manuelt på 7 siffer framfor en automatisk koding på 2-4 siffer. For de fleste arbeidstakerforhold blir det foreslått en kode i det automatiske systemet som er ment å være til hjelp i den manuell rutinen. Omtrent 35% av disse forslagene stemmer overens med den manuelle kodingen. For arbeidstakerforhold som er for uspesifikke til å kodes på 7-siffer nivå, både automatisk og manuelt, blir returnert til arbeidsgiver (med koden 0000100) med beskjed om å skrive korrekt yrkeskode eller en mer dekkende tittel.

2.2.4 Publisering

Alle arbeidstakerforhold tildeles en 7-sifret yrkeskode. Fram mot publisering av yrkesstatistikk, vil det gjøres kvalitetsundersøkelser og vurderinger av hvor detaljert nivå man kan oppgi tall. Fra AKU publiseres det årlig statistikk på yrkesfelt (1 siffer), yrkesområde (2) og yrke (4) for de vanligste yrkene. Yrkesdata fra Arbeidstakerregisteret skal også brukes til statistikk på 1-4-siffer, men kan det være å aktuelt å publisere tall for yrkestittel (7 siffer) for enkelte yrker som det er spesiell interesse for og hvor man bedømmer at kvaliteten er god nok. Eksempel på yrkestitler som det er interesse for og som krever 7 siffer for differensiering er 'ergoterapeut' og 'fysioterapeut' (3226), . Det samme gjelder formenn, spesialarbeidere, fagarbeidere og lærlinger innen samme håndverk. Her må 7 siffer benyttes for å lage statistikk over det man kan kalle "lønnsgruppene" innen hvert fagområde.

2.2.4.1 Kort om ulike kvalitetsmål

Metode	*Status*
Makronivå: fordelingen og utvikling (endringstall) av yrker sammenliknes med tilsvarende statistikk i AKU	Dette er undersøkt for de viktigste yrkene, men gir ikke mening for aggregerte nivåer inntil PAI og SST er yrkeskodet.
Skjønnsmessig vurdering: visuell gjennomgang av yrkestitteltekst, og foreslått kode.	Gjennomført for over 10.000 arbeidstakerforhold i 2001 Og noe over 9000 arbeidstakerforhold i 2002.
Mikrokonsistens: sammenlikning av yrkeskoder AA og AKU på jobbnivå.	Har vært forsøkt med litt forskjellige metoder, se notat av Leiv Solheim oktober 2002.
Omkoding: sammenlikning av manuell yrkeskode i AKU, med yrkeskode tildelt av automatisk program utfra tekstfeltet "IOs yrke".	Benyttet mye ved oppbygging av det automatiske kodesystemet, men også som mål på endelig kodeopplegg - se notat av Leiv Solheim oktober 2002..
RTV årskontrollen: bedriftene kan sende rettemelding hvis de oppfatter at tildelt yrkeskode er feil.	En har inntrykk at arbeidsgivere i svært varierende grad kontrollerer yrkeskodene. Det er også usikkert i hvilken grad arbeidsgiver endrer koder til mer korrekte.

Omkoding har vært benyttet en del i utvikling av systemet for tekstbasert koding. I denne metoden konverteres et datasett fra AKU til et datasett som er kompatibelt med system for yrkeskoding. Man velger ut "IOs yrke" samt næring, utdanning og noen andre kjennemerker. Programmet koder AKU-yrkestittelen uavhengig av den yrkeskoden som personen har fått i AKU. Til slutt sammenliknes de to yrkeskodene for hver person. Samsvaret mellom kodene gir et kvalitetsmål på den automatiske kodingen i den grad AKU-kodingen er korrekt. Følgende forhold må bemerkes ved slike sammenlikninger:

- Tilfeldige feil vil oppstå i AKU siden det kodes manuelt.
- Systematiske feil i AKU kan skyldes misforståelser eller feil i instruksen. Disse vil føre til skjevheter i sammenlikningene.
- Systematiske feil i det automatiske yrkeskodingssystemet kan skyldes programmeringsfeil og mangler i inndata.
- Tilfeldige og systemtiske ulikheter i tekstgrunnlaget for de to gruppene som sammenliknes. I AKU finnes en variabel for arbeidsoppgaver, i tillegg til yrkestittel. Det gjør at yrkeskodingen her har et mer adekvat grunnlag, men også at selve yrkestittelen i mange tilfeller skrives på en kortere måte. Intervjuerne kan føre kortversjon av yrkestittelen, fordi arbeidsoppgavene skal føres i et annet felt. I Arbeidstakerregisteret kan arbeidsgivere levere yrkestittel som en tekst på maksimalt 40 tegn.
- Teksten i AKU er den yrkestittel som IO selv har oppgitt, i AA er det arbeidsgiver som fører yrke. Dette fører nok til en god del forskjeller i hva som rapporteres.
- I automatisk koding brukes næring og utdanning systematisk for å bestemme visse yrkeskoder. I AKU benyttes disse kjennemerker mer skjønnsmessig.

Ved mikrokonsistensanalyser mellom AKU og AA vil noen av de samme feilkildene gjøre seg gjeldende. I tillegg kan det være tilfeller der det ikke er entydig kobling av arbeidstakerforhold for personer som har flere arbeidstakerforhold. Dette skyldes forhold ved kjennemerker som definerer arbeidstakerforhold i Arbeidstakerregisteret. I AKU er det ikke oppgitt organisasjonsnummer for alle, men de fleste har navn på bedriften. Kobling mellom bedriftsnavn vil være mer usikker enn kobling på organisasjonsnummer, man kan eventuelt benytte næringskode i tillegg. Videre vil tidspunkt og tidsforskyvninger kunne spille en rolle, da det er en betydelig del arbeidstakere som skifter jobb, og en viss andel som skifter yrke.

2.2.4.2 Tester av kodingssystemet ved rekoding

AKU-datamateriale ble mye brukt ved utvikling og kvalitetskontroll av yrkeskodingssystemet. Yrke kodes utfra tekstfeltet for yrkestittel og andre opplysning i AKU ved hjelp av samme metoder som skal benyttes på Arbeidstakerregisterdata. Vurderingen av kodingen baserer seg på samsvar mellom den automatiske kodingen og den koden som tidligere er manuelt yrkeskodet i AKU. Man sammenlikner 1. siffer, 1-2, 1-3, og hele yrkeskoden. Samsvar på 4-siffernivå betyr at de to kodene stemmer overens, ikke nødvendigvis at de er korrekte for det aktuelle arbeidstakerforhold. Hovedresultater:

- De fleste arbeidstakerforhold er enten kodet helt likt eller helt ulik for alle metoder.
- Omtrent to tredjedeler av arbeidstakerforholdene kan kodes automatisk med akseptabel kvalitet.
- Omtrent en tredjedel av arbeidstakerforhold med oppgitt tekst sendes til manuell koding.

Hvert arbeidstakerforhold som forsøkes kodet, kan få ulike koder utfra hvilke metoder som benyttes. Det er derfor viktig å kunne vurdere kvaliteten på de ulike metodene. Inntil man får en kontrollert årgang av Arbeidstakerregisteret, baseres undersøkelsene på AKU-materiale. Rangering av de ulike forslag til yrkeskode som de ulike deler av kodesystemet gir, vil være utslagsgivende på den totale kvaliteten. Generelt legges metoder med få treff og høy kvalitet tidlig i prosessen, mens metoder med flere treff og mindre samsvar legges lenger ut.

2.3 KODING AV YRKE VED OPPGITT STILLINGSKODE

For arbeidsgivere som registrerer arbeidsforhold i visse registre, forutsettes det at de leverer *stillingskoder* til Arbeidstakerregisteret istedenfor yrkestittel eller yrkeskode. Stillingskodene er spesielle for hvert register, og skiller seg fra yrkeskoder ved at de kan være mer knyttet til lønnsstruktur, ansiennitet, og andre forhold enn beskrivelse av arbeidsoppgaver. Aktuelle registre er:

- PAI: personelladministrativt informasjonssystem (Kommuner og visse andre bedrifter)
- SST: statens sentrale tjenestemannsregister.
- STS: sentralt tjenestemannsregister for skoleverket.
- Maritime stillinger. (eget register)

2.3.1 PAI

PAI omfatter de fleste kommunale og fylkeskommunale stillinger, samt noen få ansatte i visse andre organisasjoner. I PAI benyttes ca. 600 stillingskoder, som i større eller mindre grad tilsvarer yrker. For de fleste stillingskoder er det en såpass sterk overensstemmelse mellom stillingsbetegnelse og yrke at man tildeler yrkeskode utfra stillingskode alene. Imidlertid har mange arbeidstakerforhold stillingskoder som ikke kan kodes slik, og som må kodes ved hjelp av ulike kombinasjoner av andre kjennemerker som:

- Bedriftens næring.
- Arbeidstakerens utdanning fra "Befolkningens høyeste utdanning", BHU.
- Tjenestested, et kjennemerke fra PAI.
- Lønnstrinn og antall ansatte har også vært foreslått som mulige kjennemerker.

Ved å hente inn andre kjennemerker bringer man inn en del mulige feilkilder. BHU angir den høyeste utdanning som er registrert pr. 1.10 to år før. PAI inneholder alle arbeidstakerforhold pr. 1.10 året før. For arbeidstakerforhold som endres etter dette vil opplysninger fra PAI i mange tilfeller ikke kunne gjenfinnes. Det får litt forskjellige konsekvenser for ulike kjennemerker:

- For en del bedrifter kan det lages en koblingsnøkkel mellom bedriftsnummer og tjenestested. For andre er det ikke et entydig forhold mellom tjenestested og bedrift. Tjenestested tilsvarer etat eller fagområde og kan omfatte ulike bedrifter og ulike geografisk plasseringer.
- For lønnstrinn er en tilsvarende kobling ikke mulig, da det er knyttet til stilling pr. person og ikke kan utledes av andre variabler.

Det vil bli en viss gruppe arbeidstakere som ikke kodes, og som klassifiseres som uoppgitt. Dette skyldes blant annet manglende opplysninger fra forsøk på å koble til andre registre. Det er også noen få stillingskoder som ikke kan yrkeskodes utfra foreliggende data.

For å kontrollere overgangsnøkkelen ble det i utgangspunktet planlagt å bruke en fil hvor PAI og Yrkesregisteret var koblet sammen. Denne koblingen er programmert på KOSTRA-prosjektet. PAI-registeret for 2000 har opprinnelig 420 747 observasjoner. Etter koblingen var det imidlertid bare 126 412 observasjoner med opplysninger på både stillingskode, næring, tjenestested og utdanning. På grunn av at så få observasjoner har opplysninger på de variablene som blir brukt i omkodingen er det som en midlertidig løsning valgt å koble næring fra enhetsregisteret på organisasjonsnummeret i PAI-registeret. Denne koblingen gir næringsopplysninger på 343 641 observasjoner.

521 stillingskoder koder er kodet 1-1. Denne kodingen er foreløpig ikke sammenlignet med overgangsnøkkelen til seksjon 420. Når det gjelder 1:1-kodingen er det fortsatt en del problemer som ikke er avklart til nå.

