Afsnit III

Udpakning

Udpakning første fase.

I bilag I) 5. kan det ses hvordan udpakningssubroutinen UNPAC er blevet kaldt fra de tabellerings og beregningsprogrammer, som har bearbejdet udvandringsmaterialet på 7094 (7090).

7094 arbejdede som en såkaldt ordmaskine, hvor cellerne i kernelageret er organiseret med en fast ens bitlængde. Her med 36 bit i hvert ord.

Som det fremgår af ovenfor anførte bilag bruges følgende bit-numre som parametre til UNPAC: 1, 9, 10, 18, 19, 27, 28, 36. Dette tyder på at den programmør, som har lavet bilaget, har nummereret bittene i de enkelte ord fra 1 - 36, mens maskinen i virkeligheden betragede den første bit, som bit nummer 0 og den 36'de bit som bit nummer 35. Der er grund til at tro at pakningen af data er sket med tilsvarende parametre til pakke-subroutinen PAC, som er anvendt ved pakningen af variable. Det er således åbenbart at den fjerde variabel i hvert pakket ord er bearbejdet med ikke definerede parametre til PAC og UNPAC. De værdier, der er fremkommet gennem denne bearbejdning af data, har tilsyneladende ikke været så absurde, at der er opstået mistanke om at nogle data blev invalide på grund af programfejl.

En mulighed for at disse absurde kald af PAC og UNPAC ikke har medvirket til fejl i de tabeller og beregninger, som er lavet på 7094, kan skyldes at disse subroutiner har benyttet en skifteoperation, der er en bevægelse af samtlige bit i et ord mod den ene ende af ordet, til at placere bitstrengene, som repræsenterer værdien for en pakket variabel (eller noget andet), i det pakkede ord.

Ved en sådan skifteoperation bevæges bittene fx. et antal positioner mod "højre", hvor højre repræsenterer den mindst betydende bit.

De bit som på denne måde skubbes "ud til højre" forsvinder ikke ud i den blå luft, men placeres i de bitpositioner, som bliver ledige til "venstre".

Følgende arbejdsgang kunne tænkes ved pakningen: PAC(A,28,36,B)

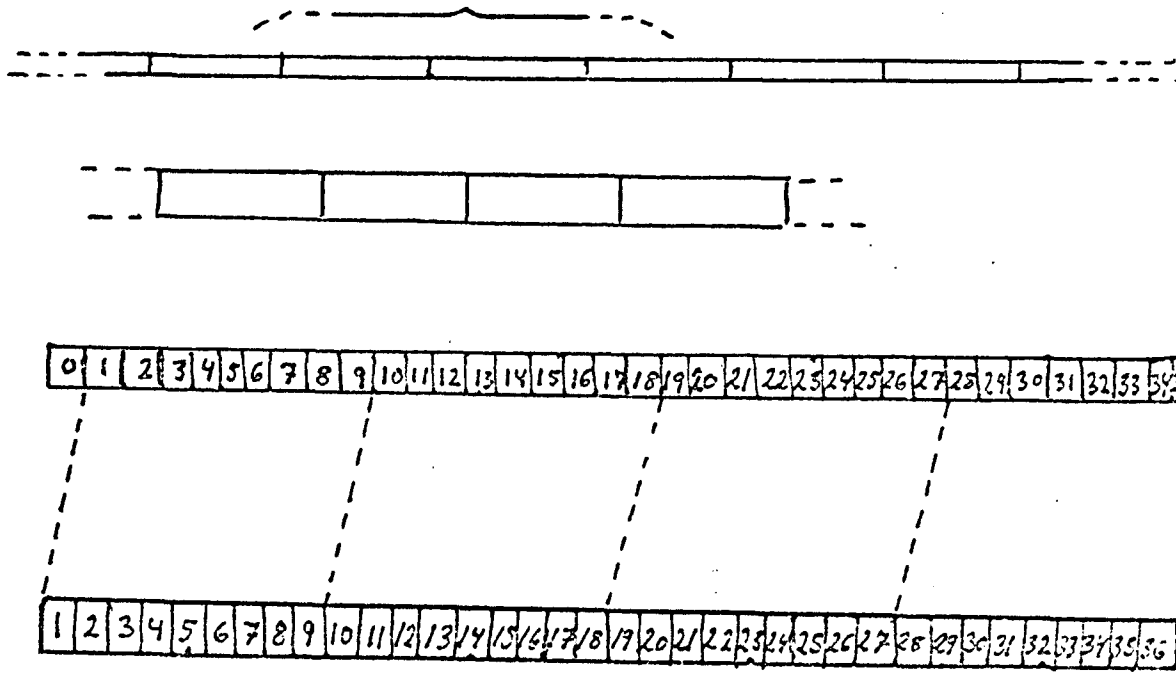
1) skift A 35-36 (flyt bittene én position mod venstre i A)
2) placer de 36-28+1 mest højrestillede bit fra B i de 36-28+1 mest højrestillede bit i A.
3) skift A -(35-36) (flyt bittene én position mod højre i A)
4) resultat: det mindst betydende bit fra B er placeret som bit nr. 0 i A

