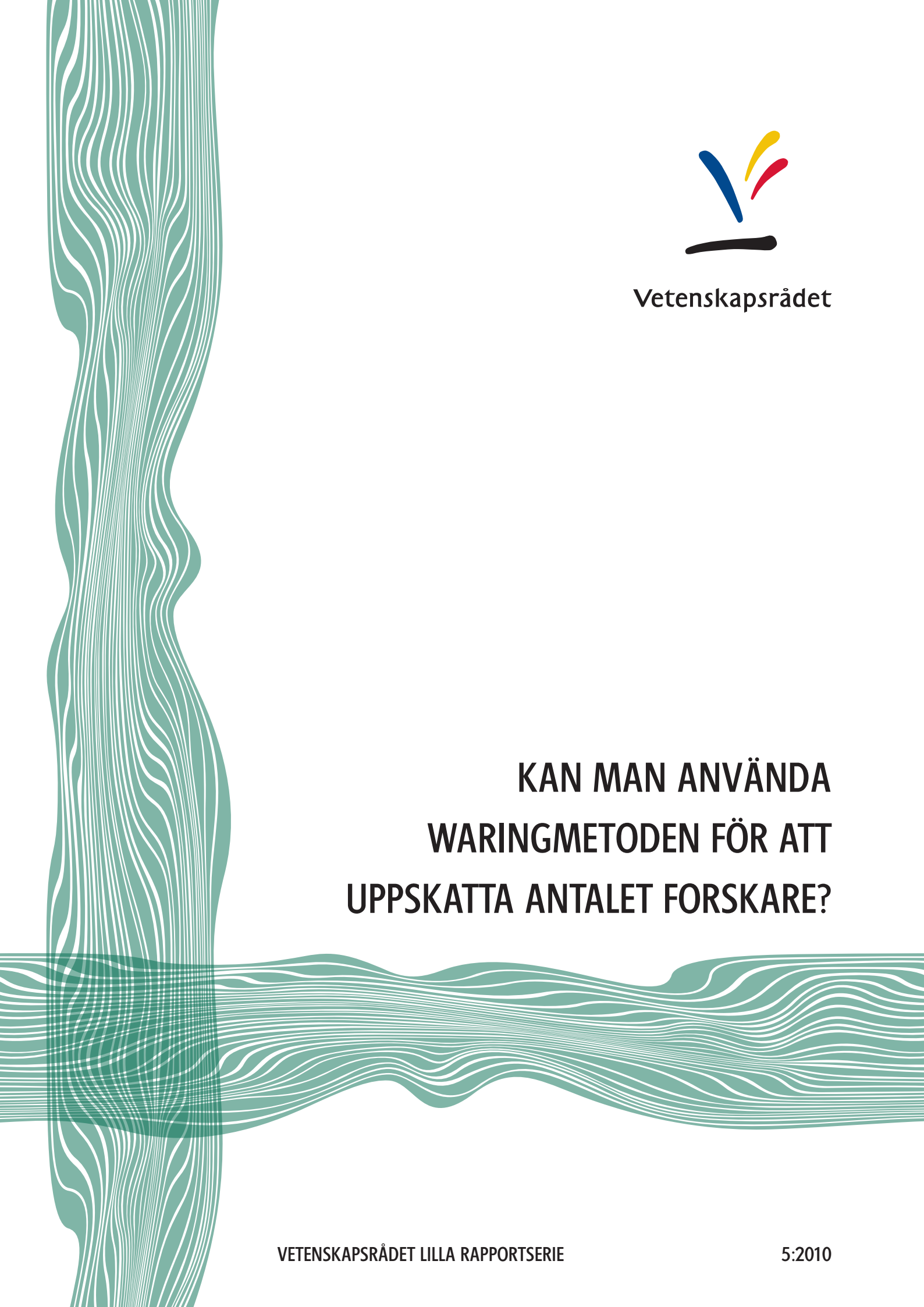


Vetenskapsrådet

# KAN MAN ANVÄNDA WARINGMETODEN FÖR ATT UPPSKATTA ANTALET FORSKARE?

VETENSKAPSRÅDET LILLA RAPPORTSERIE 5:2010

# KAN MAN ANVÄNDA WARINGMETODEN FÖR ATT UPPSKATTA ANTALET FORSKARE?

5 mars 2010

Johan Fröberg, Magnus Gunnarsson, Adam Jonsson och Staffan Karlsson

Avdelningen för forskningspolitisk analys

KAN MAN ANVÄNDA WARINGMETODEN FÖR ATT UPPSKATTA ANTALET FORSKARE?

Johan Fröberg, Magnus Gunnarsson, Adam Jonsson och Staffan Karlsson
Avdelningen för forskningspolitisk analys, Vetenskapsrådet

5 mars 2010

VETENSKAPSRÅDET
Box 1035
101 38 Stockholm


ISBN 978-91-7307-172-7

# SUMMARY

The system for funding allocation to public research institutions presented by the Swedish government in October 2008 included a bibliometric component. An important part of that component is a statistical estimation of how many active researchers there are in the Nordic countries, an estimation made with what has come to be known as the Waring method. In this report the Waring method is described and scrutinised in detail, and the conclusion is that it suffers from serious flaws:

- When looking into the published scientific literature where the Waring method has been used to estimate populations of researchers, we find conflicting results that only sometimes correspond with official statistics.
- The fine grained scheme of subject classification that is used in the Government's bibliometric component results in so small groups of publications that the Waring calculations become very unstable and in some cases completely collapse. For example, according to the Waring method the number of potential authors in the subject field of Clinical Medicine increases from 48,400 persons during the period 2000-2003 to 84,900 persons during the period 2005-2008. It seems unlikely that the actual researcher population should have changed so dramatically in such short time.
- The choice of subject classification scheme affects the end result substantially and we have found no reasons to why the classification used in the Government's bibliometric component should be fit for this particular task.
- The two reports that underlie the Government model describe a harmonisation of author names. This harmonisation requires a considerable amount of work that contains a substantial amount of subjectivity, but it does not in any critical way decrease the statistical confidence interval nor the variation between two subsequent periods.
- The average productivity is estimated by extrapolating a regression line, and from this follows that the estimated average productivity decreases if the number of highly productive authors increases. If we imagine that two or three researchers in a field step up from an average productivity to a high productivity, then this will result in a *decrease* in the average productivity of the field, contrary to what one might expect.

# INNEHÅLL

SAMMANFATTNING ..... 6
INLEDNING ..... 7
WARINGMODELLEN ..... 8
Metodens bakgrund och tidigare användning ..... 8
Waringfördelningens grund ..... 8
Beskrivning av metoden ..... 10
KRITISK DISKUSSION AV DELAR ..... 13
Ämnesklassificering ..... 13
Namnrättning ..... 13
Regressionsanalys – Extrapoleringen av regressionen ..... 14
Regressionsanalys – Extremvärden ..... 17
Regressionsanalys – Val av antal (trunkerings)punkter att ha med i analysen ..... 18
SUMMERING ..... 19
BIBLIOGRAFI ..... 20
APPENDIX A – RESULTAT EFTER OLIKA FORMER AV NAMNJUSTERING ..... 21
APPENDIX B – POTENTIELLA FÖRFATTARE ENLIGT WARINGMETODEN ..... 23
APPENDIX C – JÄMFÖRELSE MED OECD-STATISTIK ..... 24
APPENDIX D – OM VIKTERNA SOM ANVÄNDS VID BERÄKNING AV WARINGREFERENSVÄRDEN ..... 25

# SAMMANFATTNING

I det system för medelsfördelning till landets lärosäten för 2009 som regeringen presenterade i forskningspropositionen hösten 2008 ingick en bibliometrisk komponent. En viktig del av den komponenten utgörs av en statistisk uppskattning av hur många aktiva forskare som finns i Norden, en uppskattning som görs med den så kallade waringmetoden. I denna rapport beskrivs och granskas waringmetoden i detalj, och slutsatsen är att den lider av allvarliga brister:

- Den publicerade forskningslitteraturen där waringmetoden använts för att uppskatta forskarpopulationen ger motsägande resultat som bara ibland överensstämmer med officiell statistik.
- De finkorniga ämnesklassificeringar som används i regeringens bibliometriska komponent leder till så små publikationsgrupper att waringberäkningarna blir mycket instabila och i vissa fall helt kollapsar. Enligt waringmetoden ökar exempelvis antalet potentiella författare inom ämnesområdet klinisk medicin från 48 400 personer under perioden 2000-2003 till 84 900 personer under perioden 2005-2008. Det förefaller mindre sannolikt att den faktiska forskarpopulationen skulle har ändrats så dramatiskt på så kort tid.
- Valet av ämnesklassificering får betydelse för slutresultatet, och vi har inte hittat några skäl till varför just den ämnesklassificering som används i regeringens bibliometriska komponent skulle vara lämplig för uppgiften.
- Den rättning av författarnamn som gjorts i RUT2 och i HSV-rapporten är ett omfattande arbete som innehåller ett betydande mått av subjektivitet utan att på något avgörande sätt minska vare sig det statistiska konfidensintervallet eller mellanårsvariationen för beräkningarna.
- Tekniken att extrapolera fram medelproduktiviteten utifrån en regressionslinje leder till att den uppskattade medelproduktiviteten inom ett ämnesområde minskar om antalet högproduktiva forskare ökar. Om vi tänker oss att två-tre forskare inom ett ämne tar klivet upp från en genomsnittlig produktivitet till en hög produktivitet, så leder det till att den uppskattade medelproduktiviteten i ämnesområdet minskar, tvärtemot vad man skulle kunna förvänta sig.