2.3.2 SST

SST danner grunnlag for lønnsstatistikk og benyttes ved lønnsforhandlinger, personellplanlegging og budsjettering. Arbeidstakere i statlige virksomheter, som følger statens lønnsregulativ skal rapporteres inn. Tellingstidspunktet er 1.oktober, og Statistisk sentralbyrå mottar en kopi av registeret i mars påfølgende år.
Data som leveres er bl.a.: fødselsnummer, stillingskode, etatskode, tjenestestedskode, lønnstrinn og utdanning. Det benyttes omlag 1000 stillingskoder. Som med stillingskoder i PAI er det større eller mindre grad av samsvar mellom stillingskoder i SST og yrkeskoder. For de fleste stillingskoder er det en såpass sterk overensstemmelse mellom stillingsbetegnelse og yrke at man tildeler yrkeskode utfra stillingskode alene. Mange arbeidstakerforhold kan ikke kodes utfra stillingskode alene, og man benytter andre kjennemerker i tillegg: etatskode, tjenestestedskode, næring, utdanning fra BHU, antall ansatte og lønnstrinn.

2.3.3 STS

Formålet med STS er som for SST og gjelder stillinger i skoleverket. Alt undervisningspersonale og andre tjenestemenn i statlige skoler (unntatt høyskoler), fylkeskommunale og kommunale skoler, samt private folkehøgskoler, som lønnes etter statens regulativ pr. 1.oktober, skal meldes inn.
Innrapporterte data er stort sett som for SST med id-nummer for skole istedenfor etat/tjenestested. Det benyttes omtrent 100 stillingskoder, hvorav noen kun benyttes i STS og ikke i SST.

2.3.4 Maritime stillinger

Bedrifter innen sjøtransport (61) og olje- og gassutvinning (11) skal innrapportere yrkeskoder til det maritime arbeidstakerregisteret. Dette registeret har et eget system for koding av yrke som skal opprettholdes også etter at STYRK-kodingen er innarbeidet. Dette notatet beskriver arbeidet med å lage overganger mellom de maritime yrkeskodene og STYRK-koder.

2.3.4.1 Kort om det maritime yrkeskodingssystemet

De maritime kodene er delt inn i to nivåer: På det øverste nivået finnes det 24 grupper: 0100 Skipsfører - 2400 Praksisplasser. Kodene på dette nivået kjennes ved at de to siste siffer er 00. Dette er koder som kun brukes til å markere hvilken gruppering kodene på neste nivå befinner seg i , de er ikke tilgjengelige for bruk i innrapporteringen til det maritime AA-registeret. Innenfor noen av grupperingene finnes det en restgruppe. Denne har 90 som sine to siste siffer og inneholder yrker i hovedgrupperingen som ikke har en egen kode/tittel på lavere nivå.

2.3.4.2 Yrkeskoding basert på tilhørende tekst, "1-1"-koding

kode	tittel	STYRK
100	**Skipsfører**	
110	Skipsfører	3142
200	**Styrmenn**	
210	Overstyrmann	3142
220	Styrmann	3142
290	Styrmann	3142
400	**Radiopers.**	
410	Radiooffiser	3132
420	Radiooperatør	3132
430	Radiotelefonist	3132
490	Radiopers.	3132
500	**Elektropers.**	
510	Elektriker	7241
520	Elektrofagmann	7241
590	Elektropers.	7241
600	**Plattformsjef**	1222
610	Plattformsjef	1222
700	Stabilitetssjef	3116
710	Stabilitetssjef	3116
800	**Kontrollromsoperatør/-assistent**	
820	Kontrollromsassistent	8161
820	Kontrollromsassistent	8161
1200	**Sertifikatpliktig forpleiningspers.**	
1210	Forpleiningssjef	5122
1220	Kokk	5122
1230	Stuert	5122
1300	**Aspiranter**	
1310	Offisersaspirant dekk	8342
1320	Offisersaspirant maskin	8341
1330	Aspirant dekk	8341
1340	Aspirant maskin	8342
1350	Skipsmekanikeraspirant	7234
1400	**Underordnet dekkspers./skip**	
1410	Matros	8341
1420	Lettmatros	8341
1430	Arbeidsleder	8341
1440	Båtsmann	8341
1450	Matroslærling	8341
1490	Dekkspers.	8341
1500	**Underordnet maskinpers./skip**	
1510	Maskinpasser	8342
1520	Motormann	8342
1530	Smører	8342
1540	Maskinoffiserassistent	8342
1550	Motormannslærling	8342
1590	Maskinpers.	8342
1600	**Underordnet pers. dekk fl. innr.**	
1610	Arbeidsleder - plattform	7215
1620	Kranfører	7215
1630	Rigger	7215
1640	Riggerassistent	7215
1690	Annet ikke sert. pl. dekk/maskin	7215
1700	**Skipsmekaniker**	
1710	Skipsmekaniker	7234
1900	**Restaurasjons/catering pers.**	
1910	Restaurasjons/catering personale	9133
2000	**Borepers.**	
2010	Boresjef	1222
2020	Ass. Boresjef	1222
2030	Borer	8113
2040	Ass. Borer	8113
2100	**Annet pers. skip/flyttb. innr.**	
2110	Dykkerpers.	7216
2125	Lasteoffiser	9330
2130	Lege	2221
2135	Sykepleier	3231
2140	Sikkerhetsleder	3152
2145	Staff-captain	1232
2300	**Lærlinger**	
2310	Lærling catering	9133
2320	Lærling skipselektriker	7241
2330	Lærling fiske/fangst	6411

2.3.4.3 Kort om aspiranter til sjøs

På land gjelder prinsippet om at lærlinger skal kodes i det yrket de kvalifiserer seg til. Dette gjelder selvfølgelig også til sjøs. Da kan det virke underlig at aspiranter og offisersaspiranter gis samme kode, men dette er faktisk riktig. Den eneste forskjellen på aspiranter og offisersaspiranter er at offisersaspiranter har høyere utdanning enn aspirantene. Dette gir seg imidlertid ikke utslag i det arbeidet de utfører om bord, men i hvor lang fartstid de trenger før de kan tas opp til offisersutdanning, offisersaspiranter trenger kortere fartstid enn vanlige aspiranter.

2.3.4.4 Yrkeskoding med næring som tilleggsinformasjon

7 maritime yrkeskoder kan tilordnes yrkeskoder ved hjelp av opplysninger om næring. Arbeidsforhold i sjøtransport har næringskode 61 og arbeidsforhold i olje/gassutvinning har næringskode 11 ("Utvinning av råolje og naturgass. Tjenester tilknyttet olje- og gassutvinning").

maritim kode	tittel	SN94	STYRK
300	**Maskinister**		
310	Maskinsjef	61	3141
		11	3116
320	1. Maskinist	61	3141
		11	3116
390	Maskinist	61	3141
		11	3116
1000	**Teknisk sjef**		
1010	Teknisk sjef	61	3141
		11	3116
1100	**Teknisk assistent**		
1110	Teknisk assistent	61	3141
		11	3116

2.3.4.5 Næringsproblemer

Siden næring benyttes som tilleggsopplysning i omkodingen, skaper det problemer hvis enheter meldes inn med annen næring enn det som er forutsatt. Og dette skjer i et visst omfang: I en fil fra AA-reg (per00u39) var 6,1% av de maritime arbeidsforholdene registrert med næring 63 "Tjenester tilknyttet transport- og reisevirksomhet", 5,4% med registrert med næring 74 - "Annen forretningsmessig tjenesteyting" og 2,3 % registrert med næring 55 - "Hotell- og restaurantvirksomhet". I alt var 21,7% av arbeidsforholdene registrert med en annen næring enn 11 eller 61. Fordeling etter foretakenes næring gir det samme bildet av situasjonen som fordelingen etter bedriftens næring i det maritime arbeidstakerregisteret. Det er undersøkt om variabelen "fart" kan benyttes i stedet for næring for å omgå dette problemet. Det kan den ikke, ettersom det i nevnte fil ikke er noen påtagelig sammenheng mellom fart/ikke fart og næring 11/ næring 61. Noen koder er det foreløpig vanskelig å gi en entydig yrkeskode.

2.4 KONTROLL AV OPPGITT YRKESKODE

2.4.1 Nominell kontroll

For arbeidstakerforhold der yrkeskode er oppgitt av arbeidsgiver, foretas en kontroll av yrkeskode. Det undersøkes om den oppgitte yrkeskoden finnes i yrkeskatalogen ved at man slår opp direkte i yrkeskatalogdatabasen. RTV får oppdatert katalog hver uke i forbindelse med oversending av data, og skal kontrollere yrkeskoden som punches.
Generelt om endringer kan man si at man legger til flere nye yrkestitler etter som behovet melder seg, men er svært restriktiv med å fjerne yrkestitler fra katalogen.

Tabell 2.4-1: Rutine for eksistenskontroll av oppgitt yrkeskode

Kriterium	Kildevariabel	Yrkeskode SSB
7-sifret yrkestittelkode finnes i yrkeskatalogsdatabasen	1.1	Oppgitt kode
4-sifret yrkeskode finnes, men ugyldig løpenummer	1.2	4 siffer av oppgitt kode + "100"
4-sifret yrkeskode ugyldig	1.9	"0000100"

2.4.2 Realkontroll

På sikt kan det være aktuelt å kontrollere yrkeskoden i forhold til kriterier som er indikative for arbeidsoppgaver. Dette dreier seg om kjennemerker som: næring, utdanning, bedriftsstørrelse, institusjonell sektor. En slik kontroll vil også kunne implementeres for stillingskode- og tekstbasert yrkeskoding, som en intern kvalitetskontroll av disse metoder.

2.4.3 Rutine når hverken yrkeskode eller tilstrekkelig tekst er oppgitt

Arbeidstakerforhold hvor det verken er oppgitt forslag til yrkeskode eller tekst som yrkestittel, vil innledningsvis få yrkeskode til 0000100 = uoppgitt, og kilde til 99 = ikke tilstrekkelig grunnlag.

Hvis det er oppgitt tekst, men den er for generell til å kunne klassifiseres f.eks. "vikar" settes også yrkeskode til 0000100 og kilde=99, og arbeidstakerforholdet sendes ikke til manuell koding. På årskontrollskjemaet blir "0000100" tekstet "Yrkestittel må angis mer presist", for at arbeidsgivere skal få en tilbakemelding om å skrive kode eller en mer beskrivende tekst.

Også en del arbeidstakerforhold som sendes til manuell koding får yrkeskode "0000100" (uoppgitt), fordi forslaget til yrkeskode er svært usikker. Videre blir en del arbeidstakerforhold fortsatt stående med 0000100 etter manuell koding, fordi det ikke lot seg kode.

2.5 Spesifikasjon av system for yrkeskoding

2.5.1 Eksterne systemer

Programmer og systemer som lages i andre seksjoner og eksterne enheter, forutsettes dokumentert av de ansvarlige for disse. De omfatter:

- Yrkeskatalogen (Oracle)
- Manuell koding (Oracle)
- Datalinje til RTV
- SSB-web

2.5.2 UNIX-område $YRKEREG

Yrkesdata og vanlige endringsmeldinger fra Arbeidstakerregisteret leveres hver uke på linje fra RTV. Det skal kjøres et programsystem hver fredag for å sikre at RTV har data til sine oppdateringer på tirsdager.

2.5.3 Manual for yrkeskodingssystem, UNIX-delen

2.5.3.1 Før du kan begynne må du ha:

- lese- og skrivetilgang til UNIX-område **$YRKEREG**
- lese- og skrivetilgang til UNIX-område **$ARBTAKER**
- lese- og leggtil-tilgang til databasen **KPR1** og bruker: **AA-reg** (eller høyere)
- lesetilgang til **OPR6** tabell: **yrkeskat.yrkeskatalog**
- mulighet for å logge deg på UNIX-server "Ursus"
- fordel med samme passord for KPR1 og OPR6

2.5.3.2 Første gang du bruker opplegget (hvis du vil ha automatisk start)

1. Logg på UNIX.
2. **`nedit .login`**
3. Gå til slutten av fila (ctrl end).
4. **`cd $YRKEREG/prog ; sasx`**
5. Lukk fila (file, close, yes).
6. Avslutt UNIX.
7. Logg på UNIX.

