**Abstract**

In a Swedish vehicle speed survey, for a stratified multi-stage sample of road sites, data are collected by use of pneumatic tubes and a traffic analyzer. From registered pulses, vehicles are created and assigned speeds. Typically, part of the vehicles passing a chosen site will remain unobserved. The failure to observe some vehicles is indicated on one hand by imputations automatically created by the analyzer, on the other by a small reported measurement efficiency (defined as the proportion of registered pulses that have been combined into vehicles). The main survey goal is to estimate the average speed, defined as the ratio $R$ of total vehicle mileage and total travel time. An undercount of vehicles is bound to bias the estimators of the totals, whereas the impact on the estimator of $R$ is unclear.

We suggest dividing the traffic passing each spot into weighting classes. The main difficulty is to adjust the observed flow upwards. Within class, one proposal is to add the number of imputed vehicles, another to weight the observed flow by an estimated probability of registration (assumed to be constant within class). The measurement efficiency is put forward as estimator of the registration probability. Models for the errors in the number of imputed vehicles, and in the estimated registration probabilities, are used for theoretical evaluations. The models are evaluated, and the suggested estimation strategies compared, by use of some empirical data. Our model assumptions seem to agree reasonably well with this data. Also, the adjustment estimators seem to produce less biased estimates of the totals than today's unadjusted estimators. None of them show its clear superiority to the other. It remains unclear if the estimator of average speed really needs any missing data adjustments.

**Key words:** Survey, vehicle speeds, missing data, weighting classes.

## Svensk sammanfattning

Varje år genomför Vägverket en undersökning av fordonshastigheter i tätort. Huvudsyftet med undersökningen är att uppskatta medelhastigheten på vägarna. Data samlas in genom mätningar i ett stickprov av platser, utvalda genom ett trestegsförfarande. Varje utvald plats observeras under ett dygn med utnyttjande av en mätutrustning, bestående av två luftslangar spända över vägen och anslutna till en trafikanalysator. Varje fordon som passerar över slangarna ger upphov till ett antal pulsregistreringar i apparaten. Pulserna översätts till fordon med individuella hastigheter. Idealt skall alla passerande fordon på detta sätt bli observerade, men risken finns att så i verkligheten inte blir fallet. Vissa trafiksituationer och körbeteenden komplicerar nämligen mätningen så att somliga pulser inte otvetydigt kan sägas tillhöra ett visst fordon. Utifrån sådana pulser skapas automatiskt ett antal osäkra eller *imputerade* fordon, som dock inte används i senare beräkningar. Som ett mått på kvaliteten i mätdata rapporteras, för varje klocktimme, mätningens *verkningsgrad*: andel inkomna pulser som ingår i korrekt registrerade (icke-imputerade) fordon. Om den genomsnittliga timverkningsgraden bedöms som varande alltför låg, underkänns hela mätningen och speciella åtgärder vidtas. Det typiska utfallet av en mätning, och det som behandlas i rapporten, är dock att ett visst bortfall förekommer (medelverkningsgraden understiger hundra procent), dock inte i sådan utsträckning att det leder till ett underkännande. I dagsläget ignoreras sådant 'smärre' bortfall helt.

Spelar bortfallet av fordon någon roll? Den populationsstorhet man vill uppskatta, medelhastigheten, definieras som kvoten $R = t_y/t_z$ mellan totala trafikarbetet $t_y$ och totala restiden $t_z$ för vägarna under undersökningsperioden. På basis av mätdata uppskattas först totalerna $t_y$ och $t_z$. Kvoten mellan dessa estimat används sedan som en uppskattning av $R$. Om alltför få fordon ingår i beräkningsunderlaget, kommer skattningarna av $t_y$ och $t_z$ med säkerhet att ha ett negativt systematiskt fel. Riktningen på det systematiska felet i skattningen av $R$ är däremot svårare att uttala sig om – teoretiskt sett kan det vara såväl positivt som negativt.