Hvis den tilsvarende arbejdsgang er anvendt ved udpakningen, vil man kunne genskabe de oprindelige data fra de pakkede ord ved en tilsvarende bearbejdning. Dette er muligt med FLD funktionen i FORTRAN V. Udpakning af en blok fra tapen blev derfor foretaget med det i bilag III)1.viste program. For dokumentation se program i bilag III)3.-4. og FLD funktionen i FORTRAN V manualen. Programmet pakker ud efter den hypotede, at ordene er pakket med en skifteoperation som den ovenfor skitserede. Det bør bemærkes at programmeringen er lettet af det forhold at FORTRAN V også bruges på en ordmaskine med 36 bit per ord.

Ved at betragte de udpakkede data fandt vi god overensstemmelse mellem kodeinstruks og de udpakkede poster, hvilket lod formode at hypotesen var en sandsynlig forklaring. Se bilag III) 2..

Der var dog stadigvæk den mulighed at den fjerde variabel, som var pakket og udpakket med ikke definerede parametre til PAC og UNPAC, ville vise bedre eller ligeså god overensstemmelse med kodeinstruksen ved at blive udpakket efter en anden sammenstilning af bitmønstrene. Dette vil blive undersøgt i udpakningens fase 2.

Organisering af variable og poster på det pakkede bånd.



Båndet er opbygget af logiske blokke med 900 ord i hver blok.
Der bruges 4 ord til hver post således at der ialt er 900/4 = 225 poster per blok. Der er 16 variable til hver post med 4 variable pakket i hvert ord. Hvert ord skulle være pakket som vist på den nederste tegning med 9 bit til hver variabel (nederste bitmønster). Det viste sig at være forkert. Vi kender nu kun pakkemønstret for de tre variable, som vist på det øverste bitmønster.

Udpakning anden fase.

Udover den i første fase skitserede udpakningshypotese opstillede vi en række hypoteser, som ud- og indpakning kunne beskrives med.
De øvrige hypoteser var:

1) den fjerde variabel bestod af de otte sidste bit fra det pakkede ord plus et bit, som var stillet på 1, så variablen altid var ulige.

2) den fjerde variabel bestod af de otte sidste bit fra det pakkede ord plus et bit, som var stillet på 0, så variablen altid var lige.

3) den fjerde variabel bestod kun af de otte sidste bit fra det pakkede ord, så variablen altid ville få en værdi, som var halvt så stor som under punkt to.

4) den fjerde variabel bestod af de sidste otte bit fra det pakkede ord plus en bit, som havde en tilfældig værdi. Vi anslog at sandsynligheden for en 1 bit var 0,5 og at der var den samme sandsynlighed for alle udpaknihgerne.

5) den fjerde variabel bestod af de sidste otte bit fra det pakkede ord plus en bit, som havde en værdi der svarede til værdien af den første bit i det næste ord i kernelageret.
(Dette næste ord vil bortset fra det sidste ord i blok være det næste element i talsættet blok. Ved udpakningen af det sidste element fra blok valgte vi arbitrært at bruge værdien af første bit i første element i blok.)

For at sammenligne resultatet ved udpakning på de ovenfor skitserede seks måder måtte vi lave et program, som kunne udpakke posterne på forskellig måde og udskrive de fordelinger, som fremkom.
Af de 16 variable hver post består af skal nogle variable ses i sammenhæng med andre. Således er oplysningen om en persons nummer i en udvandret gruppe kædet sammen med oplysning om gruppens størrelse. Ligeledes er stedbestemmelserne fordelt på to variable, som skal ses i sammenhæng. Dette betød at vi ikke kunne nøjes med at se på de simple marginalfordelinger, men måtte krydstabellere de variable, som skulle ses i sammenhæng.

Konklusion.

Formålet med de producerede tabeller har været at skabe et grundlag for en bedømmelse af de forskellige udpakningshypoteser.

