TRAFIK ANALYS

Röjandekontroll av regional statistik om järnvägstransporter

PM 2018:5

Röjandekontroll av regional statistik om järnvägstransporter

PM 2018:5

Trafikanalys
Adress: Torsgatan 30
113 21 Stockholm
Telefon: 010 414 42 00
Fax: 010 414 42 10
E-post: trafikanalys@trafa.se
Webbadress: www.trafa.se
Ansvarig utgivare: Brita Saxton
Publiceringsdatum: 2018-05-31

Förord

Trafikanalys har regeringens uppdrag att föreslå åtgärder för att förbättra kunskapsunderlaget om järnvägstransporter med inriktning på mer detaljerad information om geografi och varugrupper.

Möjligheten att publicera mer detaljerad statistik inom området kan dock påverkas av sekretessbestämmelser. Därför finns ett behov av att studera resultat av statistisk röjandekontroll av data om järnvägstransporter.

Tom Andersson är projektledare på Trafikanalys för regeringsuppdraget. I projektgruppen har Pia Bergdahl, Gelaye Holmér, Fredrik Lindberg, Eva Lindborg, Tom Petersen, Henrik Petterson och Mats Wiklund ingått. Mats Wiklund har ställt samman denna PM och Gelaye Holmér har gjort datorkörningar med programmet τ-ARGUS.

Östersund i maj 2018

Per-Åke Vikman

Avdelningschef, Trafikanalys

Innehåll

Sammanfattning

När uppgifter samlas in för att sammanställa statistik är god svarsfrekvens och god svarskvalitet viktigt. En förutsättning för det är att ge uppgiftslämnarna en grund för förtroende att uppgifterna skyddas. Sekretess och röjandekontroll är därför viktiga för statistikens kvalitet. Det är bl.a. därför det i offentlighets- och sekretesslagen finns en bestämmelse om sekretess inom statistikverksamhet. Den sekretessen, statistiksekretessen, avser uppgift om en enskilds personliga eller ekonomiska förhållanden och som kan hänföras till den enskilde.

Statistisk röjandekontroll ska säkra att statistiksekretessen efterlevs. Den är mest välutvecklad för statistik som består av tabeller, men metoderna kan också användas för att skydda mikrodata. Röjandekontrollen griper främst in i två skeden av statistiksammanställning. Det gäller dels när tabellplan för statistiken utformas och dels när tabeller sammanställs.

När tabellplanen utformas går det att från kunskap om målpopulationen och empiriska kunskaper, göra bedömning om det finns tabellceller där det troliga utfallet är att de kommer att utgöra riskceller, d.v.s. celler vars redovisning skulle medföra en inte betydelselös risk för röjande. Tabellplanen bör utformas så att sådana potentiella riskceller undviks. Det kan också innebära att man undviker att samla in onödigt detaljerade uppgifter, vilket bidrar till minskad uppgiftslämnarbörda.

När sedan tabeller utformats enligt plan görs en systematisk bedömning av röjande- och skaderisk, d.v.s. identifikation av riskceller. Kontrollrutinen gör skillnad på frekvens- och magnitudtabeller.

En tabell med riskceller behöver skyddas. Ett sätt att undertrycka riskceller, vilket innebär att istället för att publicera cellens värde anges det att värde saknas. Ett annat sätt att skydda riskceller är att störa eller ändra värdet före publicering, d.v.s. använda sig av perturbativa metoder.

Denna PM har sin utgångspunkt i de uppgifter om regionala järnvägstransporter som ska samlas in enligt EU-förordning. Olika metoder för röjandekontroll prövas för en mer detaljerad tabellutformning för dessa data. Trafikanalys har tidigare dragit slutsatsen att publicering av tabell med på- och avlastningsregioner baserade på NUTS 2-region utgör risk för röjande av enskild, och därmed inte publicerats. Den analys som gjorts här pekar mot att det i stort är en korrekt slutsats. Dock framgår att vissa uppgifter på mer detaljerad geografisk nivå kan publiceras utan röjanderisk, vilket man bör göra när uppgifterna samlats in och finns tillgängliga. Analysen visar på olika vägar att skydda sekretessen, och diskuterar möjligheten att andra metoder kan ge ytterligare möjligheter.