# INLEDNING

I det system för medelsfördelning till landets lärosäten för 2009 som regeringen presenterade i forskningspropositionen hösten 2008 ingick en bibliometrisk komponent. Den baserades på ett underlag framtaget med den bibliometriska fördelningsmodell som beskrivs i rapporten *Resurser för citeringar* från Högskoleverket (här kallad HSV-rapporten). Denna modell är en modifierad version av den modell som presenterades i SOU 2007:81 – *Resurser för kvalitet*, här kallad RUT2.

En viktig del av fördelningsmodellen utgörs av en uppskattning av hur många aktiva forskare som finns i Norden. Denna uppskattning används i modellen för att kunna hantera det faktum att olika ämnen har olika traditioner för hur man publicerar sig i allmänhet och hur man publicerar sig i ISI-tidskrifter i synnerhet [1]. I det ena ämnet publicerar forskarna i genomsnitt fem artiklar per år, medan forskarna i det andra ämnet publicerar i genomsnitt en artikel vartannat år. I modellen antas att den arbetsinsats som ligger bakom en artikel i det andra ämnet motsvarar samma arbetsinsats som ligger bakom tio artiklar i det första ämnet.

För att beräkna det genomsnittliga antalet artiklar per nordisk forskare och år för olika ämnen krävs uppgifter dels om antalet publicerade artiklar och dels om antalet aktiva forskare. Den första uppgiften beräknas enkelt ur ISI-databasen. Den andra uppgiften är svårare att få fram, eftersom långtifrån alla aktiva forskare publicerar sig i ISI-tidskrifter under ett givet år, eller ens under en längre period. Det behövs alltså något sätt att uppskatta antalet aktiva forskare för olika ämnen under ett givet år, oavsett om forskarna har publicerat sig i ISI-tidskrifter eller inte. I den fördelningsmodell som diskuteras här har *waringmetoden* använts för denna uppskattning.

Vetenskapsrådet fick 2009-01-29 i uppdrag av regeringen att ta fram nödvändigt underlag för att tillämpa fördelningsmodellen samt att utveckla och redovisa den. Under arbetet med detta uppdrag framkom att waringmetoden har allvarliga brister, vilka redovisas i denna skrift.

[1] ISI-tidskrifter är sådana tidskrifter som Thomson Reuters inkluderar i ISI-databasen, alltså den databas som används som datakälla för HSV-rapporten och RUT2.

# WARINGMODELLEN

## Metodens bakgrund och tidigare användning

Utgångspunkten för att använda waringmetoden för att uppskatta antalet aktiva forskare är observationer som går tillbaka till 1920-talet gällande antalet författare som har ett visst antal publikationer (Lotka, 1926). Det visar sig att publikationerna är mycket ojämnt fördelade på författarna: ett fåtal författare har väldigt många publikationer och det stora flertalet författare har mycket få publikationer, om ens några. Den modell vi analyserar här använder en matematisk fördelning benämnd Waringfördelningen för att beskriva publikationsaktiviteten. Skeva fördelningar skiljer sig matematiskt i flera avseenden från mer symmetriska fördelningar. Så är till exempel olika mått som ofta används för att beskriva symmetriska fördelningar, till exempel medelvärde, median och varians, mindre användbara för att beskriva skeva fördelningar.

Herbert A. Simon (1955) använde en statistisk modell för skeva fördelningar för att beskriva ordfrekvenser, städers befolkningsstorlekar och forskares publikationsfrekvens. Det är denna modell som senare blev känd under namnet waringfördelningen. Joseph Oscar Irwin verkar inte ha känt till Simons arbete när han ett knappt decennium senare använde en tidigare känd formel för att beskriva samma typ av skeva fördelningar. Formeln hade utvecklats av Edward Waring på 1700-talet, och Irwin kallade därför fördelningstypen waringfördelning (Irwin 1963; Diodato 1994:163). Irwin använde waringfördelningen för att hantera fenomen inom biologi, citeringar till vetenskapliga verk, och i (Irwin 1968) även olycksrisk. I (Irwin 1975a; 1975b; 1975c) utvecklade och beskrev han fördelningen teoretiskt. Waringfördelningens historia beskrivs utförligt i (Xekalaki 1981).

I en serie arbeten under 1980-talet (Schubert & Glänzel 1984; Telcs, Glänzel & Schubert 1985; Schubert & Telcs 1986; Schubert & Telcs 1989; Braun, Glänzel & Schubert 1990) föreslog Wolfgang Glänzel och hans medarbetare på forskningsenheten vid biblioteket vid den ungerska vetenskapsakademin att waringfördelningen skulle kunna användas för att beräkna antalet potentiella vetenskapliga författare.

## Waringfördelningens grund

För att förstå varför publikationsfrekvens för forskare skulle kunna beskrivas med waringfördelningen behöver man förstå vilka principer som ligger till grund för waringfördelningen. Vi ska här först gå igenom dessa principer generellt, för att sedan applicera dem specifikt på forskares publikationsfrekvens.

Vi kan i Schubert och Glänzels efterföljd [2] utgå från att betrakta ett system med en oändlig följd av celler, där innehållet i cellerna kan förändras på tre sätt [3]:

A1. Nytt innehåll kan endast tillföras systemet genom den första, nollte, cellen.
A2. Innehåll kan endast flyttas i en riktning, från en cell $i$ till nästa i ordningen $i+1$ vilket sker med hastigheten $f_i$.
A3. Innehåll kan försvinna från systemet till omgivningen ur varje cell $i$ med en hastighet $g_i$.

Under vissa förutsättningar, i synnerhet att systemet betraktas efter så lång tid att ett tillstånd av konstant förändring uppnåtts, leder den matematiska behandlingen till en waringfördelning. Vid ett sådant tillstånd ökar (eller minskar) innehållet i varje cell exponentiellt med tiden.

[2] Den följande framställningen bygger på och ansluter nära till (Schubert & Glänzel 1984).

[3] Det är möjlig att det även finns andra system som kan beskrivas med waringfördelningen.

Ett specialfall av modellen är ett system vars element drabbas av en återkommande händelse. Elementen placeras i olika celler med utgångspunkt i hur många gånger denna händelse inträffat: de element för vilka ingen händelse har inträffat hamnar i den nollte cellen, de för vilka exakt 1 händelse inträffat placeras i den första cellen, osv. De ovan angivna reglerna kan då formuleras på följande sätt:

B1. Nya element tillförs systemet med en hastighet som är proportionell mot systemets hela omfattning, dvs. det totala antalet element i systemet.
B2. Intensiteten med vilken nya händelser inträffar för ett element ökar linjärt med antalet händelser som redan inträffat för det elementet.
B3. Element har samma möjlighet att lämna systemet oavsett tidigare antal händelser.

I ett system som följer dessa regler varierar det totala antalet element exponentiellt med tiden och fördelningen av element baserat på antalet inträffade händelser går mot en waringfördelning.

Är då vetenskaplig publicering ett system av denna typ? Vi betraktar först en grupp forskare, t.ex. alla forskare i Norden. Att dessa med viss regelbundenhet publicerar sina forskningsresultat är klart och det är således rimligt att se publicering som en återkommande händelse. Men följer systemet de spelregler vi just angivit?

C1. Den vetenskapliga litteraturen har visats växa exponentiellt (Price 1956). Det stämmer med att nya element tillförs systemet i en hastighet som är proportionell mot populationens storlek, dvs. med B1.
C2. Utgående från såväl Matteus-effekten (Merton 1973) som mer allmänna resonemang om att framgång föder framgång kan man hävda att ju mer en forskare publicerat, desto större är sannolikheten att han/hon publicerar igen. Det ger stöd för B2.
C3. Är det lika sannolikt att en forskare som publicerat 1 artikel helt slutar publicera sig som att en forskare som publicerat 30 artiklar gör det? Det förefaller kanske rimligt att anta att så är fallet. Schubert & Glänzel (1984:156) hävdar att även om sannolikheten för att sluta publicera sig minskar med ökande antal publikationer så är det i de flesta fallen rimligt att anta att forskare lämnar den publicerande verk-samheten i samma utsträckning oavsett hur många publikationer de redan har. Simon (1955) är av samma åsikt, om än i mindre utvecklad form. Därmed skulle även punkt B3 hålla.