2.5.3.3 Starte rutineopplegget (hvis du ikke har automatisk start)

8. Logg på UNIX.
9. Avslutt SAS hvis det er startet.
10. **`cd $YRKEREG/prog ; sasx`**

2.5.3.4 Rutineopplegget (med automatisk start)

11. Logg på UNIX og vent på at SAS skal starte.
12. Fyll ut passord i inputvindu (rosa).
13. Les loggen, og kontroller eventuelle feilmeldinger.
14. Hovedmenyen (blått vindu) kan brukes for å styre alle rutinefunksjoner.
15. Bruk menykommandoer kjør programmer | oppdater yrkeskatalogen
16. Bruk menykommandoer kjør programmer | kjøre rutinekoding pr. uke
17. Fyll ut år og uke i inputvindu (gult).
18. Programmene tar vanligvis en del tid, ikke gjør noe her før du får e-post som sier at det er ferdig.
19. Les loggen og kontakt s260 hvis det er noen unormale feil.
20. Logg av UNIX.

2.5.3.5 Litt informasjon om hva programmet gjør

1. Data hentes i ukefiler fra RTV.
2. Sjekker om vi har samme jobben med yrkeskode allerede.
3. En stor del av de med tekst blir kodet automatisk, og lagt til yrkesregisteret.
4. Datatabellen med de som skal kodes manuelt blir fylt på.
5. Henter ferdig kodede fra denne tabellen, og lagt til yrkesregisteret.
6. Data fra punkt 3 og 5 blir sendt av gårde til RTV, på en egen datalinje.

Tabell 2.5-1: Variabler

AAR	Char	4	År (når RTV)
AFID	Char	30	Identifikator arbeidsforhold SSB
ANSATT	Char	8	Ansattdato
ANSATTE	**Num**	8	Antall ansatte i bedrift. Fra "arbsum16".
ANSATTF	**Num**	8	Antall ansatte i foretak. Fra "ftk_arbsum4".
BNAVN2	Char	35	b-off. navn2 / karakteristikk Bedriftsnavn
BNAVN3	Char	35	b-off. navn3 / karakteristikk Bedriftsnavn
BRD_NAVN	Char	60	b-redigert navn/navn1 Bedriftsnavn
FERDIG	Char	7	Yrkeskode tildelt ved SSB.
FNR	Char	11	Fødselsnr.
FORG_FRM	Char	4	f-organisasjonsform
F_IOSYRK	Char	100	Yrkestittel i klartekst, bearbeidet
KILDE	**Num**	8	Kilde til yrkeskode SSB.
ORG_B	Char	11	Organisasjonsnr. Bedrift
ORG_F	Char	9	Organisasjonsnr. Foretak
P_AKOMM	Char	4	Prioritert arbeidsstedskommune
P_NACE	Char	5	Prioritert næringskode
P_SEKTOR	Char	3	Prioritert institusjonell sektor
REGSTAT	**Num**	8	Dato for registrering ved SSB.
SLUTTA	Char	8	Sluttdato
UKE	Char	2	Uke (når RTV)
YRKTEKST	Char	40	Yrkestittel i klartekst (som innrapportert)
YRK_KODE	Char	7	Kode for yrke (som innrapportert)
BHUM	Char	6	høyeste utdanning fra BHU eller igangværende utdanning; nus2000

`"Tilkod"` inneholder arbeidstakerforhold hvor **yrktekst** eller **yrk_kode** i ukefil er nytt eller forskjellig fra det som er i registeret. "`Tilkod`" splittes i datasett til delsystemer. Alle filer som brukes i de ulike kodeprogrammene skal følge filbeskrivelsen over.

Filen fra tekstsystemet yrke.man&aar.&uke benyttes til ekspertkoding ved seksjon 450. Observasjoner med kvalitet lavere enn en terskelverdi (for tiden kilde>6.3) skal kontrolleres/tildeles yrkeskode ved manuell innmating.
De legges til Oracle-databasen ved s450:

- DB = KPR2
- tablespace = S450_DATA
- table = YRKER_TIL_KODING
- rolle = AAREG

Alle kodede datasett brukes til oppdatering av yrkesregister. I register er **afid** ikke unik - personen kan bytte yrke i samme jobb. Mulige nøkler er **afid+regstat** eller **afid+ferdig**.

Kun data fra tekstbasert (maskinell og manuell) koding leveres til RTV.

Arbeidstakers utdanning kobles på fra en utvalgt kode fra "Befolkningens høyeste utdanning" (BHU) og "Igangværende utdanning". BHU angir den høyeste utdanning som er registrert pr. 1.10., og man benytter samme årsfil ved koblingen hver uke. Det tas en kopi til `$YRKEREG/prog/utdanning` hvert år.

Manuell yrkeskoding skjer kontinuerlig, og det kjøres program som eksporterer data til RTV og oppdaterer yrkesregisteret: "ferdig_s450.sas". (Dette er bygget inn i "start.sas" fra 2002, og skal kjøres hver uke).

2.5.3.6 Forholdet til databasen som brukes til yrkeskoding ved s450

2.5.3.6.1 Status for koding

Variabelen FERDIG_KODET forteller som navnet antyder om posten er ferdig kodet.
Verdien er missing "." (punktum) for nye record som kommer inn.
Når en record er kodet ved s450 settes `FERDIG_KODET = 1`
Når en record er hentet ut til s260 settes `FERDIG_KODET = 2`
Denne variabelen tas ikke med til yrkesregisteret.

2.5.3.6.2 Opphav til kode

Variabelen KILDE angir opphavet til kodingen, og brukes til kvalitetsvurdering. Kodene 6-9 settes i det rutinemessige opplegget ved programmet "start". Alle som blir kodet ved s450 skal få KILDE = 9 ved kodeprogrammet på databasen der, noen har imidlertid beholdt sin opprinnelige kode (6 - 8). NB: når record hentes ut, settes kilde = 0.9 Dette for at de skal bli lavere enn alle automatiske. koder for senere sortering. Altså: Ved s260 betyr kilde=9 "sendes til s450" og kilde=0.9 "hentet fra s450".

2.6 DETALJER I TEKSTBASERT KODING

2.6.1 Datasett med yrkestitler og yrkeskoder

a) **YRKESKATALOGEN**: Yrkeskode, løpenummer og fullstendig yrkestittel, samt merke for antall ansatte og sektor. Yrkestittelen deles i primær- og sekundærtittel. Primærtittel defineres her som alle tegn fram til første komma eller parentes, det kan være 1-5 ord. Eventuelle mellomrom og bindestreker beholdes. Primærtittel er i hovedsak slik titlene enklest vil bli skrevet. Noe over 3000 av titlene har kun én yrkeskode pr. primærtittel, og kan kodes enklest utfra dette. Sekundærtittelen skilles ut og det lages en transposisjon av primær og sekundærtittel for de tilfellene der dette ikke overstiger 50 tegn. Oppdateres hver uke

b) **TRINN**: Unike primærtitler og transponerte titler. Oppdateres hver uke.

c) **PERTINENT**: ord som identifiserer et yrke uten at det nødvendigvis utgjør en hel yrkestittel. Lager av en liste over alle ord som forekommer i titlene i Yrkeskatalogen. Analyserer denne listen for å finne ord som kan identifisere yrker i ved å finne ord som er slik at de yrkestitlene der ordet forekommer, alle har samme yrkeskode. Løpenummeret settes til 100. Oppdateres hver uke.

d) **NUKE11**: Koding basert på næring og utdanning. Inneholder 5 siffer næring og 6 siffer utdanning, samt den yrkeskoden som er mest vanlig for denne kombinasjonen. Basert på AKU-data, oppdateres ikke regelmessig.

e) **NARPUT**: Koding basert på kun næring (5 siffer). Basert på AKU-data, oppdateres ikke regelmessig.

2.6.2 Bearbeiding av den innrapporterte yrkestittelteksten

a) Venstrejustering, fjerning av sluttblanke, konvertering til store bokstaver.

b) Noen forkortelser skrives om til hele ord, noen ord fjernes:

1.	FØRSTE
ASS. (i begynnelsen)	ASSISTERENDE
ASS. (resten)	ASSISTENT
ADM. (i begynnelsen)	ADMINISTRERENDE
ADM. (i resten)	ADMINISTRASJON
AVD.	AVDELINGS
ING.	INGENIØR
PROD.	PRODUKSJONS
ARB.	ARBEIDER
VIKAR	*blank*
STUDENT	*blank*
EKSTRAHJELP	*blank*
SOMMERVIKAR	*blank*
FERIEVIKAR	*blank*

c) Tegn som erstattes med en blank: **, : ; ( ) / \ — 2 3 4 5 6 7 8 9 0**

2.6.3 Koding

1. Nødvendige programmer
    a. "$YRKEREG/prog/start.sas": hovedprogram som styrer andre, og sender data til og fra eksterne kilder.
    b. "trinn1-makroer.sas": makroer for tekstbearbeiding.
    c. "trinn1.sas": koding etter div. lister.
    d. "trinn2-makroer.sas ": makroer med valg av næring.
    e. "trinn2.sas": koding utfra tekst og næring, til dels utdanning.
    f. "trinn3.sas": valg av kode, og div. opprydding.
2. Arbeidstakerforholdene gis flere foreløpige yrkeskoder, ved ulik metoder. For de ulike kodeforslag får variabelen *kilde* en verdi som forteller om opphavet og antatt kvalitet, 6.1= best, høyere tall angir større usikkerhet.
   NB: det er nummerisk datatype, men verdien er ikke proporsjonal med kvaliteten, det er foreløpig en slags kategorivariabel.
3. Mange arbeidstakerforhold kan ikke kodes utfra skrivemåte alene, derfor forsøkes også yrkeskode utfra en rekke ulike metoder, se tabell 3.
4. Hvert arbeidstakerforhold kan få flere ulike forslag til yrkeskode på dette stadium.
5. Hvis det er flere forslag velges det med antatt best kvalitet (lavest kilde-verdi).
   Hvis det er flere forslag med samme kvalitet, velges det med høyeste yrkeskodetall, altså lavest yrkesfelt.
6. Observasjoner som er forslått yrkeskode, men har kvalitet>6.3 kontrolleres manuelt.

Tabell 2.6-1: Verdier av kildevariabel i yrkesregisteret og interne filer

Kilde for yrkeskode	Kilde
Yrkeskode oppgitt	1
PAI-kode oppgitt	2
SST/STS-kode oppgitt	3
Maritim kode oppgitt	5
Oppgitte tekst samsvarer med en unik hoved-yrkestittel i katalogen (trinn 1)	6.1
Oppgitt tekst + næring og evnt. utdanning (trinn 2)	6.2
Oppgitt tekst inneholder et karakteriserende ord fra "**pertinent**" (trinn 1)	6.3
Oppgitt tekst lyder som en hovedyrkestittel, ved **soundex**-funksjonen	6.4
Koding basert på næring og utdanning (11 siffer)	6.5
Koding basert på kun næring (5 siffer)	6.6
Reservert nye metoder	7 – 8
Sendes til koding ved seksjon 450	9
Mottatt fra seksjon 450 (altså tidligere kilde 6.4 - 9.0)	0.9
Utilstrekkelig grunnlag	99

2.6.4 Ideer om videreutvikling av automatisk tekstbasert yrkeskoding

2.6.4.1 Metoder i SAS-programmering

Kodeprogrammet benytter en rekke tekstfunksjoner i SAS. Det foretas sammenlikninger av ord og delstrenger samt beregning av lydlikhet. Det er også gjort innledende forsøk med andre metoder, som kan finne flere anvendelser ved en videre utvikling av prosjektet og eventuelt økte maskinressurser og kunnskap.