1 Inledning

När uppgifter samlas in för att sammanställa statistik är god svarsfrekvens och god svarskvalitet viktigt. En förutsättning för det är att ge uppgiftslämnarna en grund för förtroende att uppgifterna skyddas. Sekretess och röjandekontroll är därför viktiga för statistikens kvalitet. Det är bl.a. därför det i offentlighets- och sekretesslagen finns en bestämmelse om sekretess inom statistikverksamhet. Den sekretessen, statistiksekretessen, avser uppgift om en enskilds personliga eller ekonomiska förhållanden och som kan hänföras till den enskilde.

Trafikanalys samlar in underlag järnvägstransportstransportstatistik i enlighet med EU-parlamentets och rådet förordning (EU) nr 91/2003. Den föreskriver bl.a. att medlemsländerna vart femte år ska samla in uppgifter om nationella godstransporter efter NUTS2-region för pålastning och NUTS2-region för avlastning. Emellertid har Trafikanalys utgått från att publicering av tabell med på- och avlastningsregioner baserade på de underlagen utgör risk för röjande av enskild. Den slutsatsen är snarare baserad på en bedömning än på en ingående analys. Syftet med denna PM är att genomföra röjandekontroll av tabeller med regionala godstransporter på järnväg och visa tabellernas innehåll efter att de har skyddats.

2 Statistisk röjandekontroll

Statistisk röjandekontroll ska säkra att statistiksekretessen efterlevs. Den är mest välutvecklad för statistik som består av tabeller, men metoderna kan också användas för att skydda mikrodata. Röjandekontrollen griper främst in i två skeden av statistiksammanställning. Det gäller dels när tabellplan för statistiken utformas och dels när tabeller sammanställs.

När tabellplanen utformas går det från kunskap om målpopulationen och empiriska kunskaper göra bedömning om det finns tabellceller, där det troliga utfallet är att de kommer att utgöra riskceller, d.v.s. celler vars redovisning skulle medföra en inte betydelselös risk för röjande. Tabellplanen bör utformas så att sådana potentiella riskceller undviks. Det innebär också att man undviker att samla in onödigt detaljerade uppgifter, vilket kan bidra till minskad uppgiftslämnarbörda.

2.1 Bedömning av röjanderisk

När sedan tabeller utformats enligt plan görs bedömning av röjande- och skaderisk, d.v.s. identifikation av riskceller. Kontrollrutinen gör skillnad på frekvens- och magnitudtabeller. I frekvenstabeller redovisas antal individer eller objekt som har de egenskaper eller attribut som bestäms av cellens position i tabellen.

Frekvenstabeller

När cellens frekvens är låg, d.v.s. när ett litet antal individer eller objekt redovisas i cellen, ska cellen betraktas som en riskcell. Frekvensnivåer färre än tre brukar alltid betraktas som så pass låg frekvens att risk för röjande föreligger, men man kan också utgå från högre tröskelvärden.

Magnitudtabeller

I magnitudtabeller utgörs istället cellvärdena av någon form av aggregering variabelvärden för de individer eller objekt som tillhör de olika tabellcellerna. Där gäller att riskceller är sådana där ett litet antal individer eller objekt bidrar till en stor del av magnituden. Även här gäller att när ett betydande bidrag till magnituden kommer från få individer eller objekt, ska cellen betraktas som en riskcell.

Eftersom denna PM är inriktad mot att studera risk för röjande i några magnitudtabeller görs här en liten fördjupning i hur riskceller kan identifieras. Det finns två huvudsakliga metoder för detta. Den ena använder dominansregeln och den andra *p* %-regeln. Här redogörs för dessa i deras vanligaste parameteruppsättningar, men de kan varieras på olika sätt, se (ROS, 2015).

Dominansregeln anger att om det redovisade värdet i en tabellcell domineras av bidrag från två objekt (de två största) är det en riskcell. Här gäller att de två största dominerar om de bidrar med minst 90 procent av cellens redovisade värde.

p %-regeln brukar formuleras omvänt jämför med dominansregeln. *p* %-regeln utgår från hur stort bidraget från alla objekt utom de två som bidrar mest. Om det bidraget är litet för en

tabellcell gäller att den är en riskcell. Här gäller att det är en riskcell om bidraget från alla objekt utom de två största är mindre än 15 procent av bidraget från det objekt som bidrar med mest.

Dessa beslutsregler är ofta ekvivalenta, d.v.s. de kommer ofta till samma slutsats om huruvida en tabellcell är en riskcell eller inte.

