Prediktioner av antalet personer som går resp. cyklar till jobbet på basis av resvaneundersökningar och geografiskt högupplösta registerdata

Gunnar Isacsson

Utgivare:

vti

581 95 Linköping

Publikation:
Notat

Utgivningsår:	**Projektnummer:**	**Dnr:**
2014	201153	2013/0134-7.4

Projektnamn:
Utveckling av GC-kalk

Författare:
Gunnar Isacsson

Uppdragsgivare:
Trafikverket

Titel:

Prediktioner av antal personer som går resp. cyklar till jobbet på basis av resvaneundersökningar och geografiskt högupplösta registerdata

Referat

Trafikverket har under senare år utvecklat ett verktyg för samhällsekonomisk lönsamhetsbedömning av investeringar i gång- och cykelinfrastrukturen – ”GC-kalk” (Trafikverket, 2012). Dessa bedömningar baseras bl.a. på antalet resande före investeringen och i manualen för GC-kalk föreslås att detta antal baseras på räkningar eller resvaneundersökningar. I denna rapport presenteras ett mer specifikt förslag på hur resvaneundersökningar kan kombineras med högupplösta registerdata för att generera bedömningar eller prediktioner över andelen arbetsresor som sker med färdmedlen gång och cykel.

Nyckelord: Gång, cykel, prediktioner, validering, bootstrap, resvaneundersökning, registerdata

ISSN:	**Språk:**	**Antal sidor:**
0347-6030	Svenska	44

Publisher:

vti

SE-581 95 Linköping Sweden

Publication:
Notat

Published:	**Projectcode:**	**Dnr:**
2014	201153	2013/0134-7.4

Project:
Utveckling av GC-Kalk

Author:
Gunnar Isacsson

Sponsor:
The Swedish Transport Administration

Title:

Predictions of the number of persons that walk or bicycle to work using travel surveys together with register based data with high geographic resolution

Abstract

The Swedish Transport Administration has recently developed a tool for assessing investments in the infrastructure for walking and bicycling. These assessments are *inter alia* based on the number of individuals that walk or bicycle before the investment. According to the manual of this tool, the number of individuals that walk or bicycle may be assessed by using travel surveys and/or counts. This report presents a suggestion on how travel surveys can be combined with register based (administrative) data of high geographic resolution to generate predictions of the number walking or cycling to work in a specific origin and destination pair.

Keywords: Walk, bicycle, prediction, validation, bootstrap, travel survey, register based data

ISSN:	**Language:**	**No of pages:**
0347-6030	Swedish	44

Förord

VTI har på uppdrag av Trafikverket testat att använda resvaneundersökningen i kombination med registerdata över individer och arbetsställen med geokodad information om lokaliseringen av bostad och arbetsplats för att generera bedömningar av antalet person som går eller cyklar till jobbet på låg geografisk nivå. Den geokodade informationen avser en indelning av Sverige i rutor om 250 kvadratmeter i tätort och 1000 kvadratmeter utanför tätort. Kontaktperson på Trafikverket har varit Mulugeta Yilma.

Borlänge mars 2014

Gunnar Isacsson, projektledare

Process för kvalitetsgranskning

Granskningsseminarium genomfört 2014-03-12 där Reza Mortazavi var lektör. Gunnar Isacsson har genomfört justeringar av slutligt rapportmanus 2014-03-30. Projektledarens närmaste chef (namn) har därefter granskat och godkänt publikationen för publicering (datum). De slutsatser och rekommendationer som uttrycks är författarens/författarnas egna och speglar inte nödvändigtvis myndigheten VTI:s uppfattning.

Process for quality review

Review seminar was carried out on March 12, 2014 where Reza Mortazavi reviewed and commented on the report. Gunnar Isacsson has made alterations to the final manuscript of the report. The research director of the project manager (name) examined and approved the report for publication on (date). The conclusions and recommendations expressed are the author/authors and do not necessarily reflect VTI's opinion as an authority.