2.6.4.2 Vanlig strengsøk

Videre benyttes søkefunksjonen "Index" mange steder i systemet. "Index" foretar søk etter eksakte delstrenger i andre strenger. Videre har man "Indexw" som foretar ordsøk. Med ord menes her en delmengde av den totale strengen avgrenset av visse tegn (blank og en del vanlige skilletegn). Begge returnerer posisjonen som heltall, med 0 for ikke treff, slik at uttrykk kan brukes i boolske tester.

2.6.4.3 Asymmetrisk staveavstand

Funksjonen "Spedis" (Spelling Distance) kan brukes for å sammenlikne tekst og gi en justerbar toleranse for ulikheter i skrivemåte. Funksjon regner ut den asymmetriske staveavstanden mellom søkeord og nøkkelord. Staveavstanden baserer seg på typen av og antallet perturbasjoner som må til for å omvandle nøkkelordet til søkeordet, relatert til lengden av søkeordet. Staveavstanden kan inngå som et parameter i programdeler som sammenlikner den oppgitte tekst med yrkestitler i yrkeskatalogen. Parameteret kan justeres i forhold til kvalitet og kvantitet, kort sagt: øker man toleransen for ortografiske forskjeller, stiger både antall treff og antall feil. Feilene skyldes ikke bare graden av ulikhet, men også uspesifikke titler, flertydige ord og yrkestitler som ligger nær hverandre i skrivemåte men har svært ulik betydning. Metoder som gjør bruk av funksjonen er testet ut, men er foreløpig ikke implementert i rutinekodingen. Et hovedproblem er at den optimale toleransen er forskjellig for ulike titler og at det ikke er enkelt å beregne av en felles grense eller en teknisk løsning for å bruke ulike grenser er funnet.

2.6.4.4 Mønstergjenkjenning

Funksjonen "rxmatch" kan benyttes til mønstergjenkjenning i tekst (bokstavlig eller variabler), liknende *grep* i Unix for søk etter regulære uttrykk. Til forskjell fra en streng inneholder et mønster symbolske regler for gjenkjenning. I forbindelse med yrkeskoding er dette forsøkt for å identifisere nummeriske tegn der yrkeskode er skrevet istedenfor tekst i tekstfeltet, men metoden er resurskrevende og er erstattet av enklere metoder. Funksjonen "rxchange" brukes for å transformere tekstmønstre og er brukt til å lage lister med ulike rettskrivningsvarianter.
Eksempler: ede → eia , yke → juke
Rxchange kan også endre mønster av ulik lengde eks.: e→ ei , ø→au , mens funksjoner "translate" og "transwd" konverterer henholdsvis kun enkelttegn og ord. Disse funksjonene er til gjengjeld raskere, og brukes i kodesystemet til å skrive om forkortelser til hele ord, gjøre om skilletegn til blanke og erstatte visse ord.

2.6.4.5 Synonymi i yrkestitteltekst

At to ord er synonymer betyr vanligvis at det har samme betydning, dvs. samme semantiske innhold. I forhold til leveringen av yrkestittel har man observert at det brukes svært mange ulike tekster, som er ord og ordkombinasjoner, samt varianter av dette – som ulike målformer, dialekter, og rene skrivefeil. Det vil derfor være mindre nyttig å bruke eksisterende synonymordlister til våre formål. Uten å gå inn på årsaken til alle tekstvariantene, kan man definere 'yrkestittel-synonymi' som "to tekster som blir yrkeskodet på samme måte". Å finne en synonymordliste vil gi flere bruksområder:

- Sammenlikne innkommende yrkestekst med tidligere tekst på samme arbeidstakerforhold, for å unngå å kode på nytt tilfeller der det meldes ny tekst med ikke yrkessignifikante endringer.
- Sammenlikne innkommende yrkestekst med kodelister for å kode flere irregulære tekster automatisk.
- Utarbeide lister som brukes som støtte for manuell koding.
- Videreutvikle kvalitetskontroller.

Programsystemets nåværende form inneholder skjønnsmessige valg for mange av de vanligste tilfeller. Strategien har vært å bygge ut gradvis etterhvert som datagrunnlaget har økt, og til en viss grad brukt frekvensdata (fra AKU) for å støtte de valg som er gjort. I 2002 foreligger en stor mengde arbeidstakerforhold som er kodet ferdig i 2001, og som kan utnyttes til statistisk begrunnede valg. Dette innebærer i korthet at man finner kriteriene for yrkeskoding ved å regne ut sannsynligheten for hver yrkeskodene gitt et sett kriterier. Hver unik tekst får en sannsynlighetsfordeling som igjen kan sammenliknes med andre teksters struktur. En synonymliste for yrker kan lages utfra dette ved å bestemme likhetsparametere mellom disse strukturene. For tekster som klassifiseres entydig til yrke, er synonymi elementær.

2.6.4.6 Analyser av arbeidsoppgaver

I AKU registreres arbeidsoppgaver i tillegg til yrkestittel. En databasert analyse av tekstmønstre i arbeidsoppgavene kan muligens brukes ved videreutvikling yrkeskodingssystemet. En anvendelse er søk etter yrkeskarakteriserende mønstre, ikke bare enkeltord. Ideen har kommet opp etter observasjon av tekster fra Arbeidstakerregisteret. Problemområder med AKU-data er mange missingverdier og uspesifikke ord, da det i AKU opplyses mest arbeidsoppgaver der tittelen ikke er beskrivende nok. I register forekommer det ikke sjelden at det skrives arbeidsoppgaver i tillegg eller istedenfor tekst. Felles for begge er at arbeidsoppgavene ikke beskrives systematisk, og mangler ofte nøkkeldata for yrkesklassifisering.

2.6.5 Dokumentasjon av kodingen av enkelte yrker

Automatisk koding av yrkestitler som bygger på næring og eller utdanning i tillegg til tekst, baserer seg på en lang rekke valg. Detaljer om dette finnes på følgende internettadresse **http://www.ssb.no/06/90/notat_200170/**
Standard for yrkesklassifisering og Yrkeskatalogen, samt søkeverktøy for yrkestitler finnes på
http://www.ssb.no/emner/06/yrke

3 Kontroller og revisjon av yrkeskoding

3.1 Bakgrunn

Før årskontrollen 2003 sendte yrkeskoder tilbake til arbeidsgivere ble det gjennomført en rekke undersøkelser av kvaliteten på yrkeskoding, og det ble gjennomført mange omkodinger som følge av kontrollene. Dette skjedde i perioden november 2002-januar 2003, og alle detaljene her er ikke dokumentert pga. tidspress. Det ble undersøkt over 10.000 stikkprøver av enkeltarbeidstakerforhold. En rekke arbeidstakerforhold ble omkodet i henhold til en skjønnsmessig vurdering, og ved hjelp av en rekke ad-hoc SAS-programmer. Det ble trukket et utvalg etter omkodingen, og en gjennomgang viste at 7.5% av kodene var usikre eller direkte feil.

I forbindelse med ny sysselsettingsstatistikk (fra og med fob2001) er det ønske om yrkeskoder på alle sysselsatte, noe som krever imputering av betydelige antall. Vi har prioritert å forbedre kvaliteten på yrkeskodingen i Arbeidstakerregisteret, da dette utgjør den største delen sysselsatte. De metoder som er omtalt her er det vi kan kalle registerinterne metoder, altså uten kobling til andre kilder for yrkeskoder.

3.1.1 Beskrivelse av variabelen yrke

Yrke er en spesiell variabel på flere måter. Yrke er knyttet til person, men går ofte på tvers av formell utdanning og har varierende sammenheng med bedriftens virksomhet. De nedenstående definisjonene av et 'yrkeshomogent stratum' gir bakgrunnen for ulike innfallsvinkler til undersøkelser av kvalitet på yrkeskoding.

3.1.1.1 Realdefinisjon

En gruppe som har ensartede arbeidsoppgaver, noenlunde samme kompetansenivå ved formell og uformell utdanning og grad av spesialisering ved bruk av: ferdigheter, verktøy, maskiner, materialer, varer, tjenester. Krever en egen undersøkelse for å analysere. Denne definisjonen ligger til grunn ved kommunikasjon med arbeidsgiver.

3.1.1.2 Ex-post-facto definisjon

En gruppe som i et gitt (uavhengig) datasett har lik yrkeskode. Eks. en modell for yrkeskode i AKU, man beregner sannsynligheten for at en person har et yrke, gitt verdien av en rekke variabelverdier. Denne definisjonstypen brukes ved metodene i notat av Leiv Solheim oktober 2002, og i utviklingsarbeidet beskrevet i kapittel 2.

3.1.1.3 Operasjonell definisjon

En gruppe hvor en gitt vektor av antatt yrkeskarakteriserende variabler har tilstrekkelig like verdier. Eks. tittel, næring, utdanning, bedriftsstørrelse. Altså man gjør følgende antagelse: hvis det i et stratum ikke er yrkeskode-homogenitet, så er yrkesfordelingen feil – ikke stratifiseringen. Det er denne definisjonen som brukes i kontrollene som er beskrevet nedenfor.

3.2 Manuelle kontroller av arbeidstakerforhold

En metode innen det manuelle revisjonsarbeidet gjennomføres på følgende måte:

1. Programmering for å identifisere grupper som skal kontrolleres.
2. Utlisting av detaljerte opplysninger om arbeidstakerforhold i mistenkelige grupper.
3. Gjennomgåelse av disse i faggruppen.
4. Fastsettelse av instruks for koding.
5. Punching av koder ved seksjon 450:
    a. Lister for maskinell omkoding.
    b. Ren manuell omkoding.
6. Analyse av endringer.
7. Implementering i yrkesregisterfiler for arkiv.

3.2.1 Identifisering av yrkeskoder til inspeksjon

1. Det selekteres antatt yrkeskarakteriserende variabler som er tilgjengelige i Arbeidstakerregisterfiler.
    a. Variablene velges utfra faglige vurderinger av yrkesstandardens beskrivelser av de enkelte yrker.
    b. Sammensetningen av variabler er slik yrkesfordelingen i et strata skal være homogen.
2. Arbeidstakerforhold som er yrkeskodet analyseres på de valgte variabler.
3. Det fastsettes kriterier for flagging til revisjon.

3.2.1.1 Eksempel fra tekstbasert koding

1. Data:
    c. har levert tekst
    d. har fått yrkeskode
4. Variabler:
    e. tekst (grupperes etter lydlikhet)[2]
    f. næring (5)
    g. utdanning (6)
    h. antall ansatte bedrift (grupper 0-4,5-9,10-...)
8. Kriterier:
    a. Sjeldne tekster holdes utenfor, eller gruppere sammen med liknende
    b. Små grupper holdes utenfor
    c. Inkonsekvente kan defineres f.eks. ved andeler mellom 5% og 95%.
    Dette utfra en ide hvor under 5% regnes som "smågrums" og over 95% regnes som konsekvent.
    d. Lister ut tekst (yrkestittel skrevet av arbeidsgiver), næring, utdanning, størrelsesgruppe, yrke, og antall.