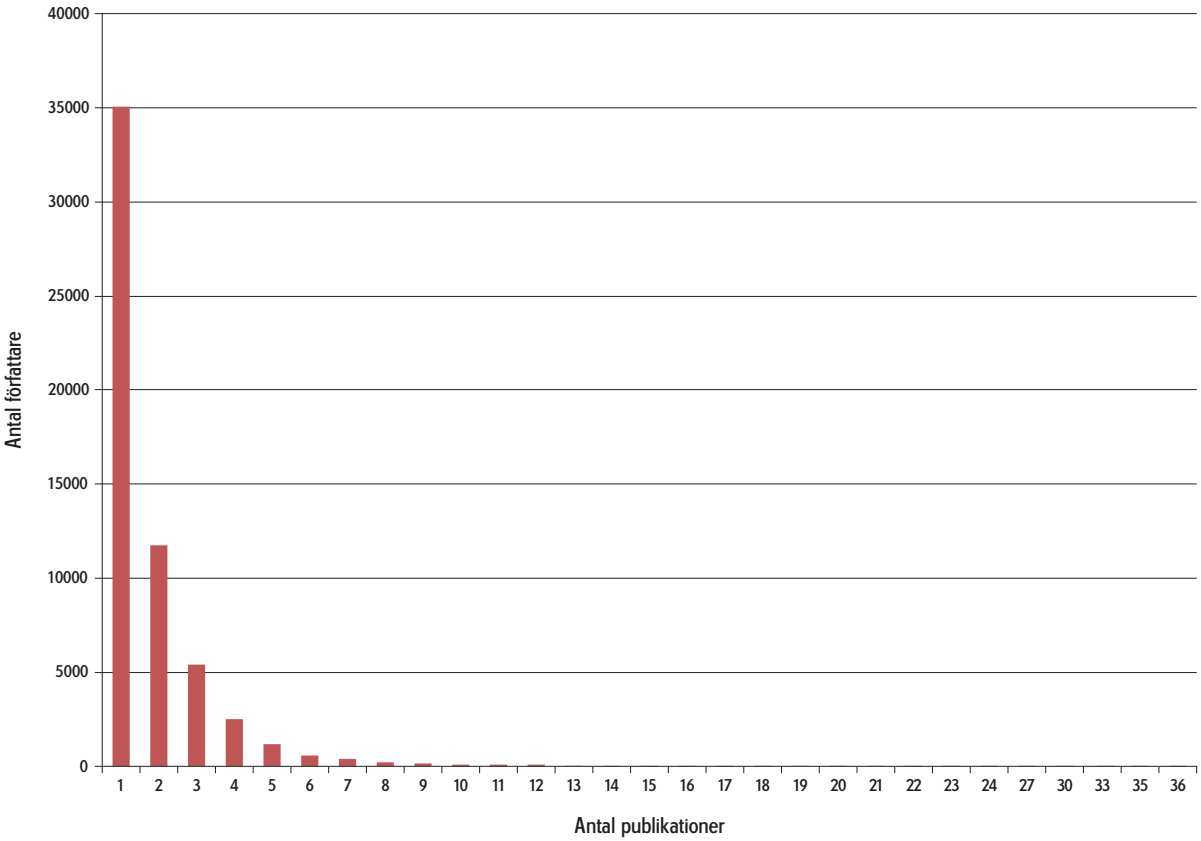
Punkterna C1-C3 kan givetvis kritiseras, vilket vi ska göra här för att visa på osäkerheten i resonemanget.

i Den vetenskapliga publiceringen i världen har uppvisat en exponentiell tillväxt under vissa studerade perioder, men det finns (naturligtvis) inga bevis för att den gör det under alla perioder och i alla ämnen eller geografiska områden.
ii Forskare som publicerat mycket kan ha en större tendens att publicera igen än forskare som publicerat lite, men det kan också finnas ett tak i den tendensen: när en forskare har publicerat ”tillräckligt” minskar publiceringstakten.
iii Vi vet egentligen inte särskilt mycket om hur sannolikheten för att en forskare helt slutar publicera är relaterad till forskarens publikationsvolym.
iv Även om C1-C3 skulle stämma vet vi inte hur stora forskar- och publikations-mängder som krävs för att den vetenskapliga produktionen ska närma sig waringfördelningen.

Punkterna C1-C3 och i-iv är naturligtvis ingen uttömmande eller slutgiltig diskussion om waringfördelningens tillämpbarhet på området vetenskaplig publikation, utan en beskrivning av problemställningen.

# Beskrivning av metoden

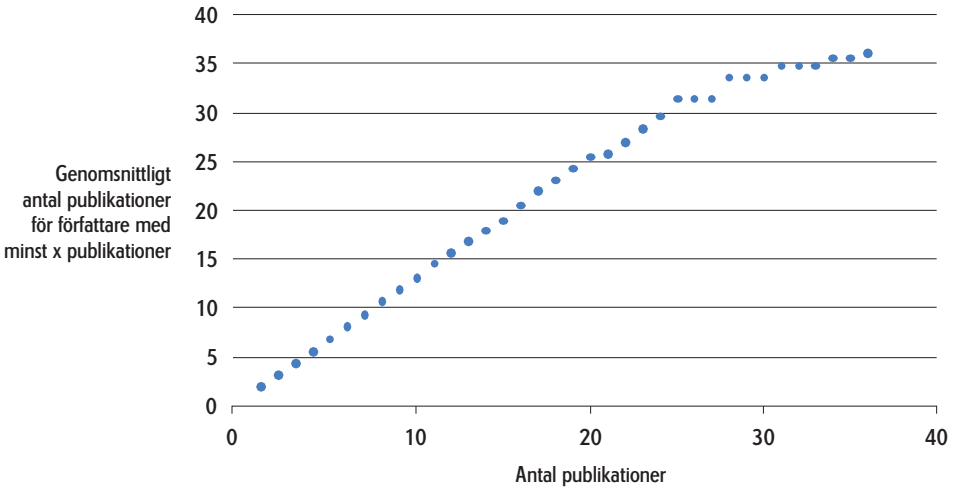
Grunddata för metoden är en publikationsdatabas av den typ som t.ex. ligger bakom Web of Science, dvs. ett någorlunda heltäckande register över de vetenskapliga publikationerna som forskare skrivit under en given tidsperiod och inom ett givet ämne. Antalet publikationer summeras för varje författare, och en frekvenstabell upprättas som visar antalet författare med 1 publikation, med 2 publikationer, med 3 publikationer, osv. Ett exempel visas i figur 1 nedan.




***Figur 1.*** *Frekvenstabell för publikationer per författare.*

Utifrån denna frekvenstabell skapas ett diagram som visar det genomsnittliga antalet publikationer för författare med en given minsta produktivitet, alltså för författare som producerat minst 1 publikation, minst 2 publikationer, minst 3 publikationer, osv. Ett exempel visas i figur 2 nedan.

Som synes i figur 2 ligger punkterna i diagrammet ovan på en någorlunda rät linje, och om man extrapolerar den linjen så att den skär y-axeln får man en uppskattning av det genomsnittliga antalet publikationer för samtliga forskare, $W_{ref}$, ett mått som alltså inkluderar även de forskare som inte publicerat något under perioden. Det totala antalet forskare beräknas sedan enkelt genom att bilda kvoten mellan det totala antalet publikationer och $W_{ref}$.




***Figur 2.*** *Den trunkerade medelvärdesserien för publikationer per författare. Data från figur 1.*

Det tillstöter här en begreppslig svårighet som gäller vad ”det totala antalet forskare” i föregående mening egentligen innebär. Vilka människor ”skulle kunna” skriva en vetenskaplig artikel? Schubert & Telcs (1989) kallar det ”number of potential authors”, medan Braun et al (1990) kallar det ”number of ’potentially productive’ scientists”; i båda skrifterna används även det mer abstrakta “(scientific) publication potential”. Här kallas det i fortsättningen ”antal potentiella författare”.

I (Schubert & Telcs 1989) används waringmetoden för att uppskatta antalet potentiella författare i 50 amerikanska delstater. Beräkningarna gjordes separat för varje delstat och resultaten summerades sedan till 647 000, vilket kan jämföras med den officiella amerikanska statistiken (Science Indicators 1980) som rapporterade 645 000 ”scientists and engineers”. I (Braun et al 1990) används metoden för att uppskatta antalet potentiella författare i tio OECD-länder, och resultatet redovisas i tabellen nedan.

***Tabell 1:*** *Antal potentiella författare i några OECD-länder, uppskattat med waringmetoden. Från (Braun et al 1990).*

| | Uppskattat antal potentiella författare | Antal forskare enligt Unesco | Antal forskare enligt OECD |
|---|---|---|---|
| Australien | 17 716 | 22 500 | 23 300 |
| Kanada | 33 482 | 26 200 | 26 300 |
| Frankrike | 39 010 | 72 900 | 72 900 |
| Tyskland | 43 622 | 122 000 | 122 000 |
| Italien | 16 894 | 40 800 | 46 400 |
| Nederländerna | 13 655 | 26 100 | 18 300 |
| Sverige | 10 696 | 14 800 | 14 800 |
| Schweiz | 10 940 | 16 400 | 10 700 |
| Storbritannien | 67 716 | 122 000 | 104 000 |
| USA | 269 649 | 637 000 | 621 000 |

Tabell 1 visar att skillnaden mellan waringmetodens uppskattningar och OECD:s respektive Unesco:s statistik skiljer sig olika mycket för olika länder. För Sverige och Schweiz ligger siffrorna tämligen nära varandra, för USA, Tyskland och Italien är OECD- och Unescostatistiken ca 2,5 ggr högre än waringuppskattningen, och för Kanada är OECD- och Unescostatistiken ca 20 % lägre än waringuppskattningen. Man kan också notera att det waringuppskattade antalet potentiella författare för USA är ca

40 % av det antal som uppskattades av Schubert & Telcs (1989), där delstaterna beräknades för sig och sedan summerades.

Det är inte självklart att waringmetoden borde komma fram till samma svar som den officiella statistiken. I den amerikanska statistiken rapporteras ”scientists and engineers”, medan waringmetoden, såsom den här tillämpats, uppskattar ”potentiella författare av publikationer i ISI-tidskrifter”. Den stora variationen i hur väl resultaten från waring-beräkningarna passar den officiella statistiken leder dock till den försiktiga slutsatsen att den officiella statistiken inte ger något stöd till waringmetodens användbarhet.

I RUT2, där författarna har använt waringmetoden för att uppskatta antalet potentiella författare vid Linköpings universitet och vid Kungliga tekniska högskolan, anges att ”metoden ger ganska precisa skattningar av personalomfattningen”. Inga närmre detaljer lämnas dock om detta.