2.2 Skydd av tabeller

En tabell med riskceller behöver skyddas. Det kan göras på olika sätt. Vid sidan om att aggregera tabellutformningen till färre redovisningsgrupper används två metoder för att skydda tabeller. Det är dels undertryckning och dels perturbation.

Undertryckning

Dels kan riskceller undertryckas, vilket innebär att istället för att publicera cellens värde anges det att värde saknas. Eftersom man vill redovisa rad- och kolumnsummor i tabeller blir det då oftast nödvändigt att undertrycka fler tabellceller än riskcellerna, s.k. sekundärundertryckning. Om inte det görs kan det finnas möjlighet att röja uppgiften i den undertryckta riskcellen genom att räkna baklänges m.h.a. rad- eller kolumnsumman.

Valet av tabellceller som ska sekundärundertryckas kan göras på många olika sätt. Varje alternativ ger upphov till informationsförluster, som kan mätas på olika sätt. När man har bestämt metod för att mäta informationsförlust kan med matematiska optimeringsmetoder välja ut de tabellceller som ska sekundärundertryckas genom att minimera informations-förlusten. Det finns ett stort utbud av datorprogramvaror för att lösa matematiska optimeringsproblem. Det finns de som är mer avancerade (och dyrare) som oftast kommer närmare en optimal lösning än de enklare varianterna.

Perturbation

Ett annat sätt att skydda riskceller är att störa eller ändra värdet före publicering, d.v.s. använda sig av perturbativa metoder. För frekvenstabeller innebär det helt enkelt att frekvensvärden avrundas uppåt eller nedåt.

Även för magnitudtabeller innebär perturbation att cellvärden ändras uppåt eller nedåt, men då så att det publicerade värdet inte längre domineras av ett litet antal individer eller objekt. Detta är dock mer komplicerat än avrundning i frekvenstabeller. För att ge en indikation på hur denna form av perturbation går till kan man utgå från en tabellcell där bidraget från de två största utgör mer än 90 procent av cellens värde. Den tabellcellen är då en riskcell. Säg att de två största bidrar med 97 procent av cellens värde, då går det att visa att genom att istället ange ett värde som är minst 8 procent större än det egentliga värdet blir risken för röjande liten. Det går också att bestämma ett värde som är lägre än tabellcellens egentliga värde som är tillåtet att ange.

Precis som vid undertryckning medför perturbation av riskcellers värden att det måste göras sekundär perturbation för att förhindra baklängesidentifikation. Det innebär, precis som vid undertryckning, att man måste finna bra lösningar på matematiska optimeringsproblem.

Perturbation för att skydda magnitudtabeller är en förhållandevis ny metod och utvecklingen av metoden startades bl.a. med (Dandekar, 2002).

3 Röjandekontroll för tabeller regionala järnvägstransporter

I detta kapitel görs röjandekontroll av tabeller över nationella godstransporter med järnväg efter region för pålastning och region för avlastning. Regionindelningen är dels NUTS 2 och dels NUTS 1[1]. De data som har använts avser inrikes godstransporter på regional nivå som Trafikanalys samlade in avseende år 2015 i enlighet med EU-parlamentets och rådet förordning (EU) nr 91/2003.

Syftet att beskriva om det överhuvudtaget går att publicera några uppgifter och samtidigt ge en fingervisning om hur stor informationsförlusten kan tänkas bli.

Röjandekontrollen har gjorts med stöd av datorprogramvaran τ-ARGUS. Den matematiska optimeringen för sekundärundertryckningen och för sekundär perturbation har gjorts med stöd av ett enklare datorprogram som är fritt tillgängligt. Det innebär att den informationsförlust som uppstår med denna röjandekontroll förmodligen kan reduceras med stöd av bättre programvara.

Vid publiceringen av den officiella statistiken om järnvägstransporter år 2015 redovisades mängden transporterat gods på järnväg utan annan geografisk upplösning än inrikes respektive utrikes transport (ref#). I Tabell 1 redovisas samma data uppdelad mellan NUTS 2-regioner[1], med undertryckning i röjandekontroll enligt τ-ARGUS. Hela 53 av tabellens 64 celler (förutom rad- och kolumnsummor) har undertryckts (primärt eller sekundärt). Det totala transporterade godset i de undertryckta tabellcellerna uppgick till 25 851 tusen ton, vilket är 73 procent av den totala mängden gods som redovisas i tabellen. Informationsförlusten är alltså ganska påtaglig, men å andra sidan finns ju en del information fortfarande tillgänglig.