Tryckt på VTI, Linköping, 2014

Innehållsförteckning

Prediktioner av antal personer som går resp. cyklar till jobbet på basis av resvaneundersökningar och geografiskt högupplösta registerdata

av Gunnar Isacsson
VTI, Statens väg- och transportforskningsinstitut
581 95 Linköping

Sammanfattning

Trafikverket har under senare år utvecklat ett verktyg för samhällsekonomisk lönsamhetsbedömning av investeringar i gång- och cykelinfrastrukturen – ”GC-kalk” (Trafikverket, 2012). Dessa bedömningar baseras bl.a. på antalet personer som går eller cyklar i en viss OD-relation i det jämförelsealternativ (JA) som används. Denna rapport använder den svenska resvaneundersökningen (RES) i kombination med SCBs registerdata över den svenska befolkningen och alla arbetsställen för att på ett *enkelt* sätt producera en prediktion av antalet fotgängare resp. cyklister i ett litet geografiskt område. Grundproblemet är alltså att registerdatamaterialet inte innehåller information om vilket färdmedel individerna väljer för att ta sig till jobbet. Därför används RES för att ”fylla i” denna information. Detta görs med hjälp av en statistisk modell som estimeras på RES och sedan tillämpas på registerdatamaterialet.

En utgångspunkt för denna rapport är att investeringar i gång- och cykelinfrastruktur är så pass små att mer genomarbetade och resurskrävande modellkörningar som brukar användas för större investeringar i järnvägs- och vägnäten inte är aktuella för investeringar i detta sammanhang. Då färdmedlen gång och cykel används i relativt hög utsträckning för korta resor är troligen en relativt fin geografisk indelning också att föredra när man utreder åtgärder i gång- och cykelinfrastrukturen. I denna rapport används därför en geografisk indelning som baseras på ett rutnät med 250 kvadratmeter stora rutor i tätort och en kvadratkilometer stora rutor utanför tätort. En viktig restriktion för denna rapport är att det i dagsläget inte finns nationellt heltäckande information om nätverken för gång- och cykelvägar, därför har sådan information inte kunnat användas.

En av huvudfrågorna för denna rapport är hur bra prediktioner för färdmedelsval man kan få från en modell som estimerats på en nationell resvaneundersökning. Därför utvärderas ett antal olika modeller med en form av korsvalidering som baseras på s.k. ”bootstrap” metod. Denna metod innebär också att den del av prediktionsfelet i registerdata vilket beror på stickprovsvariation i resvaneundersökningen enkelt kan simuleras för varje individ i registerdatamaterialet.

Resultaten tyder på att prediktionskvaliteten från den modell som bedöms ge bäst prediktioner sett över hela landet varierar mellan olika län. Den valda modellen används också för att illustrera hur prediktioner på låg geografisk nivå kan genereras. Dessutom diskuteras hur osäkerheten i de individuella prediktionerna kan användas för att väga in annan information som kan finnas när antalet personer som går eller cyklar till jobbet ska bestämmas i en specifik utredningssituation och i en specifik OD-relation.

Predictions of the number of persons that walk or bicycle to work using travel surveys and register based data with high geographic resolution

by Gunnar Isacsson
The Swedish National Road and Transport Research Institute (VTI)
SE-581 95 Linköping

Summary

The Swedish Transport Administration has recently developed a tool ("GC-kalk") for assessing investments in the infrastructure for walking and bicycling. These assessments are *inter alia* based on the number of individuals that walk or bicycle before the investment. This report uses the Swedish national travel survey combined with register based (administrative) data with high geographic resolution to generate predictions of the number walking or cycling to work in a small geographic area. The basic problem is that the register based material that includes the entire Swedish population and all work places does not include information about mode of transport for the journey to work. Therefore the travel survey is used to fill in this information. This is done with a statistical model estimated on the travel survey and used on the register based material.

A presumption of this report is that investments in the infrastructure for walking and cycling are so small that models usually used for investments in the infrastructure for road and rail transports are too resource demanding to be applicable in the context of walking and cycling. Since walking and cycling primarily are used for short distances it is, furthermore, likely preferable to use a high geographic resolution when assessing investments in the infrastructure for walking and cycling. In this report the geographic delineation of Sweden is based on squares with an area of 250 square-meters in urban areas and an area of one square-kilometer in rural areas. An important restriction of this report is that there is currently no information available on the networks for walking and bicycling with national coverage. Thus, no such information has been used here.

A main question of this report concerns the quality of predictions on mode choice that you get from a model estimated on a national travel survey. Therefore a set of models are evaluated by cross validation based on the so-called "bootstrap" method. This method also implies that the uncertainty in mode choice predictions that depend on the sample variation in the travel survey easily can be simulated for all individuals in the register based data.