# KRITISK DISKUSSION AV DELAR

I detta avsnitt detaljgranskas några problem med waringmetoden.

## Ämnesklassificering

De båda studier som tidigare använts för att jämföra waringmetoden med officiell statistik har sett till hela delstater/länder utan att göra någon ämnesuppdelning, vilket innebär många publikationer och stora forskargrupper. Resultaten från (Schubert & Telcs 1989:295) pekar på att när forskargrupperna blir mindre (de befolkningsmässigt små delstaterna Minnesota med 5,1 miljoner invånare; Wyoming med 0,5 miljoner invånare; och Idaho med 1,4 miljoner invånare) fungerar waringmetoden sämre. I RUT2 och i HSV-rapporten gjordes waringberäkningen på danska, finska, norska och svenska publikationer tillsammans, vilket ger ett befolkningsunderlag på ca 24,7 miljoner, men samtidigt användes i dessa båda rapporter ämnesindelningar med 23 respektive 34 klasser. Underlagen för waringberäkningarna blev då i snitt 4,3 % respektive 2,9 % av alla publikationer, men eftersom ämnesområdena inte var lika stora blev vissa områden så små som 0,87 % (RUT2) respektive 0,12 % (HSV-rapporten) av alla publikationer. Översatt till Schubert & Telcs studie skulle det motsvara en delstat med en befolkning på 220 000 respektive 63 000.

Problemet med små publikationsmänger som underlag till waringberäkningarna innebär att ämnesklassificeringen i sig kan få stor betydelse. Det finns modern forskning om nästan allt som människan känner till, vilket gör att det är nästan lika svårt att ämnesklassificera forskning som det är att ämnesklassificera all mänsklig kunskap [4]. Det finns sålunda många olika ämnesklassificeringar för vetenskapliga publikationer, men ingen större konsensus. Den ämnesklassificering som följer med ISI-databasen är tämligen spridd utan att för den skull nödvändigtvis vara omtyckt.

Den ämnesklassificering som användes för waringberäkningarna i RUT2 var en makroindelning av ISI-klasserna [5], medan den klassificering som användes i HSV-rapporten var helt ny. Den hade sitt ursprung i en klustring av tidskrifter i ISI-databasen baserad på citeringar mellan tidskrifterna; en metod som är känd men som innehåller stora variationsmöjligheter. Det finns ingen anledning att anta att klassificeringen i HSV-rapporten på något sätt skulle vara sämre än någon annan klassificering, men det *finns* anledning att anta att valet av ämnesklassificering påverkar waringberäkningarna. För så små publikationsmängder som det i sammanhanget rör sig om kan även små justeringar av ämnesgränser ge stora utslag i det beräknade waringvärdet $W_{ref}$. Valet av ämnesklassificering måste därför motiveras mycket väl.

## Namnrättning

I ISI-databasen finns information om publikationernas författare, men dessvärre enbart i form av efternamn följt av en eller flera förnamnsinitialer. Forskarna Lars Eriksson och Leif Eriksson listas alltså båda som *Eriksson, L* i ISI-databasen. Många forskare använder, väl medvetna om ISI-databasens begränsningar, flera förnamnsinitialer i sina publikationer, men tyvärr inte på ett helt konsekvent sätt. En forskare med det fullständiga namnet Per Erik Rune Lindberg kan då finnas i ISI-databasen som t.ex. *Lindberg, P; Lindberg, PE* och *Lindberg PER*.

I RUT2 och i HSV-rapporten gjordes en manuell justering av danska, finska, norska och svenska författarnamn med målet att föra samman olika namnformer för samma forskare till en enda form samt att dela upp namn som innehas av två eller flera forskare i flera former så att varje namnform används för endast en forskare. I rapporterna nämns ingenting om vilka principer som användes för att avgöra

[4] En längre diskussion om ämnesklassificeringar finns i (HEFCE 2009, p. 36 ff).

[5] Indelningen byggde på en klustring av ISI-klassernas utifrån citering mellan klassernas tidskrifter.

vilka namn som skulle delas upp och vilka som skulle slås ihop, men avsaknaden av dokumentation tyder på att det var en delvis subjektiv process. Det rör sig om ca 50 000 forskare.

Eftersom waringmetoden är tänkt att användas för medelsfördelning till landets lärosäten är det problematiskt att använda en manuell och delvis subjektiv metod för namnrättning, särskilt som denna måste göras om varje år. En manuell process försämrar metodens genomskinlighet och försvårar för lärosätena att upprepa metoden i kontrollerande och utbildande syfte. Detta kan i sin tur leda till att tilltron sjunker till systemet för medelsfördelning som helhet.

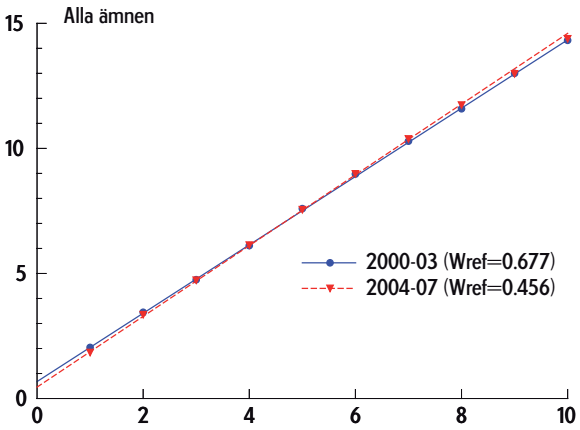
Under Vetenskapsrådets utredningsarbete användes en maskinell metod för namndisambiguering, dvs. att dela upp namn som innehas av två eller flera forskare. Den metoden innebar att lärosätesnamnet lades till författarnamnet, så att t.ex. *Eriksson, L* vid Uppsala universitet skiljdes från *Eriksson, L* vid Lunds universitet. För att undersöka effekten av namnrättning gjordes tre olika waringberäkningar med utgångspunkt i HSV-klassificeringen. I den första varianten användes den ovan nämnda maskinella metoden för namndisambiguering på omodifierade ISI-data. I den andra varianten användes manuellt harmonierade lärosätesnamn [6] i kombination med den maskinella metoden för namndisambiguering. I den tredje varianten gjordes en manuell genomgång av författarnamnen inom ett ämnesområde, Statistics. Undersökningen visade att skillnaden mellan den första och den andra varianten var betydanden, i det avseendet att felmarginalerna för Wref-värdena i de flesta fall minskade påtagligt när harmonierade lärosätesnamn användes. Den tredje varianten hade ungefär samma stabilitet och mellanårsvariation som den första varianten, men resultatet är något svårtolkat eftersom den andra varianten lyckdes betydligt sämre än den första varianten i det undersökta ämnet Statistics. Resultaten redovisas i större detalj i Appendix A.

I Vetenskapsrådets undersökningar har, om inget annat nämns, den andra varianten använts.

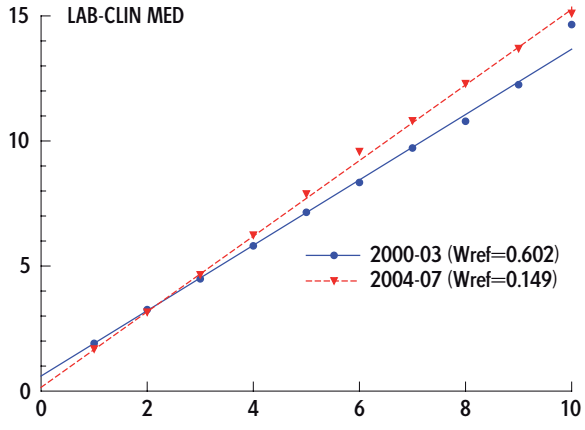
## Regressionsanalys – Extrapoleringen av regressionen

En grundläggande idé i waringmetoden är att punkterna i den trunkerade medelvärdes-serien (se figur 2) inte avviker särskilt mycket från en rät linje och att serien därför på ett trovärdigt sätt kan extrapoleras till x=0 (dvs. skärningspunkten med y-axeln). Figur 3 visar hur serierna ser ut för några olika ämnen. Den blå serien visar värden för perioden 2000-2003 och den röda serien visar värden för perioden 2004–2007.

Det ska påpekas att regressionen är viktad så att de första punkterna i serien, som har betydligt fler underliggande observationer än de sista punkterna, påverkar regressionslinjen mest. Viktningen beskrivs i appendix D.

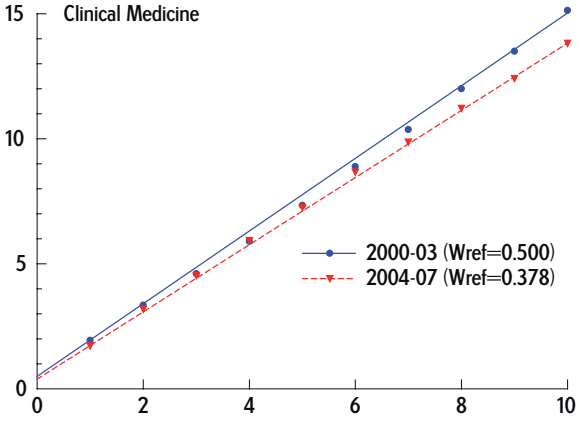



***Figur 3a.*** *Alla ämnen sammantaget*

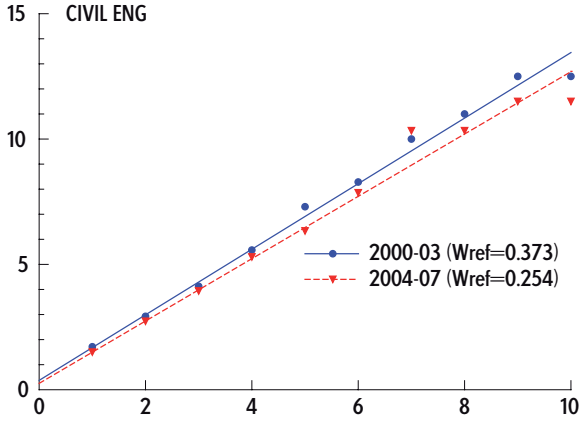



***Figur 3b.*** *Lab clin med*

[6] Denna harmoniering justerade t.ex. "University of Gothenburg" och "Univ Goteborg" till en och samma adress.