Tabell 1. Antal tusen ton transporterat gods på järnväg i inrikes trafik år 2015 mellan NUTS 2-regioner[1], där tabellceller har undertryckts för att förhindra röjande.
Källa: Trafikanalys

	Till								
Från	*SE11*	*SE12*	*SE21*	*SE22*	*SE23*	*SE31*	*SE32*	*SE33*	*Totalt*
SE11	..	..	..	..	..	..	..	..	741
SE12	..	812	..	..	..	1 647	..	..	3 965
SE21	..	..	..	..	..	..	..	..	1 104
SE22	..	566	..	..	..	..	..	..	2 432
SE23	..	..	428	..	..	1 276	94	321	4 412
SE31	..	1 642	..	..	2 328	..	..	..	7 417
SE32	..	..	..	..	191	..	..	..	1 914
SE33	..	..	..	..	347	..	..	..	13 519
Totalt	1 867	3 677	904	2 769	5 232	7 167	1 350	12 538	35 504

[1] NUTS nivå 1-3 i Sverige (2018-05-09).

Det förtjänar att upprepas att informationsförlusten i Tabell 1 möjligen kan reduceras med stöd av bättre programvara för matematisk optimering.

Ett sätt att minska informationsförlusten vid röjandekontroll kan vara att minska informationsinnehållet från början i den tabell man avser att publicera. Det kan man göra genom att slå samman redovisningsgrupper. I Tabell 2 redovisas godstransporterna på NUTS 1-nivå, istället för på NUTS 2-nivå som i Tabell 1. Andelen undertryckta celler minskar då till fyra av nio. Det transporterade godset i de undertryckta cellerna uppgår till 11 924 tusen ton, vilket svarar mot 33 procent av den totala mängden transporterat gods. Notera att den totala mängden transporterat gods är något högre i Tabell 2 än i Tabell 1, vilket beror på att det i grunddata enbart finns uppgift om NUTS 1-region för på- och avlastning.

Tabell 2. Antal tusen ton transporterat gods på järnväg i inrikes trafik år 2015 mellan NUTS 1-regioner[1], där tabellceller har undertryckts för att förhindra röjande.
Källa: Trafikanalys

	Till			
Från	*SE1*	*SE2*	*SE3*	*Totalt*
SE1	1 005	2 135	1 881	5 022
SE2	..	..	2 467	7 989
SE3	..	..	16 869	23 270
Totalt	5 699	9 366	21 217	36 281

Som nämnts ovan är perturbation ett alternativ till undertryckning för att skydda tabeller. I Tabell 3 redovisas godstransporter med järnväg mellan NUTS 2-regioner, där tabellen skyddats genom perturbation. Hur informationsförlust ska mätas vid perturbation verkar fortfarande vara ett aktuellt forskningsområde, se (Hundepool, 2012). För att beräkna informationsförlusten här har kvadratsumman beräknats för skillnaderna mellan verkligt och redovisat värde för de 64 cellerna i Tabell 3. Roten ur den summan är 1 216 ton, vilket svarar mot 3 procent av tabellens totalsumma. Kvadratsumman kan normeras med antalet frihetsgrader i tabellen. Eftersom rad- och kolumnsummor är fixerade i tabellen är antalet frihetsgrader $(8 – 1) \times (8 – 1) = 49$. Det innebär att roten ur medelkvadratfelet är 174 ton, vilket kan tolkas som ett medelfel per tabellcell.

Tabell 3. Antal tusen ton transporterat gods på järnväg i inrikes trafik år 2015 mellan NUTS 2-regioner[1], tabellceller har ändrats (perturberats) för att förhindra röjande.
Källa: Trafikanalys

	Till								
Från	*SE11*	*SE12*	*SE21*	*SE22*	*SE23*	*SE31*	*SE32*	*SE33*	*Totalt*
SE11	0	14	0	410	106	0	0	211	741
SE12	0	812	0	1 084	230	1 647	0	192	3 965
SE21	0	135	78	93	249	230	0	318	1 104
SE22	725	566	92	157	442	192	35	224	2 432
SE23	236	353	428	458	1 339	1 276	0	321	4 412
SE31	534	1 680	147	170	2 328	1 923	82	552	7 417
SE32	0	0	0	78	191	288	920	437	1 914
SE33	372	117	158	318	347	1 611	313	10 282	13 519
Totalt	1 867	3 677	904	2 769	5 232	7 167	1 350	12 538	35 504

Den uppmärksamme läsaren kanske noterar att enligt Tabell 3 transporterades inget gods från region SE23 till SE32, medan Tabell 1 anger 94 ton för samma relation. Det behöver inte vara något konstigt med det, men det understryker ett behov av att se över om ett bättre optimeringsprogram behöver användas.