"Smågrumset" er ment å skille ut tilfeldige småfeil, med ser at det utgjør ganske mange arbeidstakerforhold i testdata med de valgte grenseverdiene. Dette bør undersøkes nærmere. Det samme gjelder sjeldne yrker, hvor det er vanskelig å finne mønster for disse under ett, men hvor de enkelte yrker kan være av interesse for statistikkbrukere og oppdragsgivere. Begge disse forhold krever betydelig manuelt revisjonsarbeid.

3.3 Kontroller av yrkeskoding pr. bedrift

3.3.1 Undersøkelser av frafall

Tilfeldig frafall i Arbeidstakerregisteret har i utgangspunktet liten betydning for grupper av de størrelser som det er planlagt å lage statistikk for. Arbeidstakerregisteret er et forvaltningsmessig verktøy, og det er et krav fra RTV at alle arbeidsgivere skal levere yrkesdata på alle sine arbeidstakere. Frafall er derfor hvertfall et administrativt problem, i tillegg til de forhold som nevnes under. Med frafall i denne sammenheng menes her også andre typer problemer med inndata.

3.3.1.1 Manglende innlevering

Det mangler fortsatt endel yrkesdata, og det meste er fra bedrifter som ikke har levert noe yrkesdata. Skjevheter i dette materialet kan by på problemer, særlig i den grad frafallet avhenger av yrkesrelaterte kjennemerker. Det er ikke så mange bedrifter som har levert yrkesdata for bare deler av arbeidsstokken. Det er flere årsaker til mangler:

1. bedrifter som ikke leverer meldinger
2. meldinger/årskontroller som ikke blir punchet
3. maskinelle rutiner som ikke fungerer (lønnssystemer, etc.)

3.3.1.2 Uspesifikke titler

Bruken av ord som *vikar*, *assistent* og *konsulent* skaper problemer for yrkeskoding, særlig i næringer med uensartede arbeidsoppgaver. Næringsstandarden er blitt noe endret fra 2003, med noe mer detaljering i enkelte tertiærnæringer. Fortsatt er mange arbeidstakerforhold i næringer der arbeidsoppgavene er så varierte at de ikke kan yrkesklassifiseres utfra dette. Arbeidstakerforhold som blir levert med uspesifikk tekst vil ved årskontrollen i Arbeidstakerregisteret returneres med koden 0000100 og teksten "Angi yrkestittel mer fullstendig".

3.3.1.3 Uventede titler

Bruken av nye ord og engelske versjoner er økende. Dette kan til dels avhjelpes ved å lage nye koder, samt å hjelpe arbeidsgivere med oversettelser. Når arbeidsgiver leverer slik tekst uten å kontakte oss, får vi som regel en negativ reaksjon på koden/teksten som sendes tilbake også i de tilfeller der yrkeskoden (4 siffer) er korrekt.

Bruken av titler på uventede steder er langt mer problematisk, f.eks. at man kaller en som selger bensin for *stasjonsbetjent,* en tittel som normalt brukes innen om helt andre arbeidsoppgaver i jernbanenæringen. Forutsetningene for automatisk tekstbasert yrkeskoding er at teksten forteller noe om reelle arbeidsoppgaver. Når teksten ikke lenger er adekvat for dette formål, må andre metoder benyttes. En konsekvens er at stadig færre titler kan kodes 1-1. Denne metoden er i utgangspunktet den sikreste, raskeste og billigste.

Misvisende tekstbruk er vanskelig å oppdage, og kan gi betydelige systematiske feil. Mye er blitt rettet opp med manuell kontroller, og man kan i ettertid si at det kunne ha vært en like god ressursbruk å kode mer manuelt med en gang.

[2] Lydlikhetsfunksjonen har svakheter når det gjelder yrkestitler generelt, og brukes med forsiktighet ved innkoding. Her dreier det seg om strata som har store likheter (jf. valgte variabler), og da anses det forsvarlig å gruppere utfra denne funksjon.

Et annet problem er titler som tildels brukes om liknende arbeidsoppgaver, men som skal ha yrkeskoder på ulike yrkesfelt. F.eks. skal *salgskonsulent* og *selger* skal til yrkesfelt 3 og 5, mens enkelte andre selgere skal til 4. Det er grunn til å spørre om arbeidsgiver faktisk bruker yrkestitlene på den måte som er forutsatt.

3.3.1.4 Yrkeskoder levert av arbeidsgiver

Stadig flere arbeidsgivere leverer yrkeskoder istedenfor tekst. Denne koden kan være basert på yrkeskode fra Statistisk sentralbyrå utfra tidligere levert tekst, eller en yrkeskode arbeidsgiver har kommet fram til. Det er ifølge RTV stadig flere arbeidsgivere som leverer data til Arbeidstakerregisteret på diskett. Pr. mai 2003 er det omlag 2500 foretak med noe over 600.000 arbeidstakerforhold som leverer diskett, og som derfor skulle levert yrkeskode.

Når arbeidsgiver leverer yrkeskode er den maskinelle behandling enklere, men det foreligger ingen holdepunkter for arbeidstakerens reelle arbeidsoppgaver. Det må derfor kontrolleres mot avledede variabler, men her gjenstår mye arbeid.

3.3.1.5 Status for yrkesdata og frafall Arbeidstakerregisteret

Pr. uke 13-2003, er det 12% av arbeidstakerforholdene som mangler en brukbar yrkeskode.

3.3.2 Undersøkelser av avvik i yrkesstruktur

For kontrollformål er det interessant å se på yrkesfordelingen på enkeltbedrifter sammenliknet med andre bedrifter i samme strata (eks. nærings- og størrelsesgrupper). En slik sammenlikning kan identifiserer områder som bør undersøkes nærmere, uavhengig om det skyldes at enkelte bedrifter er blitt kodet feil, eller at bedriften har levert feil yrkeskoder. Nedenfor nevnes noen tilnærminger til undersøkelser på bedriftsnivå.

3.3.2.1 Avvik fra yrkesstruktur i den enkelte næring

Man kan beregne yrkesfordelinger pr. næring, og pr. bedrift, og sammenlikne andelene i hver bedrift med andelene i næringen bedriften tilhører. Det er flere forhold som må vurderes:

- En bedrift må være ganske stor for å kunne sammenliknes med totaltallene. De fleste bedrifter er små, slik at metoden vil utelukke mange bedrifter. På den annen side befinner de fleste arbeidstakerforhold seg i store bedrifter, slik at nytteverdien kan være stor.
- Det er store forskjeller mellom næringene hvor ensartede bedriftene er med hensyn til arbeidsoppgaver. I næringer hvor bedriftene er uensartede vil avvik i yrkesfordelingen kunne avspeile korrekte yrkeskoder. Avviksberegninger vil da gi for mange falske positive, og man kan vurdere å kutte ut en slik næring.
- I små næringer kan noen reelt avvikende bedrifter være årsaken til den store variasjonen, og vil ikke bli oppdaget hvis man utelukker næringen.

3.3.2.2 Beregning av yrkesfordeling og avvik

Vi begynner med å se på andeler av hvert enkelt yrke innen hver næring. Definerer den totale yrkesfordelingen etter næring som en endelig matrise:

$$\vec{P}_{næringer} = \begin{bmatrix} p\ (yrke_1)_1, p\ (yrke_2)_1, ..., p\ (yrke_n)_1 \\ p\ (yrke_1)_2, p\ (yrke_2)_2, ..., p\ (yrke_n)_2 \\ ... \\ p\ (yrke_1)_m, p\ (yrke_2)_m, ..., p\ (yrke_n)_m \end{bmatrix}$$

De gjeldende standarder gir n=353 og m=658.
Hver bedrift har kun èn næringskode *j* og i denne næringen er yrkesfordelingen:

$$\vec{P}_{næring=j} = \left[p\ (yrke_1)_j, p\ (yrke_2)_j, ..., p\ (yrke_n)_j \right]$$

Hver bedrift *k* har sin egne yrkesfordeling:

$$\vec{P}_{bedrift=k} = \left[p_k(yrke_1)\ , p_k(yrke_2)\ , ..., p_k(yrke_n)\ \right]$$

Vi kan da beregne et vektet avviksmål for bedrift *k* i næring *j* med X_k ansatte:

$$B_{bedrift=k} = X_k \cdot \sum_{i=1}^{n} \frac{(p_k(yrke_i)\ - p\ (yrke_i)_j)^2}{p\ (yrke_i)_j}$$

Vi kan da lage lister på detaljert nivå med mikrodata fra de bedrifter som viser størst avvik, innen egnede næringer. For å finne de uegnede næringer, kan vi se etter stor variasjonen av yrkesandelene innen hver næring – eller enklere: svært mange bedrifter som (tilsynelatende) er avvikende i en næring.

3.3.3 Avvik fra AKU-struktur

Det har også vært diskutert muligheten for å bruke AKU-data til å identifisere områder for mikrokontroller. Det er laget strukturtabeller utfra AKU-data som brukes som grunnlag for å liste ut arbeidstakerforhold med "problemyrker".

3.3.3.1 Datagrunnlag

I forsøkene ble det brukt AKU-data med yrkeskode 1-4 kvartal, 2000-2002, 12 perioder. Dette er valgt for å få en viss stabilitet og få med små yrker. Det foretas en lett revisjon av næringskoder. Det beregnes gjennomsnittlige andeler av hvert yrke i diverse stratifiseringer utfra næring.
Definerer to kontrolltabeller:

- yrker som overveiende finnes i få næringer.
- næringer hvor de ansatte har få ulike yrker.

Kobler dette tilbake til Arbeidstakerregister-data og finner bedrifter som har "mistenkelige" yrker i forhold til næringen.
Lister ut detaljerte opplysninger om bedriften og de enkelte arbeidstakerforhold:

- Navn på bedrift, næring, størrelse.
- Antall med levert tekst, utdanning, yrkeskode.

Slike lister er forsøkt med en viss nytte for å oppdage feilkodinger, og iverksette revisjon.

3.4 Manuell koding

3.4.1 Fordeler

Kvaliteten på manuell koding vil i mange tilfeller kunne bli av bedre kvalitet enn automatisk koding. Eksempler på forhold gjøre at manuell koding er bedre for en viss mengde arbeidstakerforhold:

- Vurdering av semantiske forhold f.eks. synonymer, hvor automatiske metoder kommer til kort.
- Helhetsinntrykk av et arbeidstakerforhold sett utfra flere variabler enn ved automatisk koding. Eksempler på dette er alder, bedriftsnavn, organisasjonsstruktur.
- Oppdage feil og mangler i inndata.
- Ved problematiske titler som er kodet automatisk og senere er korrigert for feilkodinger, ville det vært raskere å kode manuelt med en gang. Kodeprogrammet justeres derfor slik at noen flere titler kodes manuelt.

3.4.2 Feilkilder

3.4.2.1 Inndata

Tekst som sendes til manuell koding er i utgangspunktet vanskeligere å kode enn det som gjøres automatisk, med alle de problemer som er nevnt i pkt. 3.1.2 og –3

3.4.2.2 Systematiske feil

Uvaner, feil i instruks, misforståelser av yrkesstandarden kan gi systematiske skjevheter. Eksempel på dette er:

- Koding til de største grupper fordi man husker de best.
- Koding av uspesifiserte til "særgruppene" (xxx9-kodene) som en samlegruppe.