***Figur 3c.*** *Clinical medicine*




***Figur 3d.*** *Civil engineering*

Som figurerna 3a-3d visar finns det en variation mellan ämnen i hur väl det går att anpassa en regressionslinje till datapunkterna. För alla ämnen sammantaget ligger regressionslinjerna mycket nära punkterna, och regressionslinjerna från de två jämförda perioderna skiljer sig heller inte mycket åt. Även för *Clinical medicine* ligger punkterna nära regressionslinjen. För *Lab clin med* är punkterna något mer utspridda i relation till regressionslinjerna, och linjerna skiljer sig också ganska mycket åt mellan de två perioderna. För *Civil engineering* ligger regressionslinjen ännu längre från punkterna.

Ämnesklassningen som använts i figurerna ovan kallas SPRU19, en makroklassning av ISI-ämnena som är en lätt modifierad variant av den ämnesklassning som utvecklades av Katz & Hicks (1995). Den delar in de 255 ISI-klasserna i 19 större grupper, och har använts vid Vetenskapsrådet i många tidigare analyser. Under utredningsarbetet testades ett flertal olika ämnesklassificeringar [7], och mönstret är detsamma oavsett klassificering: för vissa områden passar regressionslinjen sina punkter väl, för andra områden gör den det inte. Den generella principen är att ju fler klasser ämnesindelningen har (dvs. ju färre författare ämnesklasserna innehåller), desto sämre passar regressionslinjen sina punkter.

Det är värt att notera att trots att en mänsklig betraktare kan tycka att regressionslinjen passar punkterna tämligen väl i alla diagrammen i figur3 så är det regressionslinjens skärning med y-axeln som är det viktiga, och även små skillnader i linjens anpassning kan ge stora effekter på denna skärning. För att få en bättre förståelse för hur stor variationen är kan den översättas till antal uppskattade potentiella författare – $W_{ref}$ är ett mått på antalet publikationer per person, eftersom vi vet antalet publikationer kan antalet personer lätt beräknas. Tabell 2 visar waringberäkningens resultat i termer av antalet uppskattade potentiella författare.

[7] Förutom de två klassificeringar som används i RUT2 (23 klasser) respektive HSV-rapporten (34 klasser) undersöktes även SPRU14, SPRU19, VETOMR, ESI och ESIPRIM. SPRU14 och SPRU19 är två grupperingar av ISI-klasserna som tagits fram vid University of Sussex, med 14 respektive 19 ämnen. VETOMR är en gruppering av ISI-klasserna i fyra stora områden motsvarande vetenskapsområden, framtagen vid Vetenskapsrådet. ESI är en klassificering från ISI med 23 ämnesgrupper, och ESIPRIM är en modifiering av ESI där den lilla klassen Space Science har slagits samman med Physics och där Humaniora har lagts till.

***Tabell 2:*** *Resultat av waringberäkningar för två perioder, översatt till antal potentiella författare. Ämnesklassning SPRU19.*

| Ämne | 2000–2003 | | | 2005–2008 | | |
|---|---|---|---|---|---|---|
| | Uppskattat antal potentiella författare | 95% konfidensintervall | | Uppskattat antal potentiella författare | 95% konfidens-intervall | |
| | | Min | Max | | Min | Max |
| Agriculture | 12 000 | 9 000 | 17 900 | 19 300 | 17 200 | 22 000 |
| Art | 6 000 | 2 500 | ∞ [1] | 7 000 | 3 700 | 58 900 |
| Biology | 5 000 | 4 700 | 5 500 | 12 700 | 11 200 | 14 700 |
| Biomedicine | 26 100 | 23 000 | 30 200 | 42 100 | 36 300 | 50 100 |
| Chemistry | 7 100 | 6 500 | 7 800 | 11 400 | 10 600 | 12 300 |
| Civil Eng | 1 400 | 1 000 | 2 100 | 3 600 | 2 200 | 8 700 |
| Clinical Medicine | 48 400 | 40 000 | 61 100 | 84 900 | 73 900 | 99 800 |
| Data | 6 100 | 4 000 | 12 600 | 9 400 | 7 900 | 11 600 |
| Eng Other | 600 | 500 | 1 000 | 1 000 | 800 | 1 500 |
| Eng Sci | 2 100 | 1 700 | 2 800 | 4 400 | 3 200 | 6 700 |
| Eng Tech | 1 600 | 1 300 | 2 000 | 2 400 | 1 800 | 3 700 |
| Geosciences | 2 700 | 2 500 | 2 900 | 8 300 | 6 400 | 11 900 |
| Lab-Clin Med | 10 800 | 9 400 | 12 600 | 57 400 | 41 600 | 92 600 |
| Materials Science | 1 600 | 1 400 | 1 700 | 3 700 | 3 100 | 4 500 |
| Mathematics | 2 000 | 1 700 | 2 500 | 3 400 | 2 700 | 4 300 |
| Physics | 100 | 0 | 100 | 100 | 0 | 400 |
| Soc | 7 600 | 7 200 | 7 900 | 11 900 | 11 000 | 12 900 |
| Systems | 9 900 | 2 800 | ∞ [1] | 34 200 | 26 700 | 47 500 |

[1] *Konfidensintervallets lägsta värde för* $W_{ref}$ *var mindre än eller lika med noll, vilket gör att antalet potentiella författare egentligen inte kan beräknas. Vi har tolkat det som att modellen uppskattar antalet potentiella författare till oändligt många.*

Tabellen ovan visar att de uppskattade antalen potentiella författare för vissa områden varierar kraftigt mellan de två perioderna. Det gäller framför allt de små områdena (t.ex. *Geosciences*, *Civil Enginerering* och *Eng Other*) men även de stora områdena *Lab-Clin Med* och *Systems*. Eftersom det verkar orimligt att det faktiska antalet forskare inom dessa discipliner skulle variera på det sättet måste slutsatsen vara att variationen beror på ett metodologiskt problem. Det stora konfidensintervallet för $W_{ref}$, som i tabell 2 kommer till uttryck i Min- och Max-kolumnerna, visar också på metodens osäkerhet. En jämförelse med OECD:s statistik finns i Appendix C.

Appendix B innehåller motsvarande data för den ämnesklassning som användes i HSV-rapporten.

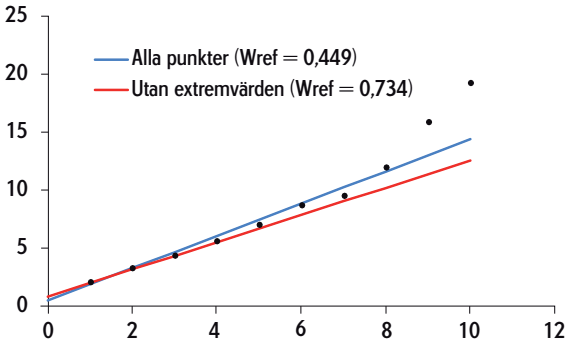
Som nämnts ovan bygger Waringmetoden på att man extrapolerar fram skärningspunkten mellan regressionslinjen och y-axeln, dvs. interceptet, för ett material där det minsta observerade x-värdet är 1. Som diskuterats ovan ökar variabiliteten kring linjen generellt med ökande värden på x. Eftersom regressionslinjen balanserar kring sin tyngdpunkt (oftast i närheten av x=2) orsakar en förändring i y-värdena för höga x-värden en motsatt förändring i interceptet när detta beräknas med minsta-kvadratmetoden. Regressionslinjen fungerar därför som en gungbräda med sin axel kring x=2, och om antalet högproducerande författare ökar så sjunker linjens skärning med y-axeln, och vice versa. Det betyder att om antalet högt produktiva forskare ökar från en period till en annan, minskar modellens predicerade medelproduktivitet. Detta mönster framgår av flera av figurerna i denna rapport – ett tydligt exempel är figur 3b. En eventuell eliminering av extremvärden (se nästa avsnitt) påverkar interceptet, dvs. $W_{ref}$ värdet, på samma sätt.