The results suggest that the quality of predictions derived from the model that gives the best predictions at the national level vary across different counties. The chosen model is also used to illustrate how predictions of the shares of individuals walking or bicycling to work in a small area can be generated. In addition, the report discusses how the uncertainty in individual predictions can be used to consider other available information when the number of persons walking or bicycling shall be assessed in a specific situation and in a specific origin-destination pair.

1 Inledning

Trafikverket har under senare år utvecklat ett verktyg för samhällsekonomisk lönsamhetsbedömning av investeringar i gång- och cykelinfrastrukturen – ”GC-kalk” (Trafikverket, 2012). Dessa bedömningar baseras dels på hur många individer som berörs av en investering, dels på en monetär bedömning av olika typer av värden för dessa individer av investeringen. Antalet berörda individer avser här de som cyklar resp. går mellan en start- och en målpunkt (en OD-relation) före investeringen genomförs samt motsvarande antal efter det att investeringen genomförts, d.v.s. efterfrågan på resp. färdmedel före och efter investeringen.

Ett par svenska studier av hur man ska bedöma värdet av åtgärder i gång- och cykel-infrastrukturen har nyligen genomförts (Börjesson & Eliasson, 2012, Björklund & Carlén, 2012, Björklund m.fl., 2013 och Björklund & Mortazavi, 2013). Dessa studier behandlar bl.a. värdet av insparad restid och hälsoeffekter av cykling. Dessutom finns det numera ett par studier av hur efterfrågan för cykel förändras då en specifik typ av investering genomförs (Wardman m.fl., 2007 samt Björklund & Isacsson, 2013). Dessa studier baseras bl.a. på enkäter där respondenterna har fått ange hur de skulle välja mellan tydligt specificerade alternativ om valet hade varit på riktigt. På basis av dessa går det alltså att göra en bedömning av hur efterfrågan förändras efter investeringen genomförts. Tidigare sammanställningar av hur efterfrågan påverkas av olika åtgärder i cykelinfrastrukturen tyder på stor osäkerhet i hur den förändras (Naturvårdsverket, 2005, samt WSP, 2007).

För att kunna avgöra hur efterfrågan i en OD-relation förändras av en specifik investering måste man ha en bedömning av hur många personer som går eller cyklar i relationen från början, d.v.s. i det jämförelsealternativ (JA) GC-kalkmanualen diskuterar. Enligt manualen ska man ”Ange hur många resor som görs på varje länk i JA, totalt i båda riktningar.” I manualen föreslås vidare att: ”Underlagsdata kan ibland tas från räkningar eller resvaneundersökningar, men i de flesta fall måste de kompletteras med (och ibland ersättas av) bedömningar.” (Trafikverket, 2012, sid. 10). Relevanta frågor i detta sammanhang är hur resvaneundersökningar mer specifikt kan användas och hur bra bedömningar man därigenom kan få.

Syftet med denna rapport är att illustrera hur resvaneundersökningar i kombination med SCBs registerdata över hela den svenska befolkningen och alla arbetsställen på ett *enkelt* sätt kan användas för att producera en bedömning/prediktion av antalet fotgängare resp. cyklister i olika delar av landet. Eftersom resvaneundersökningar baseras på urval av befolkningen uppstår även frågan hur bra prediktionerna blir när de ska användas för en specifik OD-relation. Ett generellt problem i detta sammanhang är att urvalen är så små att antalet observationer i en specifik relation inte räcker till för att göra en bra bedömning av hur antalet personer som reser i relationen fördelar sig på olika färdmedel. Istället måste man förlita sig på en estimerad modell som relaterar färdmedelsval till ett antal observerbara egenskaper för resan, orten och individen och använda denna modell för att göra bedömningen. SCBs registerdata innehåller en stor mängd socio-ekonomiska karaktäristika på individnivå och kan även kombineras med detaljerad information om var varje individs bostad och arbetsplats är lokaliserad. I

registerdatamaterialet känner man dock inte till vilket färdmedel en individ väljer för att resa till arbetet. Men genom att välja variabler som finns i både resvaneundersökningen och i registerdatamaterialet går det att estimera en färdmedelsvalsmodell på resvaneundersökningen och använda den för att göra en prediktion för varje individs färdmedelsval i registerdatamaterialet. Eftersom SCBs registerdatamaterial innehåller uppgifter för hela den svenska befolkningen innebär detta alltså att man kan få en individbaserad prediktion för varje individs val av färdmedel för resan till jobbet. Syftet med denna rapport är alltså att genomföra detta.