3.4.2.3 Tilfeldige feil

Lesefeil, punchefeil (manuelle feil) vil sannsynligvis være tilfeldig fordelt. Tilfeldig fordeling betyr ikke jevn fordeling, så ved stor feilprosent og/eller små grupper vil dette kunne få spesielle utslag.
All manuell koding skjer ved PC. Tastaturforhold kan gi muligens gi noen systematiske utslag. Kontroller og oppslagsfunksjoner minsker vel dette i stor grad. Det er gjort et betydelig arbeid med brukergrensesnittet ved s450.

3.4.3 Kvalitetskontroll

Det er gjennomført forsøk på analyser av manuell koding liknende de beskrevet i punkt 2. Det er også skilt ut ved konsistensanalyser mot AKU-data, for å se på forskjellene mellom automatisk og manuell koding. Det er flere ulemper ved disse metodene, det vesentligste er forskjellen i inndatamaterialet som beskrevet i notat av Leiv Solheim oktober 2002.
En mer vanlig reliabilitetsanalyse av manuell yrkeskoding kan gjennomføres på følgende måte:

1. Trekker tilfeldig et utvalg arbeidstakerforhold som er manuelt kodet.
2. Lager et nytt datasett av disse, separerer yrkeskodene.
3. Sender på nytt til manuell koding.
4. Sammenlikner de to yrkeskoder pr. arbeidstakerforhold ved en eller flere metoder:
   - Binær korrelasjonskoeffisient.
   - Sammenlikner ulike nivåer (1-4 siffer).

En billigere variant er å trekke parvis "like" arbeidstakerforhold som allerede er kodet. To problemer:

- Det er vanskelig å finne tilstrekkelig mange som er tilstrekkelig like.
- Har ikke kontroll med hvem som har kodet, og når de er kodet.

3.4.3.1 Prosedyrer

Man kan beskrive den nåværende manuelle yrkeskodingen som streng på den måten at man enten setter en 7-sifret kode eller ingen kode (0000100). Dette har ulemper hvorav noen er:

- Forsterker de systematiske feilene som nevnt i pkt. 4.2.2
- Flere blir kodet uoppgitt enn kanskje nødvendig

Hensikten med dette i forhold til "normal koding" (som f.eks. etterarbeid ved en undersøkelse) er å mobilisere arbeidsgivere til å levere en gyldig kode til Arbeidstakerregisteret. Det er derfor ikke aktuelt å kode uspesifikke yrkestitler til et mer overordnet nivå, med tilbakesending til Arbeidstakerregisteret.
Det som kunne vært aktuelt er å gi en foreløpig kode på et mer overordnet nivå i første omgang, for å la 2.linje-kodere forsøke en nærmere koding. En annen mulighet er at 1.linje-koderen markerer hvert arbeidstakerforhold i hvilken grad man er sikker på koden eller ikke. Denne markeringen kan brukes av 2.linje-kodere, eller databaserte analyser.

3.4.3.2 Kodelister

Yrkeskatalogen er offisiell publikasjon i **notat**-serien. Dette har lagt en del restriksjoner på hvilke titler som tas med og hvordan teksten formuleres her. Yrkeskatalogen er mer tenkt som et oppslagsverk for arbeidsgivere og andre enn for de som koder yrke. Den er med andre ord ikke det som kalles "coding index" eller kodeliste. En praktisk kodeliste vil skille seg fra den offisiell Yrkeskatalogen på en del punkter som kan ha betydning for manuell koding, og eventuelt annen yrkeskoding :

- tekster som svarer til det som forekommer i inndata
- nøkkelord, ikke formelle titler
- mer diskriminant beskrivelse
- restgrupper (andre innen ...)
- overordnete grupper (..., uspesifisert)

Yrkeskatalogen fyller bare delvis disse funksjonene, og i automatisk koding er det laget en del andre lister som ikke er direkte egnet for manuell koding. Utvikling av en kodeliste som er direkte formålsrettet for manuell koding ville derfor være ønskelig. Det foreligger et betydelig datamateriale som kan utnyttes i dette arbeidet.

4 Levering av tekst i yrkesdata til Arbeidstakerregisteret

4.1 Innledning

Dette er en oppsummering om tekstdata som er mottatt i "yrke"-feltet på meldinger til Arbeidstakerregisteret i tiden uke 8-2000 – uke 26-2003. Når årskontrollen til Arbeidstakerregisteret for 2003 er levert inn, regner vi med at det vanligvis vil bli levert yrkeskode til Arbeidstakerregisteret, og at tekst i mye mindre grad vil benyttes. Det er derfor greit å foreta en analyse av den tekst som er levert. Denne oppsummeringen forholder seg kun til teksten, og berører i liten grad selve yrkeskodingen (yrkesklassifisering utfra tekst).

```
Antall meldinger mottatt med tekst:   2.462.669
Antall distinkte person/tekst:        1.884.631
```

Mange personer forekommer flere ganger pga. av andre endringer enn yrkestittel. I den videre analyse brukes datasett med distinkte person / tekst kombinasjoner.
Over 22% av personene finnes med to ulike tekster. Dette omfatter både korrigeringer, feilstavinger og reelle endringer i yrke, samt ulike arbeidstakerforhold.
Nær 80% av tekstene består kun av 1 ord. Bruk av flere enn 2 ord er ganske sjelden (4%)
De vanligste tekstene og enkeltord er stort sett innen de vanligste yrkene også, selv om det nok ikke er tilfeldig hvilke bedrifter som leverer tekst og hvilke som leverer kode.. Det som er mest påfallende er at selv de aller vanligste ord og tekster utgjør en svært liten del av hele massen (max 3%). Dette sier mye om hvor stor variasjon og hvor mange muligheter det er når man arbeider med tekst. En nærmere analyse viser et typisk mønster når det gjelder tekst: få ord forekommer ofte, de fleste forekommer svært sjeldent. F.eks. 95% av ordene forekommer færre enn 50 ganger, minst halvparten forekommer kun 1 gang. Antall ulike tekster: 163.931, antall ulike ord: 59.876.
Det er stor forskjell i hvilke ord som er vanligst på de ulike plassene i teksten. Dette har betydning for søk etter titler og yrkeskoding utfra tekst.

Kommentar i forhold til skillet mellom 'tekst' og 'ord':
I tabellene opptrer "butikkmedarbeider" som 1 ord mens "daglig leder" er 2 ord. Dette skyldes skrivemåten (med mellomrom) og er egentlig et kunstig skille i denne analysen. Enten så burde ord som "butikkmedarbeider", "bussjåfør" vært splittet, eller uttrykk som "daglig leder", "formann verksted" burde vært samlet. Problemet har vært at dette vanskelig kan gjøres automatisk.

4.2 Tekstbruk i forhold til yrkeskatalogen

Arbeidstakerregisteret: antall ulike tekster: 163.931 antall ulike ord: 59.876

Yrkeskatalogen: antall titler: 5.613 antall ulike ord: 4.409
Ord som forekommer begge steder: 3.155 (5% av leverte ord)
Ord som kun forekommer i Arbeidstakerregisteret: 56.721
Ord som kun forekommer i Yrkeskatalogen: 1.254 (28% av distinkte ord i katalogen)

4.2.1 Utnyttelse av tekstfeltet "yrke"

I RTVs database over Arbeidstakerregisteret var det satt av 40 tegn til registrering av yrkestittel. Gjennomsnittlig lengde på teksten er 13 tegn. De aller fleste (over 90%) har benyttet kun halvparten av den tilgjengelige plassen, 99% har brukt ¾ eller mindre. Bruk av diverse tegn: 312.268 titler (17%) inneholder andre tegn enn mellomrom og bokstaver.

4.2.2 Metoder for å gruppere liknende tekster

Tabellene i innledningen viser at det er stort behov for å gruppere tekster. Den enorme variasjonen skyldes ikke bare språklig forhold som ved normale tekstbaserte metoder. En stor del skyldes at det ikke foretas noen kontroll av inndata ved registrering på trygdekontorene. Med enkle midler for inndatavalidering ville man fått en ensartet staving, og mange flere arbeidstakerforhold kunne kodes automatisk. Vi vet ikke årsakene til manglende på datavalidering på registreringssiden, men det kan i ettertid vurderes om totalkostnadene kunne vært lavere med en mer rigid kontroll av inndata.

De metodene som nevnes her er kun basert på teksten selv. Se avsnitt 2.6.2 for metoder for å måle om ord som brukes har samme betydning utfra yrkeskoding og eventuelt andre variabler. Ved å benytte 'soundex'-funksjonen i SAS reduseres antallet tekstgrupper fra 163931 til 102085, hvor den gjennomsnittlige frekvensen øker fra 11.5 til 18.5. Det er med andre ord en viss forbedring, og funksjonen har stor nytte i kombinasjon med andre variabler. Benyttet alene vil 'soundex' kunne gi følgende feilmuligheter:

- unnlater å gruppere ord som gir samme yrkeskode
- grupper ord som skulle gitt forskjellig yrkeskode

Dette gjelder også andre metoder.

Ved 'Spedis'-funksjonen i SAS kan man, i motsetning til 'soundex', justere nivået for grupperingen. En viktig problemstilling er å finne en optimal grense slik at man minimaliserer både type 1- og type 2 feil som nevnt ovenfor. Dette er vanskelig av minst to grunner:

- en kort staveavstand kan gi stor semantisk avstand.
- det er ulike grenser for ulike ord.

I kodesystemet brukes slike metoder bare i kombinasjon med andre kjennemerker eller ytterligere kontrollrutiner.

5 Forsøk med estimering av antall ledere

Dette avsnittet bygger på en oppgave til et kurs i statistiske metoder (SM03/2003), og gjennomgår metoder som kan videreutvikles i forbindelse med produksjon av yrkestatistikk. Andre utviklingsoppgaver som kan være aktuelle er generelt kvalitetsmål på yrkesvariabelen i Arbeidstakerregisteret og imputering av yrke i Registerbasert sysselsettingsstatistikk. Metodikken bygger på kursnotater av Leiv Solheim 2003.

5.1 Faglig bakgrunn

Arbeidstakerregisteret skal inneholde yrkeskode for alle arbeidstakerforhold. Det er foreløpig ikke laget statistikk utfra dette fordi det i oppstartfasen er usikker kvalitet, pga. frafall og andre forhold. Jeg ønsker å finne et mål på usikkerheten pga. frafall ved å beregne varians for en begrenset gruppe. Jeg ønsker også å finne to estimatorer for å sammenlikne metodene. I AKU publiseres tall for yrkesfelt pr. fylke. Fra Arbeidstakerregisteret er det aktuelt å publisere tall for kommune. Oppgaven avgrenses til å finne antall personer i yrkesfelt 1 (adm. ledere) pr. kommune, fylke og totalt, og anslå usikkerheten for disse tallene.
Data er gyldige hoved- og biarbeidstakerforhold i Arbeidstakerregisteret periodefil uke 13-2003. Tallene vil da tilsvare aktive arbeidstakerforhold ved utgangen av 2002. Yrkesdata er koblet ved en sikker identifikator, i motsetning til en del andre forsøk ved påkobling av yrkesdata. Dette gjør at frafallet blir noe høyere, pga. endringer av orgnr., og feil i datoer.
Typiske grupperingsvariabler for denne undersøkelsen vil være geografiske områder, kjønn, alder, næring, utdanning, bedriftsstørrelse.Fordelen med å stratifisere er mindre total usikkerhet så lenge ikke stratumsstørrelsen blir altfor liten. Man må forsøke å identifisere grupper som har stor innbyrdes likhet og utvendig store skiller. Dette kan operasjonaliseres ved at variansen skal være liten innen gruppen, og de beregnede tallene skal være forskjellige mellom gruppene.