I denne rapport vil der kun blive udført en nødtørftig analyse af tabellerne, hvor der ikke vil blive taget stilling til datas overensstemmelse eller mangel på samme med tidligere produktioner af tabeller fra den foreliggende tape. Kun en sammenligning mellem tabeller og kodeinstruks og en eventuel bedømmelse af datas konsistens vil danne grundlag for konklusionerne.

Tabellerne vil blive kommenteret i den orden, som de er udskrevet. Se tabeller i bilag III) 5.

Den første tabel viser god overensstemmelse med kodeinstruksen. Der er to forhold at bemærke. Det første er den manglende forekomst af individer i årene 00-14, men disse er på et tidligere tidspunkt slettet fra registret. Det andet er en manglende konsistens mellem antallet af udvandrere i 1968 i første tabel og antallet i tabel nummer to under form 3, som skulle være ens. Forskellen må skyldes en manglende rensning for invalide koder af den foreliggende tape.

Den anden tabel har udover den ovenfor nævnte afvigelse en del poster, som er kodet med 6 og 7, men ingen som er kodet med 5. Dette stemmer ikke med kodeinstruksens angivelse af koder.

I den tredie tabel kan man se at mange poster er kodet med et bogstav, som ikke er defineret i kodeinstruksen(8931 poster opført under gruppe 27).

Tabel nummer fire strækker sig over seks sider, da den tabulerede variabel er udpakket efter seks hypoteser. Ved at betragte denne tabel kan hypoteserne to, tre og fem med det samme udelukkes, mens de øvrige hypoteser giver en fordeling, som ikke afviger overvældende fra hinanden.

Den femte tabel giver en fordeling, som stemmer godt med de oplysninger, som gives i kodeinstruksen vedrørende nummerering af personer på sider

uden sidenummer, hvor det indirekte angives, at der normalt er 16 personer på en side. Dette stemmer med at koderne 1 - 16 rummer langt de fleste personer jævnt fordelt. Kode 00 er ikke nærmere defineret i kodeinstruksen, så det er ikke klart om de 62 personer i tabellens gruppe 100 skal tolkes som personer med nummer 100 eller som invalide koder.

Den sjette tabel viser at 264 personer er udvandret i en måned, som ikke er defineret, hvilket må skyldes den manglende rensning af registret for invalide koder.

Den syvende tabel viser at 4 personer har en kode for køn, som ikke er defineret.

I den ottende tabel er variablen udpakket med seks hypoteser. Da vi allerede har afskrevet hypotese nummer to, tre og fem, er det kun den fordeling, som de øvrige hypoteser giver, der har interesse. Her er det kun udpakningen efter hypotese nummer et, som giver en tilfredsstillende fordeling, da de andre har for mange poster i gruppe 10, der står for kode 0, der står enten for grupper over 9 personer eller uoplyst.

Tabel nummer ni viser at de seks hypoteser giver den samme fordeling, hvad de også skal. Tabellen viser en fordeling på nummer i gruppen, som stemmer godt overens med den fordeling over størrelsen af grupperne, som den foregående tabel giver med hypotese 1.

Den tiende tabel er en krydstabellering af antal og nummer i gruppen. Tabellen viser at en del personer er kodet med et nummer, som er større end antallet i gruppen. (Når udpakningshypotese nummer 1 antages at være korrekt.

Tabel elleve og tolv viser en ventet fordeling.

Tabel tretten viser at 336 personer er kodet med en ikke defineret kode for amt, når udpakningen foretages efter hypotese nummer ét. For hypoteserne 3 - 6 er antallet i denne kategori 10 000 - 35 000, hvilket viser at disse hypoteser må forkastes.

Tabel fjorten viser fordelingen på distrikter. Det bør bemærkes at tabellen viser at ingen personer har en invalid kode for distrikt (her gruppe 7).

Tabel femten viser fordelingen på distrikter i hvert amt, men tabellen er for omfattende til en kort kommentar.

Tabel seksten viser fordelingen på kommuner. Den viser at 156252 er kodet med 00 (her gruppe 100), hvilket stemmer overens med oplysningen om at kun to amter er kodet med opdeling på sognekommuner.

Tabel sytten viser fordelingen på landområder, hvortil udvandrerne rejste. Den viser at 8 personer er kodet med en invalid kode (her gruppe 58).

Tabel atten viser opdelingen på lokaliteter. Det ses at 1 person er kodet med en invalid kode (her gruppe 7).

I tabel nitten er hvert område fordelt på lokaliteter, men tabellen er for omfattende til at kommenteres kortfattet.

Den sidste tabel viser at opdateringerne i de forskellige matricer har været korrekt, da summen for hver variabel giver antallet af poster i registret.