I rapporten diskuteras ett par förslag till bortfallsjusteringar av skattningarna av $t_y$, $t_z$ och $R$. Vi har valt att kalla förslagen Strategi 1 respektive 2. Förhoppningen är att man, genom utnyttjande av någon av

dessa, skall kunna motverka bortfallets eventuella snedvridande effekter på undersökningsresultaten. Båda strategierna förlitar sig på en enkel modell för den mekanism som gör att vissa fordon registreras: vår så kallade *registreringsmodell*. Registreringsmodellen säger att den trafik som passerar en mätplats kan delas in i homogenitetsgrupper sådana att

- alla fordon i gruppen har samma sannolikhet att registreras, och
- registreringen av ett givet fordon i gruppen är oberoende av registreringarna av övriga fordon.

Betrakta en given mätplats och homogenitetsgrupp. (För enkelhetens skull avstår vi här ifrån att indexera för plats och grupp.) Helst skulle vi vilja veta det sanna antalet passerande fordon $y$ och summan av deras restider $z = \sum_U x_v$, där $x_v$ är den tid fordon $v$ använder för att passera platsen, och vi summerar över mängden $U$ av passerande fordon. Om vi har bortfall i mätningarna känner vi endast antalet registrerade fordon samt summan av restiderna för dessa. Under Strategi 1 görs följande bortfallsjusteringar. Till det registrerade antalet fordon läggs det antal fordon som imputerats av trafikanalysatorn. Den resulterande uppskattningen av $y$ kallar vi $\hat{y}^{(1)}$. Resultatet används för att uppskatta $z$ med $\hat{z}^{(1)} = \hat{y}^{(1)}\bar{x}_r$, där $\bar{x}_r$ är de registrerade fordonens medelrestid. Under Strategi 2 används verkningsgraden som skattning av den okända registreringssannolikheten. Genom att dividera antalet registrerade fordon med verkningsgraden erhålles en alternativ skattning $\hat{y}^{(2)}$ av $y$. Enligt samma princip som under Strategi 1 uppskattas därefter $z$ med $\hat{z}^{(2)} = \hat{y}^{(2)}\bar{x}_r$. Vi noterar att Strategi 2 fordrar att homogenitetsgruppen utgörs av en enhet för vilken verkningsgraden är känd: i praktiken en klocktimme eller en grupp av klocktimmar.

Om vi inte hade något bortfall skulle vi uppskatta $t_y$, $t_z$ och $R$ genom att räkna upp de observerade $y$- och $z$-värdena för utvalda platser med hänsyn till hur urvalet gjordes. Våra bortfallsjusterade skattningar av populationsstorheterna utnyttjar samma formler, med den enda skillnaden att de (nu okända) $y$ och $z$-värdena ersätts med uppskattningar av desamma. För att kunna analysera Strategi 1-skattningarnas statistiska egenskaper (uttrycka deras väntevärde och varians) använder vi en *imputeringsmodell*: en enkel modell för hur antalet imputerade fordon förhåller sig till det sanna antalet oregistrerade fordon. För analys av Strategi 2-skattningarna använder vi

istället en modell för hur verkningsgraden förhåller sig till den sanna registreringssannolikheten: den så kallade *felmodellen för* $\hat{\theta}^{(2)}$. Om uppskattningarna av $y$ och $z$-värdena under respektive strategi är väntevärdesriktiga, blir även skattningarna av $t_y$ och $t_z$ väntevärdesriktiga (med avseende på såväl urvalsdesign som relevanta modeller). Bortfallet ger då inte heller upphov till något systematiskt fel i skattningen av $R$. För Strategi 1:s del skulle denna gynsamma situation inträffa om, för varje utvald plats och homogenitetsgrupp, det förväntade antalet imputerade fordon (givet antal registrerade fordon) sammanfaller med antalet registrerade fordon; för Strategi 2:s del om den förväntade verkningsgraden sammanfaller med den sanna registreringssannolikheten.