5.2 Estimering og variansberegning

Det teoretiske variansen for estimatet er avhengig av følgende faktorer:

- Den absolutte størrelsen på utvalget, her: de som har levert yrkeskode til register.
- Utvalgets størrelse i forhold til populasjonen, her: totalregister.
- Andelen innen hver yrkeskategori.
- Den estimerte variansen for andelen.
- For ratemodellen: den reelle variansen av ratevariabelen.

Den største usikkerheten får vi ved små grupper, uensartede grupper, stort frafall og andeler nær 50%.

5.3 Modell 1

Modellen beregner antall ledere i et stratum utfra hvor mange som har yrkesdata, altså leveringsgraden. Metodene brukes som om leverte yrkesdata er et utvalg og hele registeret er populasjonen. Vi kjenner ikke mekanismene bak frafall, men regner i første omgang med at det er tilfeldig – selv om det fra en registerfaglig synspunkt neppe er holdbart. Definerer:

- $T_h \equiv$ antall ledere i stratum $h \in H$
- $L_h \equiv$ antall arbeidstakerforhold med leder-yrkeskode i stratum h
- $n_h \equiv$ antall arbeidstakerforhold med yrkeskode i stratum h

  $n = \sum_{h=1}^{\eta} M_h$ tilsvarer "antall i utvalget", $\eta = |H|$ er antall strata.
- $N_h \equiv$ antall arbeidstakerforhold totalt i stratum h

  $N = \sum_{h=1}^{\eta} N_h$ tilsvarer "antall i populasjonen"

Vi definerer reelt antall ledere i stratum h:

$$T_h = \sum_{i=1}^{N_h} Y_{i,h}$$

for en dummyvariabel Y gitt ved:

$$Y_{i,h} = \begin{Bmatrix} 1 \ hvis \ person \ \ i \ er \ leder \\ 0 \ hvis \ ikke \end{Bmatrix}$$

Modell 1 kan formaliseres på følgende måte for estimat av antall ledere:

$$\hat{T}_h = \sum_{i=1}^{n_h} Y_{i,h} + \sum_{i=n_h+1}^{N_h} Y_{i,h}$$

Andel ledere i leverte data ("utvalget") er:

$$\hat{p}_h = \frac{\sum_{i=1}^{n} Y_{i,h}}{n_h}$$ og estimat for totalen i stratum h er

$$\hat{T}_h = \sum_{i=1}^{n_h} Y_{i,h} + \sum_{i=n_h+1}^{N_h} \hat{p}_h = N_h \cdot \hat{p}_h = \frac{N_h}{n_h} \cdot L_h$$

I praksis regner vi ut:

$$\hat{T} = \sum_{h=1}^{\eta} \hat{T}_h = \sum_{h=1}^{\eta} \left(\frac{N_h}{n_h} \cdot L_h\right)$$

Sammenlikner to enkle stratifiseringer:

- Etter fylke, og bruker samme vekt for alle kommunene i et fylke. Registreringen av yrkesdata ved Trygdeetaten skjer fylkesvis, slik at det vil være regionale forskjeller i mengden registrerte yrkeskoder. En begrunnelse for en slik stratifisering er at det blir et håndterlig antall strata, og ingen strata blir ekstremt små.
- Vekter for hver kommune. Fordelen med dette er at det faglig sett bør gi bedre kommunetall, fordi leveringsgraden kan være ganske forskjellige i noen kommuner innen samme fylke. Det gjør at variansen for hvert kommunetall er mindre. En ulempe er at det i endel tilfeller blir svært små tall, som er mer følsomme for ikke-tilfeldige faktorer.

Variansen beregnet utfra forutsetning om tilfeldig frafall:

$$\text{var}(T_h - \hat{T}_h) = \frac{N_h(N_h - n_h)}{n_h} \cdot \sigma_h^2$$

hvor standardavviket estimeres:

$$\hat{\sigma}_h^2 = \frac{\hat{T}_h}{N_h} \cdot \left(1 - \frac{\hat{T}_h}{N_h}\right)$$

Tabell 5.3-1 Resultat pr. fylke

	Arb.taker	Med yrke	Med lederkode	Vekt	Estimert ledere	Estimert andel ledere	Varians	std	rel.std
01 ØSTFOLD	106036	71596	3971	1.481032	5881	6%	2672	52	0.9%
02 AKERSHUS	205952	161746	13645	1.273305	17374	8%	4348	66	0.4%
03 OSLO	397477	318058	32613	1.2497	40756	10%	9133	96	0.2%
04 HEDMARK	75156	62194	3718	1.208412	4493	6%	880	30	0.7%
05 OPPLAND	73934	66737	3816	1.107841	4228	6%	430	21	0.5%
06 BUSKERUD	103869	91449	6256	1.135813	7106	7%	899	30	0.4%
07 VESTFOLD	90633	81202	5651	1.116142	6307	7%	682	26	0.4%
08 TELEMARK	68118	59368	3798	1.147386	4358	6%	601	25	0.6%
09 AUST-AGDER	39403	34313	2129	1.14834	2445	6%	340	18	0.8%
10 VEST-AGDER	67002	59071	3792	1.134262	4301	6%	540	23	0.5%
11 ROGALAND	183872	152098	9190	1.208905	11110	6%	2181	47	0.4%
12 HORDALAND	201484	173135	10685	1.163739	12435	6%	1910	44	0.4%
14 SOGN OG FJORDANE	47171	39640	2119	1.189985	2522	5%	453	21	0.8%
15 MØRE OG ROMSDAL	102964	87259	5611	1.179981	6621	6%	1115	33	0.5%
16 SØR-TRØNDELAG	126849	113666	6474	1.11598	7225	6%	790	28	0.4%
17 NORD-TRØNDELAG	49311	40005	2251	1.232621	2775	6%	609	25	0.9%
18 NORDLAND	94912	77517	4806	1.224402	5884	6%	1239	35	0.6%
19 TROMS	70693	50559	3095	1.398228	4328	6%	1618	40	0.9%
20 FINNMARK	29579	23445	1629	1.261634	2055	7%	500	22	1.1%
	2134415	1763058	125249		152202	7%			

Tabell 5.3-2 Resultat pr. kommune (vist Østfold), med estimat etter fylke-vekt og kommune-vekt.

	Arbeids takere	Fylke	Kommune	Ledere estimat 1	Ledere estimat 2	Forskjell	Rel. Std (for nr.2)
0101 Halden	11876	1.481032	1.491772	628	633	1 %	2.71
0104 Moss	13721	1.481032	1.43992	834	811	3 %	2.26
0105 Sarpsborg	21765	1.481032	1.432851	1 134	1 098	3 %	1.94
0106 Fredrikstad	33240	1.481032	1.566742	1 724	1 824	6 %	1.71
0111 Hvaler	571	1.481032	1.590529	27	29	7 %	14.00
0118 Aremark	298	1.481032	1.806061	6	7	20 %	33.00
0119 Marker	1138	1.481032	1.814992	70	85	20 %	9.40
0121 Rømskog	166	1.481032	1.307087	9	8	12 %	19.31
0122 Trøgstad	1106	1.481032	1.215385	78	64	20 %	5.61
0123 Spydeberg	1350	1.481032	1.478642	129	129	0 %	5.80
0124 Askim	5327	1.481032	1.384356	298	278	7 %	3.62
0125 Eidsberg	4338	1.481032	1.459623	228	225	1 %	4.40
0127 Skiptvet	704	1.481032	1.259392	34	29	16 %	9.27
0128 Rakkestad	2487	1.481032	1.608668	116	125	8 %	6.79
0135 Råde	1763	1.481032	1.60711	102	111	8 %	7.16
0136 Rygge	4416	1.481032	1.351285	355	324	9 %	3.17
0137 Våler	1018	1.481032	1.39071	67	63	6 %	7.65
0138 Hobøl	752	1.481032	1.303293	43	38	13 %	8.73

5.3.1 Kommentar til resultatene

Ser at det er en god del forskjeller i tallene avhengig av hvilken stratifisering som brukes. De kommunene som er mest forskjellig fra gjennomsnittet vil ha de største avvikene. En typisk fordeling i norske fylker er at de har noen få kommuner som er mye større en gjennomsnittet og mange små. For de fleste fylker vil det derfor gi tilsvarende utslag som for det viste (Østfold). Når ulikheten i estimatene overskrider den usikkerhet som hvert estimat har, så er det grunn til å nøye vurdere hvilken estimeringsmetode som skal brukes. En kan ikke si på bakgrunn av dette forsøket at en geografisk basert modell er særlig egnet.

5.4 Modell 2

Denne modellen estimerer antall ledere ved en ratemodell, utfra ansatte pr. foretak. Gjennomsnittlig andel ledere beregnes, og stratifiseringen er i 4 størrelsesgrupper, da en utfra yrkesklassifiseringen postulerer at lederandelen er forskjellig i store og små bedrifter. Det er imidlertid ikke enkelt å finne naturlige avgrensinger, velger følgende strata: 5-9, 10-99, 100-999, 1000+ .

Foretak med færre enn 5 ansatte tas ikke med i beregningene, enten de har levert lederkoder eller ikke. Dette utfra en operasjonell definisjonen av "administrativ leder" der man gjør antagelser om arbeidsoppgavene utfra antall ansatte.

Definisjoner:

F_h = foretak som har levert yrkeskode på de fleste av sine arbeidstakerforhold

L_h = antall arbeidstakerforhold med leder-yrkeskode i stratum h der $f \in F_h$

$\tilde{y}$ = gjennomsnittlig andel ledere i stratum h der $f \in F_h$

k_h = antall foretak i stratum h der $x \notin F_h$

$Y_{f,h}$ = antall ledere i foretaket f

$x_{f,h}$ = antall arbeidstakerforhold i foretaket f

X_h = antall arbeidstakerforhold i stratum h

x_n = antall arbeidstakerforhold i foretak som har levert yrkeskode på de fleste av sine arbeidstakerforhold

Modell:

$$Y_{f,h} = \beta_h \cdot x_{f,h} + \varepsilon_{f,h}$$

Estimerer raten:

$$\hat{\beta}_h = \frac{\sum Y_{f,h}}{\sum X_{f,h}} = \bar{y}_h$$

Formelt estimat av totalt antall ledere i stratum h:

$$\hat{T}_h = \sum_{f \in F} Y_{f,h} + \sum_{f \notin F} \hat{Y}_{f,h} = \hat{\beta}_h \cdot X_h$$

Regner ut estimatet i praksis:

$$\hat{T}_h = L_h + \sum_{f=1}^{k_h} \bar{y}_h \cdot x_{f,h}$$

Estimering av varians (til differansen av estimat og reelt tall) som et mål på modellusikkerheten:

$$\text{var}(T_h - \hat{T}_h) = X_h^2 \cdot \frac{X_h - x_h}{X_h} \cdot \frac{\sigma_h^2}{x_h}$$

hvor standardavviket må estimeres utfra standardavvik i "utvalget":

$$\hat{\sigma}_h^2 = \frac{1}{F-1} \cdot \sum_{f=1}^{F} \frac{(Y_{f,h} - \hat{\beta}_h x_{f,h})^2}{x_{f,h}}$$