# Regressionsanalys – Extremvärden

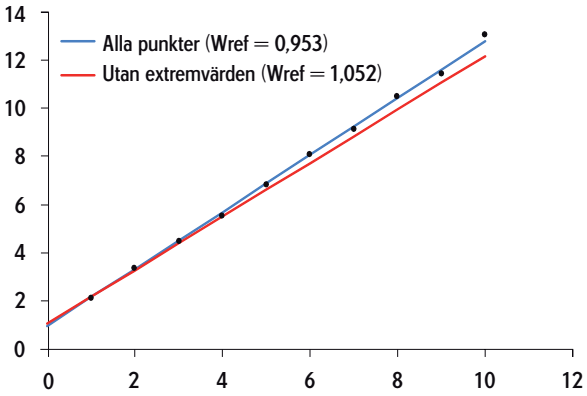
I RUT2 görs en manuell justering av den trunkerade medelvärdesserien:

> Extremvärden som tydligt inte är representativa för klassen tas bort (t.ex. om en författare publicerat dubbelt så många artiklar som den som publicerat näst flest).
>
> *RUT2, sid. 438.*

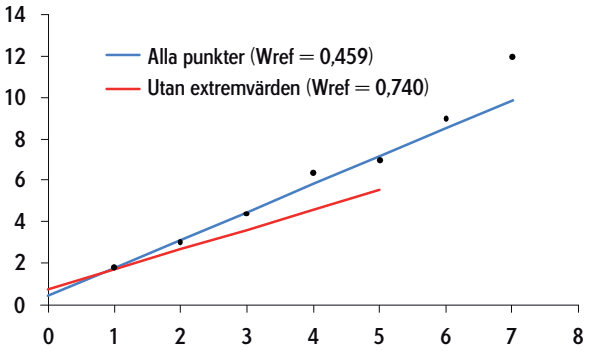
I HSV-rapporten nämns inget om detta. Under Vetenskapsrådets utredningsarbete har ett filter applicerats vid beräkningarna så att den mest produktive författaren tas bort om han/hon producerat mer än 75 % fler publikationer än den näst mest produktive författaren under perioden. Detta påverkar två ämnesklasser när HSV-klassificeringen används och 1-2 klasser om Sprul9-klassificeringen används (beroende på tidsperiod).



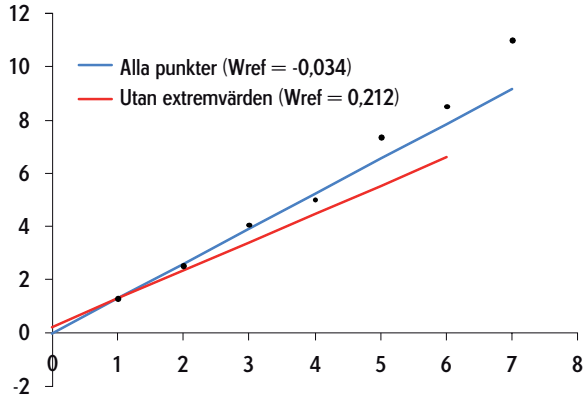

***Figur 4a.*** *HSV, Agriculture, 2000–2003*




***Figur 4b.*** *Sprul9, Biology, 2000–2003*




***Figur 4c.*** *HSV, Information Science, 2000–2003*




***Figur 4d.*** *HSV, Environmental studies, 2004–2007*

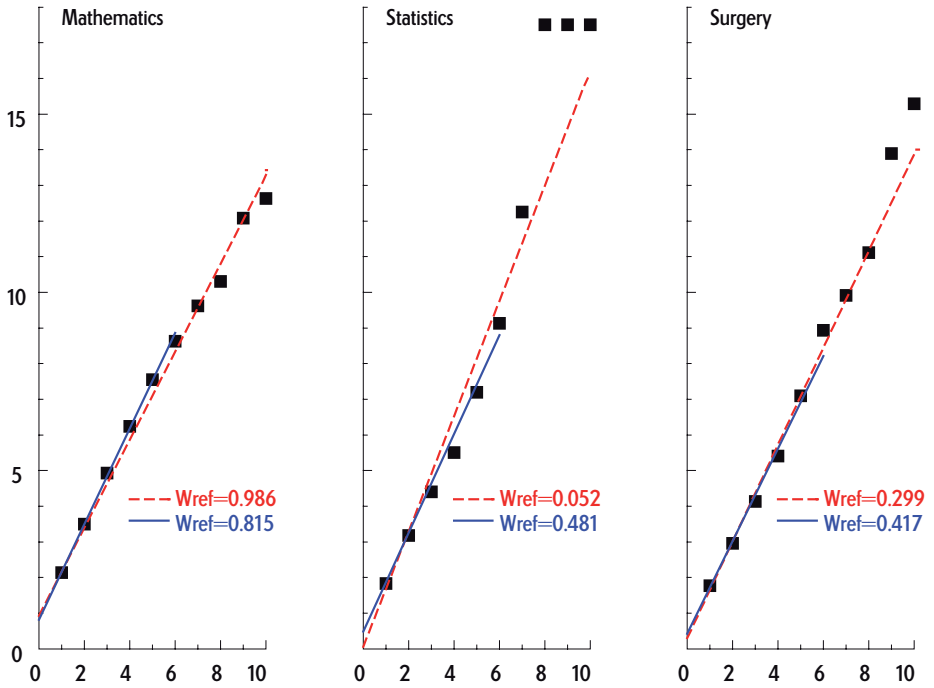
Observera att i figur 4d skär regressionslinjen y-axeln under nollpunkten när alla punkter tas med.

Det finns inte någon motivering i RUT2 eller HSV-rapporten till varför de mest produktiva författarna ska ignoreras i beräkningen, annat än att de utgör extremvärden i regressionsanalysen.

## Regressionsanalys – Val av antal (trunkerings)punkter att ha med i analysen

Ju längre till höger man kommer i den trunkerade medelvärdesserien desto mindre stabila är värdena, eftersom färre observationer ligger bakom dessa. Som nämnts ovan viktas regressionen så att de första punkterna påverkar linjen mer än de sista, vilket leder till en stabilare regression. Ett sätt att förstärka detta (eller ett sätt att justera viktningen, om man så vill) är att bara inkludera de första punkterna i serien vid regressionen. I figur 2 ovan skulle man t.ex. kunna nöja sig med de 20 första punkterna, som alla ligger mycket nära en tänkt rät linje. Schubert & Telcs (1989) använde alla punkter i sin analys, medan Braun et al (1990), RUT2 och HSV-rapporten verkar ha begränsat sig till de 10 första punkterna. [8] I de beräkningar som Vetenskapsrådet gjort i samband med denna rapport har bara de 10 första punkterna använts, om inget annat anges.

Valet av antal punkter har viss betydelse. Figurerna nedan visar regressionslinjen för tre områden i HSV-klassificeringen med 6 respektive 10 punkter.




***Figur 5.*** *Regression med olika antal punkter i den trunkerade medelvärdesserien för tre områden i HSV-klassificeringen. Perioden 2004-2007. Röd linje visar regression med 10 punkter och blå linje visar regression med 6 punkter.*

Att begränsa antalet punkter i regressionen innebär att man bortser från de mest produktiva forskarna, och vi har inte hittat några andra skäl i litteraturen för detta än att det beräknade $W_{ref}$-värdet då blir mer stabilt.

[8] Ingenting nämns om detta i texterna, men alla i diagram visas endast de 10 första punkterna.

# SUMMERING

Att använda waringmetoden för att uppskatta antal forskare på det sätt som föreslagits i RUT2 och HSV-rapporten är en mycket osäker metod med stora felkällor:

- Den publicerade forskningslitteraturen där waringmetoden använts för att upp-skatta forskarpopulationen ger motsägande resultat som bara ibland överensstämmer med officiell statistik.
- De finkorniga ämnesklassificeringar som används i RUT2 och HSV-rapporten leder till så små forskargrupper att beräkningarna blir mycket instabila och i vissa fall helt kollapsar [9]. Enligt waringmetoden ökar exempelvis antalet potentiella författare inom klinisk medicin från 48 400 under perioden 2000-2003 till 84 900 under perioden 2005-2008. Det förefaller mindre sannolikt att den faktiska forskarpopulationen skulle ha ändrats så dramatiskt på så kort tid.
- Valet av ämnesklassificering får betydelse för slutresultatet, och det finns inga ämnesindelningar som i sammanhanget kan sägas vara ”opartiska” eller ”korrekta”. Vi hittar inga skäl till varför just de ämnesklassificeringar som används i RUT2 och HSV-rapporten skulle vara lämpliga.
- Den namnrättning som gjorts i RUT2 och i HSV-rapporten är ett omfattande arbete som innehåller ett betydande mått av subjektivitet utan att på något avgörande sätt minska vare sig det statistiska konfidensintervallet eller mellanårsvariationen för waringvärdet ($W_{ref}$).
- Den rensning av extremvärden i den trunkerade medelvärdesserien som i RUT2 och HSV-rapporten görs inför waring-regressionen innebär att de mest produktiva forskarna ignoreras utan att det finns något stöd för detta i den bakomliggande teorin.
- Den begränsning som görs av antalet datapunkter i regressionen innebär, på samma sätt som vid bortrensning av extrempunkter, att de mest produktiva forskarna ignoreras utan att det finns något stöd för detta i den bakomliggande teorin
- Tekniken att extrapolera fram medelproduktiviteten utifrån en regressionslinje leder till att den uppskattade medelproduktiviteten inom ett ämnesområde minskar om antalet högproduktiva forskare ökar. Om vi tänker oss att två-tre forskare inom ett ämne tar klivet upp från en genomsnittlig produktivitet till en hög produktivitet, så leder det till att den uppskattade medelproduktiviteten i fältet *minskar*, tvärtemot vad man skulle kunna förvänta sig.

[19] När Wref är mindre än eller lika med noll.

# BIBLIOGRAFI


Braun, Tibor; Glänzel, Wolfgang & Schubert, András. (1990) Publication productivity: from frequency distriutions to scientometric indicators. I *Journal of Information Science*, 16, sid. 37-44.

Diodato, Virgil Pasquale (1994). *Dictionary of Bibliometrics*. Haworth Press, New Tork, USA.