I syfte att utvärdera våra modeller, och i förlängningen de föreslagna justeringsmetoderna, genomfördes sommaren 2001 ett experiment. Data samlades in för fem mätplatser, fördelade på olika tätortsvägar i Linköpingsområdet, under ett dygn. Platserna valdes inte ut slumpmässigt, utan för att representera olika vägtyper. Vi begränsade oss till dubbelriktade vägar med hastighetsbegränsningen 50 km/h. I varje plats användes dubbla par av slangar. Det ena slangparet användes för att registrera trafiken på det sätt som normalt görs i hastighetsundersökningen. Det andra paret slangar, som monterades tätt intill det första, användes för att försöka samla in "sanna" värden (eller åtminstone data av högre kvalitet) att jämföra med. I mitten av dessa slangar monterades en *ventil*, och slangändarna på var sida av vägen anslöts till en trafikanalysator. Mätning utfördes på detta sätt separat för varje körfält; ett förfarande som förenklar översättningen av inkommande pulser till fordon (dels eftersom pulserna blir färre, och dels eftersom fordonens körriktning är känd). Trots detta uppstod ett visst bortfall, manifesterat i form av ett antal imputationer, även i ventilmätningarna. I analysen av experimentdata görs alla beräkningar såväl inklusive som exklusive dessa imputerade fordon: vi arbetar alltså med två olika uppsättningar av "sanna" värden.

Varken imputationsmodellen eller felmodellen för $\hat{\theta}^{(2)}$ kan utvärderas empiriskt i sin generella formulering, utan vi begränsar oss till några enkla specialfall. Betrakta först imputeringsmodellen med ett multiplikativt fel. Enligt denna modell ges antalet imputerade fordon som det sanna antalet oregistrerade fordon gånger ett slumpfel. Slumpfelen antas ha konstant väntevärde och varians, och dessa moment antas vara oberoende av antal reg-

istrerade fordon och av plats. Om slumpfelen har väntevärde ett är $\hat{y}^{(1)}$ och $\hat{z}^{(1)}$ väntevärdesriktiga för $y$ respektive $z$. Våra data tyder inte på beroende mellan slumpfelsmomenten och antalet registrerade fordon. Den osäkerhet i analysen som följer av att det förekommer imputationer i ventilmätningarna, gör att vi inte vågar uttala oss om eventuellt platsberoende. Av samma skäl har vi svårt att säga om $\hat{y}^{(1)}$ och $\hat{z}^{(1)}$ är väntevärdesriktiga för $y$ och $z$ eller ej.

Vi utvärderar såväl en additiv som en multiplikativ felmodell för $\hat{\theta}^{(2)}$. Under den additiva modellen ges den skattade registreringssannolikheten som den sanna sannolikheten plus ett slumpfel; under den multiplikativa modellen som den sanna sannolikheten gånger ett slumpfel. I båda fallen antas slumpfelen ha konstant väntevärde och varians, och dessa antas vara oberoende av antal registrerade fordon och av plats. Om slumpfelen under den additiva modellen har väntevärde noll är $\hat{y}^{(2)}$ och $\hat{z}^{(2)}$ approximativt väntevärdesriktiga för $y$ respektive $z$; detsamma gäller för den multiplikativa modellen om slumpfelen har väntevärde ett. Vi ser inga tydliga tecken på beroende mellan slumpfelsmomenten och antalet registrerade fordon. Däremot verkar det finnas ett platsberoende, vilket kan tyda på att modellerna är för enkla. Oavsett modell verkar $\hat{y}^{(2)}$ och $\hat{z}^{(2)}$ vara approximativt väntevärdesriktiga för $y$ och $z$.

Vi skattade $y$ och $z$ för var och en av de fem platser som ingick i experimentet, och jämförde med de ”sanna” värdena. Som väntat blir såväl flöde som restid genomgående underskattade om bortfallet helt enkelt ignoreras. Båda de föreslagna justeringsstrategierna verkar ge uppskattningar som kommer litet närmare ”sanningen”. Såväl Strategi 1 som 2 verkar alltså ha potential att minska bortfallsfelet.