Tabell 4 redovisas godstransporter med järnväg mellan NUTS 1-regioner, där röjandeskyddet har gjorts med perturbation. Roten ur summan av kvadratavvikelser mellan redovisat och verkligt värde är 481 ton, vilket svarar mot 1 procent tabellens totalsumma. Medelkvadratfelet är 241 ton.

Tabell 4. Antal tusen ton transporterat gods på järnväg i inrikes trafik år 2015 mellan NUTS 1-regioner[1], där tabellceller har ändrats (perturberats) för att förhindra röjande.
Källa: Trafikanalys

	Till			
Från	*SE1*	*SE2*	*SE3*	*Totalt*
SE1	1 170	1 971	1 881	5 022
SE2	1 972	3 550	2 467	7 989
SE3	2 557	3 845	16 869	23 270
Totalt	5 699	9 366	21 217	36 281

4 Slutsatser

Trafikanalys har tidigare dragit slutsatsen att publicering av tabell med på- och avlastningsregioner baserade på NUTS 2-region utgör risk för röjande av enskild. Den analys som gjorts här pekar mot att det i stort är en korrekt slutsats. Dock framgår att vissa uppgifter kan publiceras utan röjanderisk, vilket man bör göra när uppgifterna samlats in och finns tillgängliga.

Underlagen om järnvägstransportstransportstatistik har samlats in i enlighet med EU-parlamentets och rådet förordning (EU) nr 91/2003. Om det inte hade funnits någon sådan förordning, vore det inte rimligt att Trafikanalys samlar in underlagen genom en enkät. Orsaken är det innebär en onödig uppgiftslämnarbörda, se (SCB, 2016). Eftersom man kunnat förutse att uppgifterna i liten utsträckning kommer att vara möjliga att publicera efter röjandekontroll, bör man avstå från att samla in dem rutinmässigt. För kvalitetsuppföljning eller statistikutveckling kan det dock ibland vara motiverat att samla in uppgifter på en finare nivå än den som används vi publicering av statistik.

Om däremot uppgifter om transporter mellan regioner finns tillgängliga i ett register, kan Trafikanalys inhämta dessa och pröva i vilken utsträckning de går att publicera. Ett sådant register har då skapats för andra ändamål än statistik. Det är givetvis viktigt att ett sådant register håller tillräckligt hög kvalitet för statistikanvändning.

5 Referenser

Dandekar, R. A., och Cox, L.H. (2002). *Synthetic Tabular Data: An Alternative to Complementary Cell Suppression*. Opublicerat manuskript. Retrieved from https://www.researchgate.net/profile/Ramesh_Dandekar/publication/233817782_syn_tab/links/0912f50bdf09d5f8eb000000.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., och de Wolf P.-P. (2012). *Statistical Disclosure Control*. Chichester, IK: John Wiley & Sons.

ROS. (2015). *Handbok i statistisk röjandekontroll*. Retrieved from https://www.scb.se/contentassets/0cd92207266d40eb8829244d51d90b94/handbok-i-statistisk-rojandekontroll.pdf

SCB. (2016). *Kvalitet för den officiella statistiken - en handbok*. Retrieved from https://www.scb.se/contentassets/b3cf91d501e9466fae5bf1eae2a5484f/ov9999_2016a01_br_x99br1701.pdf

TRAFIK ANALYS

Trafikanalys är en kunskapsmyndighet för transportpolitiken. Vi analyserar och utvärderar föreslagna och genomförda åtgärder inom transportpolitiken. Vi ansvarar även för officiell statistik inom områdena transporter och kommunikationer. Trafikanalys bildades den 1 april 2010 och har huvudkontor i Stockholm samt kontor i Östersund.

Trafikanalys
Torsgatan 30
113 21 Stockholm

Tel 010 414 42 00
Fax 010 414 42 10
trafikanalys@trafa.se
www.trafa.se