HEFCE (2009). *Report on the pilot exercise to develop bibliometric indicators for the Research Excellence Framework.* Higher Education Funding Council for England, Storbritannien. Elektronisk resurs: http://www.hefce.ac.uk/pubs/hefce/2009/09_39/

Irwin, Joseph Oscar (1963). The Place of Mathematics in Medical and Biological Statistics. I *Journal of the Royal Statistical Society*. Series A (General), vol. 126, no. 1, sid. 1-45.

Irwin, Joseph Oscar (1968). The Generalized Waring Distribution Applied to Accident Theory. I *Journal of the Royal Statistical Society.* Series A (General), vol. 131, no. 2, sid. 205-225.

Irwin, Joseph Oscar (1975a). The Generalized Waring Distribution. Part I. I *Journal of the Royal Statistical Society*. Series A (General), vol. 138, no. 1, sid. 18-31.

Irwin, Joseph Oscar (1975b). The Generalized Waring Distribution. Part II. I *Journal of the Royal Statistical Society.* Series A (General), vol. 138, no. 2, sid. 204-227.

Irwin, Joseph Oscar (1975c). The Generalized Waring Distribution. Part III. I *Journal of the Royal Statistical Society*. Series A (General), vol. 138, no. 3, sid. 374-384.

Katz, J. Sylvan & Hicks, Diana (1995). *The Classification of Interdisciplinary Journals: A New Approach*. The Fifth Biennial Conference of The International Society for Scientometrics and Informatics, Rosary College, River Forest, Il., USA.

Lotka, Alfred James (1926) The frequency distribution of scientific productivity. I *Journal of the Washington Academy of Sciences*, 16, sid. 317-323.

Merton, R. K. (1973). The Matthew Effect in Science. I Storer, N.W. (ed) *The Sociology of Science*, sid. 439-459. University of Chicago Press, Chigago, USA.

Price, Derek J. (1956). The Exponential Curve of Science. I *Operational Research Quarterly*, vol. 10, no. 3 (Sep), sid. 179.

Schubert, András. & Telcs, András. (1986) Publication potential – an indicator of scientific strength for cross-national comparisions. I *Scientometrics*, vol. 9, nos 5-6, sid. 231-238.

Schubert, András & Telcs, András (1989). Estimation of the Publication Potential in 50 U.S. States and in the District of Columbia Based on the Frequency Distribution of Scientific Productivity. I *Journal of the American Society for Information Science*, 40(4), sid. 291-297.

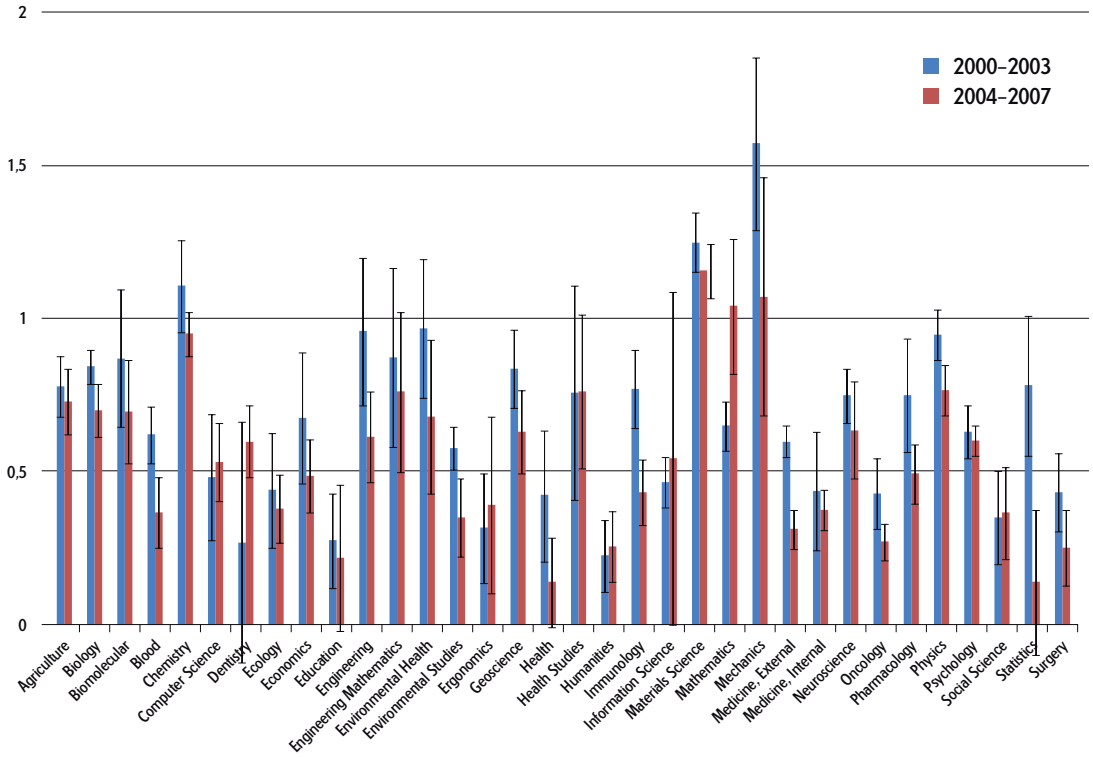
Schubert, András & Glänzel, Wolfgang. (1984) A dynamic look at a class of skew distributions. A model with scientometric applications. I *Scientometrics*, vol. 6, no. 3, sid. 149-167.

Simon, Herbert A. (1955) On a Class of Skew Distribution Functions. I *Biometrika*, vol 42, no. 3/4 (Dec), sid. 425-440.

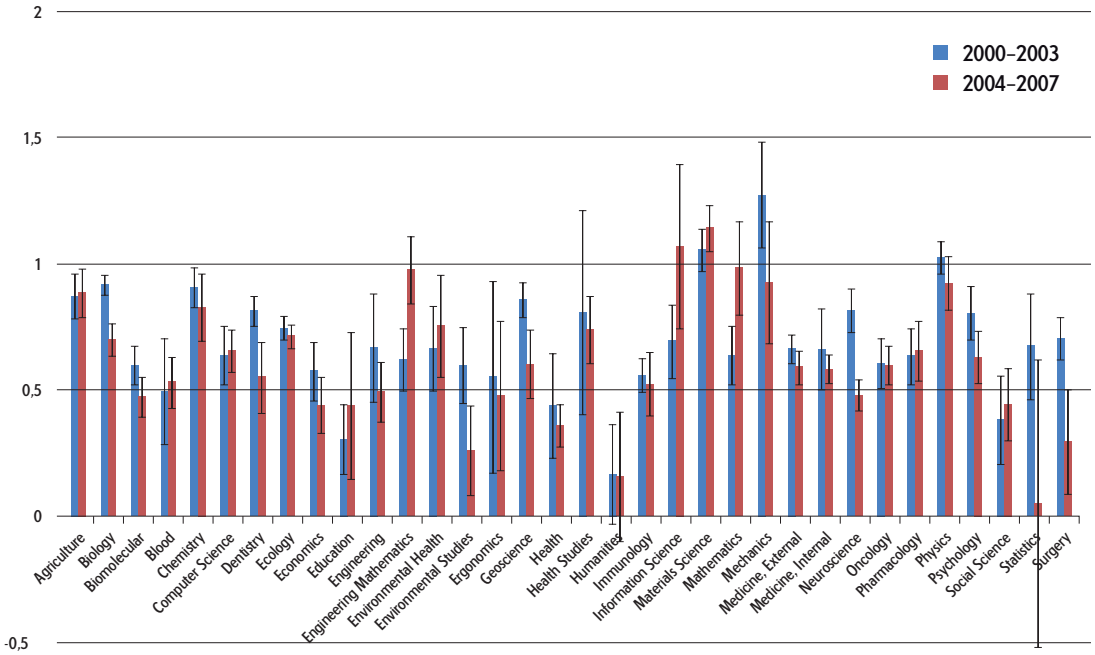
Telcs, András; Glänzel, Wolfgang & Schubert, András. (1985) Characterization and statistical test using truncated expectations for a class of skew distributions. I *Mathematical Social Sciences* 10, sid. 169-178.

Xekalaki, E. (1981). Chance Mechanisms for the Univariate Generalized Waring Distribution and Related Characterizations. I Taillie, C & Patil, G. P. (eds) *Statistical Distributions in Scientific Work*, Vol. 4 - Models, Structures, and Characterizations, sid. 157-171. D. Reidel Publ. Co., Dordrecht, Tyskland.