5.4.1 Resultater

Tabell 5.4-1 Estimert antall ledere etter fylke

	Arbeidstakere	Ledere	Andel ledere
I alt	1 976 557	154 819	8 %
01 ØSTFOLD	97 496	7 235	7 %
02 AKERSHUS	189 816	15 919	8 %
03 OSLO	372 841	34 413	9 %
04 HEDMARK	68 830	5 104	7 %
05 OPPLAND	67 728	4 901	7 %
06 BUSKERUD	95 469	7 572	8 %
07 VESTFOLD	83 072	6 636	8 %
08 TELEMARK	62 965	4 847	8 %
09 AUST-AGDER	36 024	2 798	8 %
10 VEST-AGDER	61 642	4 731	8 %
11 ROGALAND	172 074	11 992	7 %
12 HORDALAND	187 782	13 432	7 %
14 SOGN OG FJORDANE	43 040	3 084	7 %
15 MØRE OG ROMSDAL	94 997	7 240	8 %
16 SØR-TRØNDELAG	118 497	8 316	7 %
17 NORD-TRØNDELAG	44 614	3 315	7 %
18 NORDLAND	87 297	6 577	8 %
19 TROMS	65 367	4 590	7 %
20 FINNMARK	26 884	2 108	8 %

Tabell 5.4-2 Estimert antall ledere i Østfoldkommuner, sammenliknet med forrige modell

	Modell 1			Modell 2		
	arb.taker	leder	andel	arb.taker	leder	andel
0101 Halden	11876	633	5.3 %	10 900	809	7.4 %
0104 Moss	13721	811	5.9 %	12 700	960	7.6 %
0105 Sarpsborg	21765	1 098	5.0 %	20 255	1 317	6.5 %
0106 Fredrikstad	33240	1 824	5.5 %	31 062	2 258	7.3 %
0111 Hvaler	571	29	5.1 %	452	44	9.8 %
0118 Aremark	298	7	2.3 %	224	21	9.3 %
0119 Marker	1138	85	7.5 %	1 001	91	9.1 %
0121 Rømskog	166	8	4.8 %	143	12	8.4 %
0122 Trøgstad	1106	64	5.8 %	931	79	8.5 %
0123 Spydeberg	1350	129	9.6 %	1 182	139	11.8 %
0124 Askim	5327	278	5.2 %	4 830	358	7.4 %
0125 Eidsberg	4338	225	5.2 %	3 908	307	7.9 %
0127 Skiptvet	704	29	4.1 %	596	38	6.5 %
0128 Rakkestad	2487	125	5.0 %	2 169	186	8.6 %
0135 Råde	1763	111	6.3 %	1 566	136	8.7 %
0136 Rygge	4416	324	7.3 %	4 008	363	9.1 %
0137 Våler	1018	63	6.2 %	926	70	7.5 %
0138 Hobøl	752	38	5.1 %	643	46	7.2 %

De ujusterte tallene for arbeidstakerforhold i modell 2 er noe lavere enn for modell 1 og publiserte tall for arbeidstakere. Hvis denne metoden skal anvendes for statistikkformål, kan man kalibrere slik at kommunetotalene stemmer med øvrig registerbasert statistikk. Det er også mulig å kalibrere yrkesfelt-tallene slik at den stemmer på publiserte tall i AKU, men da kun på fylkesnivå. Alle kommuner i et fylke kalibreres da med samme faktor for hvert yrkesfelt. Fordelen er at vi gir et mer ensartet bilde overfor brukerne, ulempen er å bringe inn flere usikkerhetsmomenter. Det er minst to kilder til usikkerhet: variansen av estimatet fra AKU og skjevheten mellom kommunene innen et fylke i yrkesdatalevering.

5.4.2 Vurdering av modell 2

Estimerer av standardfeil som et mål på modellusikkerheten:

$$\text{std}(T_h - \hat{T}_h) = X_h \cdot \sqrt{\frac{X_h - x_h}{X_h}} \cdot \frac{\sigma_h}{\sqrt{x_h}}$$

Tabell 5.4-3 Estimert antall ledere i hvert stratum, og beregnet usikkerhet

	Ledere est. andel	Standardavvik	Sum utvalg	Sum populasjon	Ledere est. total	Standardfeil	Variasjons-koeffisient
Totalt	0.784457	.	1550524	1976557	.		
Gruppe 5-9	0.194481	0.29632	69622	160255	31166.57	135.34	0.43 %
Gruppe 10-99	0.092221	0.49968	439900	570479	52610.11	205.62	0.39 %
Gruppe 100-999	0.05666	0.91887	624614	700839	39709.35	268.72	0.68 %
Gruppe 1000+	0.032226	1.61027	416388	544984	17562.52	660.62	3.76 %

Standardavviket i hver gruppe forteller hvor stor spredning ratevariabelen har. Standardfeil er et mål på usikkerheten i estimatet. At dette tallet er lavt forteller at modellen gir lav usikkerhet på det estimerte antall ledere, forutsatt at antall ledere reellt sett er en funksjon av antall ansatte.

Det er flere momenter som må vurderes for å gi et totalt inntrykk av kvaliteten her:

1. Modell 2 kan sies å gi et mer "yrkesfaglig" korrekt tall, da man kutter ut små foretak. Dette henger sammen med definisjonen av 'administrative ledere' hvor man forutsetter et visst minimum av ansatte for å anta at arbeidsoppgavene er overveiende administrative.
2. For andre yrkesfelt vil nok en stratifisering utfra en eller annen næringsinndeling være nødvendig. Dette fordi andre yrkesfelt enn ledere vil være mer avhengig av bedriftens produksjonen enn størrelsen.
3. Ved å dele inn etter både regionale, størrelsesmål og næring, står man i fare for å *overstratifisere.* Det betyr at man får svært små grupper, og mange tomme strata, noe som er metodisk problematisk.
4. Å modellere utfra levering/frafall forutsette et tilfeldig utvalg. Estimatene er derfor beheftet med større usikkerhet enn det som framkommer i disse beregningene. Dette på grunn av skjevhet i frafallet, og feil i yrkeskodingen som til sammen kan være mer utslagsgivende enn ulikheter i modellene og usikkerheten i en valgt modell.

5.5 Vurdering av modellene

For å kunne sammenlikne modellene, må vi finne ut om:

- de gir forskjellige tall, og i så fall hvilke tall som er mest riktig.
- hvilken modell som gir minst usikkerhet, uttrykt ved variansen til estimatet.
- om forskjellen mellom modellene er merkbart større en usikkerheten i tallene selv.

Formelt sett kan vi definere:

- Hvilken modell har minst usikkerhet:

$$u = \text{Min}(\ \text{var}(T_{\text{mod}\,I,h} - \hat{T}_{\text{mod}\,I,h})\ ,\ \text{var}(T_{\text{mod}\,II,h} - \hat{T}_{\text{mod}\,II,h})\ )$$

- Hvor stor forskjell er det på modellene, her relatert til størrelsen:

$$v = \frac{2(\hat{T}_{\text{mod}\,I,h} - \hat{T}_{\text{mod}\,II,h})}{\hat{T}_{\text{mod}\,I,h} + \hat{T}_{\text{mod}\,II,h}}$$

- Forskjellen mellom modellene i forhold til usikkerheten:

$$w = v - 2\sqrt{u}$$

Hvis $w > 0$ kan man si at forskjellen på modellene sannsynligvis skyldes noe annet enn tilfeldigheter.
På grunn av at den siste modellen kutter ut de minste foretakene, er ikke en direkte sammenlikning mulig på det nåværende stadium. Det mest metodisk korrekte er i første omgang å kutte ut de samme foretakene og kjøre modell I på samme gruppe. En annen metode ville vært å kalibrere tallene i modell II, men man bringer inn en nye usikkerhetsfaktorer som f.eks. forskjeller i yrkesfordelingen i bedrifter av ulik størrelse.

Et eget notat om justering for partielt frafall og estimering av yrkesandeler er planlagt utgitt høsten 2003.

De sist utgitte publikasjonene i serien Notater

2003/48 E. Siig Meen og O. Rognstad: Jordbrukstelling 1999- dokumentasjon. 105s.

2003/49 L.Rogstad: Statistiske temakart og X-Map. 32s.

2003/50 E. Holmøy: Velferdsregnskap - et mulig teoretisk rammeverk.35s.

2003/51 C. Wiecek: Undersøkelse om fremtidsplaner, familie og samliv. Dokumentasjonsrapport. 59s.

2003/52 KOSTRA: Arbeidsgrupperapporter 2003. 153s.

2003/53 A. Haglund: Rapport fra arbeidsgruppa om forslag til arbeidsdeling mellom Brønnøysundregistrene (BR) og Statistisk sentralbyrå (SSB). 40s.

2003/54 E. Eng Eibak: Forventningsindikator - konsumprisene. Mai - november 2003. 19s.

2003/55 G. Daugstad: Levekår for ungdom i større byer. 80s.

2003/56 A. Vedø og D. Rafat: Sammenligning av utvalgsplaner i AKU. 17s.

2003/57 L. Belsby: Frafall og vekter i Tidsbruksundersøkelsen 2000-2001. 20s.

2003/58 L.Belsby: Vekter i Forbruksundersøkelsen. 28s.

2003/59 M. Mogstad og L.C. Zhang: På veien fra familie- til husholdningsregister. En metode for prediksjon av samboere uten barn .53s

2003/60 A. Vedø og D. Rafat: Redigering av husholdningsfilen fra Kvalitetsundersøkelsen. 13s.

2003/61 M. Mogstad: Analyse av fattigdom basert på register- og folketellingsdata. 75s.

2003/62 T. Eika og J.A. Jørgensen: Makroøkonomiske virkninger av høye strømpriser i 2003. En analyse med den makroøkonometriske modellen KVARTS.16s

2003/63 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 2001. 32s.

2003/64 E. Røed Larsen og D.E. Sommervoll: Til himmls eller utfor stupet? En katalogisering av forklaringer på stigende boligpriser. 31s.

2003/65 P.E. Tønjum: Tilbakemelding/ dokumentasjon av prosjektet: Avstemming av KNR mot nye årstall ifølge tallrevisjonen.43s.

2003/66 B.A. Holth: Arbeids- og bedriftsundersøkelsen 2003. Dokumentasjon. 67s.

2003/67 H. Tønseth: Kommuneale helseforskjeller -de finnes, men kan de måles? 15s.

2003/68 T.M. Normann: Omnibusundersøkelsen mai/juni 2003. Dokumentasjonsrapport. 50s.

2003/69 KOSTRA (Kommune- Stat- Rapportering) Rutinebeskrivelse og dokumentasjon. 60s.

2003/70 E. Holmøy og B. Strøm: Fordeling av tjenesteproduksjon mellom offentlig og privat sektor i MSG-6. 25s.

2003/71 J.K. Dagsvik: Hvordan skal arbeidstilbudseffekter tallfestes? en oversikt over den mikrobaserte arbeidstilbudsforskningen i Statistisk sentralbyrå. 67s.

2003/72 A. Steinkellner: Inntektsstatistikk for personer og familier 1999-2001. Dokumentasjon av datagrunnlag og produksjonsprosess. 43s.

2003/73 F. Tverå, I. Sagelvmo: Beregning av næringene fiske eget bruk, fiske og fangst og fiskeoppdrett i nasjonalregnskapet. 19s.

2003/74 K.H. Grini: Lønnsstatistikk privat sektor 1997-2001. Dokumentasjon av utvalg og beregning av vekter. 36s.

2003/75 A.H. Foss: Grafisk revisjon av nøkkeltallene i KOSTRA. 16s.

2003/76 K. Hansen: Ideelle organisasjoner i nasjonalregnskapet. 30s.