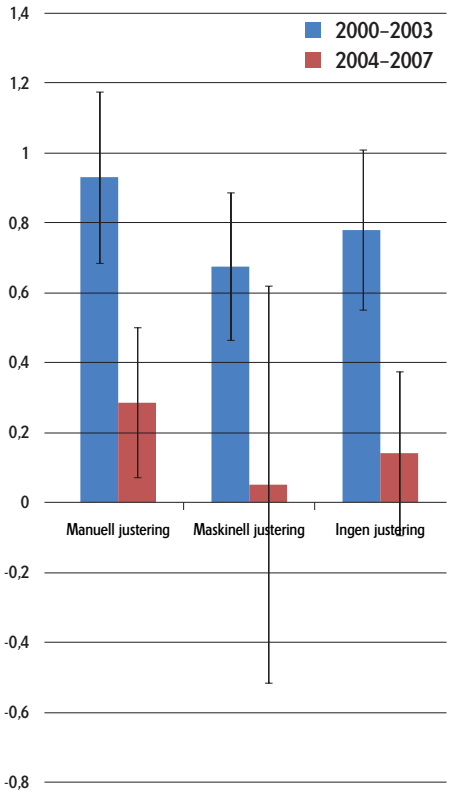
# APPENDIX A
# RESULTAT EFTER OLIKA FORMER AV NAMNJUSTERING




***Figur A1:*** *Resultat av waringberäkningar helt utan manuell namnrättning; maskinell namndisambiguering har dock gjors genom tillägg av organisationsnamnet. De blå (vänstra) staplarna visar $W_{ref}$-värden beräknade för perioden 2000-2003, och de röda (högra) staplarna visar $W_{ref}$-värden beräknade för perioden 2004-2007. De svarta linjerna visar de 95-procentiga konfidensintervallen för $W_{ref}$.*




***Figur A2:*** *Resultat av waringberäkningar när namndisambiguering har skett genom att manuellt standardiserade lärosätesadresser har lagts till författarnamnet. De blå (vänstra) staplarna visar $W_{ref}$-värden beräknade för perioden 2000-2003, och de röda (högra) staplarna visar $W_{ref}$-värden beräknade för perioden 2004-2007. De svarta linjerna visar de 95-procentiga konfidensintervallen för $W_{ref}$.*




***Figur A3:*** *Resultat av waringberäkning för området* ***Statistics*** *efter manuell adresstandardisering och manuell justering av författarnamn, tillsammans med samma område med maskinell justering respektive helt ojusterad. Den blå (vänstra) stapeln visar $W_{ref}$-värdet beräknat för perioden 2000-2003, och den röda (högra) stapeln visar $W_{ref}$-värdet beräknat för perioden 2004-2007. De svarta linjerna visar det 95-procentiga konfidensintervallet för $W_{ref}$.*

# APPENDIX B
# POTENTIELLA FÖRFATTARE ENLIGT WARINGMETODEN

***Tabell B1:*** *Resultat av waringberäkningar för två perioder, uttryckt i antal potentiella författare. HSV-klassificering.*

| | 2001–2004 | | | 2005–2008 | | |
|---|---|---|---|---|---|---|
| Ämne | Uppskattat antal potentiella författare | 95% konfidensintervall | | Uppskattat antal potentiella författare | 95% konfidens-intervall | |
| | | Min | Max | | Min | Max |
| Agriculture | 2 100 | 2 000 | 2 300 | 2 900 | 2 400 | 3 600 |
| Biology | 9 900 | 9 200 | 10 800 | 11 800 | 10 100 | 14 200 |
| Biomolecular | 10 500 | 6 600 | 25 100 | 7 700 | 6 800 | 9 000 |
| Blood | 4 800 | 3 500 | 7 600 | 4 900 | 4 200 | 5 800 |
| Chemistry | 7 600 | 7 000 | 8 400 | 10 000 | 7 800 | 13 800 |
| Computer Science | 8 900 | 6 900 | 12 800 | 9 800 | 8 000 | 12 800 |
| Dentistry | 1 500 | 1 400 | 1 700 | 1 600 | 1 200 | 2 500 |
| Ecology | 7 900 | 7 100 | 8 800 | 9 000 | 8 000 | 10 300 |
| Economics | 3 800 | 2 800 | 5 800 | 7 700 | 5 800 | 11 400 |
| Education | 1 700 | 700 | – [1] | 3 000 | 800 | – [1] |
| Engineering | 1 800 | 1 400 | 2 500 | 1 800 | 1 300 | 2 900 |
| Engineering Mathematics | 600 | 400 | 1 100 | 600 | 500 | 900 |
| Environmental Health | 1 000 | 700 | 1 400 | 1 200 | 1 000 | 1 800 |
| Environmental Studies | 900 | 600 | 1 800 | 1 600 | 900 | 9 700 |
| Ergonomics | 400 | 200 | 900 | 600 | 300 | – [1] |
| Geoscience | 2 200 | 1 600 | 3 400 | 2 400 | 2 000 | 2 900 |
| Health | 5 300 | 4 300 | 7 100 | 6 200 | 4 800 | 8 900 |
| Health Studies | 600 | 300 | 2 000 | 500 | 300 | 1 200 |
| Humanities | 11 700 | 1 900 | – [1] | 4 900 | 2 200 | – [1] |
| Immunology | 5 700 | 4 900 | 6 800 | 6 900 | 5 900 | 8 300 |
| Information Science | 200 | 100 | 700 | 300 | 200 | 600 |
| Materials Science | 5 000 | 4 300 | 5 800 | 4 900 | 4 500 | 5 400 |
| Mathematics | 1 700 | 1 200 | 2 800 | 1 200 | 900 | 1 700 |
| Mechanics | 900 | 100 | – [1] | 200 | 100 | 300 |
| Medicine, External | 7 500 | 6 700 | 8 700 | 11 300 | 8 000 | 19 200 |
| Medicine, Internal | 16 400 | 14 000 | 19 900 | 25 600 | 19 900 | 36 000 |
| Neuroscience | 6 200 | 5 300 | 7 600 | 9 500 | 8 500 | 10 800 |
| Oncology | 9 300 | 7 500 | 12 100 | 8 700 | 6 800 | 12 100 |
| Pharmacology | 1 700 | 1 200 | 2 800 | 1 900 | 1 500 | 2 600 |
| Physics | 5 600 | 4 900 | 6 400 | 7 500 | 5 600 | 11 200 |
| Psychology | 1 800 | 1 500 | 2 200 | 2 500 | 2 000 | 3 300 |
| Social Science | 2 000 | 1 200 | 5 000 | 2 600 | 1 800 | 4 800 |
| Statistics | 800 | 500 | 1 600 | 1 900 | 1 000 | 27 300 |
| Surgery | 4 300 | 3 400 | 6 000 | 9 000 | 5 200 | 35 400 |

[1] *Konfidensintervallets lägsta värde för* $W_{ref}$ *var mindre än 0.*

# APPENDIX C
# JÄMFÖRELSE MED OECD-STATISTIK

| Uppskattat antal potentiella författare | 2001–2004 | 2005–2008 |
|---|---|---|
| Waring hsv-indelning | 152 000 | 182 300 |
| Waring spru19-indelning | 159 400 | 321 600 |
| Waring ej ämnesindelad | 146 900 | 293 900 |

| Uppskattat antal forskare | 2001–2004 | 2005–2007 |
|---|---|---|
| HE[1] researchers heads count | 83 000 | 89 800 |
| HE[1] researchers FTE | 42 600 | 44 100 |
| HE[1] tot R&D personnel FTE | 55 900 | 56 700 |
| Tot. researchers heads count | 193 100 | 213 700[2] |

[1] *HE = Higher Education*

[2] *2005 års värde.*

OECD-statistiken är uttagen från OECD.Stat 2008-12-17 för länderna Danmark, Finland, Norge och Sverige. För perioden 2001-2004 har medelvärdet för de fyra åren använts för OECD-uppgifterna. Eftersom det inte finns uppgifter för 2008 har medelvärdet för åren 2005-2007 använts för perioden 2005-2008.

I R&D personnel inkluderas även TA-personal dvs. alla personal som på ett eller annat sätt har med R&D verksamheten att göra.

# APPENDIX D
# OM VIKTERNA SOM ANVÄNDS VID BERÄKNING AV WARINGREFERENSVÄRDEN

Syftet med detta dokument är att tydliggöra hur de vikter som används vid beräkning av waringreferensvärden tagits fram.

Låt *N(k)* beteckna antalet författare med *k* publikationer, *k* = 0,1,2, . . . Den trunkerade medelvärdesserien *e(0)*, *e(1)*, . . . definieras

$$e(k) = \frac{\sum_{j=k}^{\infty} jN(j)}{\sum_{j=k}^{\infty} N(j)}, k \geq 0$$

Antag att N*(k)* genererats från ett antal observationer $X_1, X_2, \ldots, X_N$ på en waringfördelad stokastisk variabel *X* med fördelningsfunktion *F*, dvs. *N(k)* = # { *j* : *Xj* = *k*}, och där *N* är antalet potentiella författare. Standardavvikelsen av *e(k)*, säg $\sigma_k$, är ett mått på osäkerheten i skattningen *e(k)*. Om $\mu_k$ betecknar väntevärdet *E[e(k)]* så gäller ([1], s id. 173) att $\sigma_k$ = C · 1/w(k) där

$$w(k) = \sqrt{\frac{1 - F(k-1)}{\mu_k(\mu_k + 1)}}, k \geq 0$$

Proportionalitetskonstanten C beror av waringfördelningens parametrar. Vi har $\sigma_1 < \sigma_2 < \sigma_j < \ldots$ , dvs. osäkerheten i skattningen *e(k)* ökar då *k* ökar. (Det gäller även att $\sigma_k \to \infty$ då $k \to \infty$.) En naturlig skattning av $\sigma_k$ är *1/W(k)* där

$$W(k) = \sqrt{\frac{\sum_{j=k}^{\infty} N(j)}{e(k)(e(k) + 1)}}, k \geq 0$$

För att ta hänsyn till att (vårt mått på) osäkerheten *1/w(k)* i skattningen *e(k)* är växande i *k* så används *W(k)*, $k \geq l$ som vikter då vi anpassar en regressionslinje till *e(l)*, *e(2)*, . . . ([1], sid. 173). Vikterna *W(k)*, $k \geq 1$ beror inte av *N(0)* och kan därför beräknas utifrån *N(1)*, *N(2)*, . . . .

## Referenser

[1] Glänzel, W.; Telcs, A. & Schubert, A. (1985). Characterization and statistical test using truncated expectations for a class of skew distributions. In *Mathematical Social Sciences*, Vol. 10, sid. 169-178.