En utgångspunkt för denna rapport är att investeringar i gång- och cykelinfrastruktur är så pass små att mer genomarbetade och mer resurskrävande prediktioner av den typ som genomförs med exempelvis SAMPERS (Algers & Beser, 2000) inte är aktuella i sammanhanget. Dessutom används färdmedlen gång och cykel i relativt hög utsträckning för korta resor. Därför är sannolikt en finare geografisk indelning än den som används i SAMPERS att föredra i detta sammanhang. Därför ligger betoningen i denna rapport på att bedömningar/prediktioner baserade på resvaneundersökningar för JA i GC-kalk ska vara *enkla* att ta fram och att de ska avse prediktioner i högupplösta geografiska data över var individer bor resp. jobbar. I denna rapport används en geografisk upplösning som baseras på ett rutnät med 250 kvadratmeter stora rutor i tätort och en kvadratkilometer stora rutor utanför tätort.

En viktig förutsättning för denna rapport är att det i dagsläget inte finns heltäckande information om nätverken för gång- och cykelvägar. Därmed kan man inte veta hur utformningen av infrastrukturen för gång och cykel ser ut för respondenterna i den resvaneundersökningen som används här. Det pågår ett arbete med att lägga in information om gång- och cykelinfrastrukturen i den nationella vägdatabasen. Men då detta arbete inte är avslutat då denna rapport skrivs har inte den informationen kunnat tas med i de analyser som presenteras i det följande. Men det kan vara värt att undersöka möjligheterna att koppla samman denna information med resvaneundersökningar i framtida utvecklingsarbete av GC-kalk. Detta skulle kunna bidra till mer precisa prediktioner än vad som kan åstadkommas utan denna information. Men information om infrastrukturens utformning i en utredning av en specifik åtgärd i ett visst område kan vägas in genom att anpassa prediktionen beroende på hur infrastrukturen är utformad i utgångsläget (JA). Detta diskuteras i uppsatsen.

En av huvudfrågorna för denna rapport är alltså hur bra lokala prediktioner för färdmedelsval man kan få från en modell som estimerats på en nationell eller regional resvaneundersökning. Med ”lokala” avses här prediktioner av befolkningens färdmedelsval i ett relativt litet avgränsat geografiskt område. Resvaneundersökningar är som sagt urvalsundersökningar och stickprovsstorleken på dessa utgör en begränsning för hur väl man kan prediktera färdmedelsandelar på lokal nivå. Ytterst handlar detta om hur väl modellen lyckas prediktera färdmedelsval för varje individ. Det finns ett antal faktorer som kan ge skillnader mellan den faktiska fördelningen av antal resande med färdmedlen gång och cykel dels ren slumpmässig variation (”stickprovsvariation”), dels systematiska skillnader mellan faktiska och modellberäknade val (”bias”). Systematiska skillnader kan uppstå p.g.a. faktorer som inte beaktas i modellen men som är viktiga för val av färdmedel i praktiken.

I allmänhet gäller att stickprovsvariationen är lägre ju större stickprov som används för att estimera modellen men det är däremot inte självklart hur bias påverkas av stickprovsstorleken. Men det kan finnas en trade-off mellan stickprovsstorlek och bias i de modeller som presenteras i denna rapport. Skälet till detta är att icke-observerade, platsspecifika faktorer som inte har beaktats i modellen kan ge upphov till systematiska skillnader mellan predikterade och faktiska färdmedelsandelar. Icke-observerade faktorer inkluderar t.ex. den grundläggande benägenheten att gå eller cykla i den lokala befolkningen, turtäthet för kollektiva färdmedel och den exakta utformningen av nätverken för gång och cykel. Man kan förmoda att ju finare geografisk uppdelning av resvaneundersökningen man använder desto mindre blir den bias som beror av icke-observerade platsspecifika faktorer. Men samtidigt blir stickprovsvariationens betydelse för modellens prediktionsfel större ju finare geografisk uppdelning man försöker använda. Detta beror alltså på att stickprovsstorleken per geografisk enhet minskar ju finare geografisk indelning man vill använda. Detta är i huvudsak ett exempel på s.k. överanpassning (jfr ”overfitting”) av en modell till data vilket innebär att modellen är bra på att prediktera valen i stickprovet men samtidigt ger dåliga prediktioner utanför stickprovet.

Överanpassning av modeller till data är ett generellt problem då man estimerar modeller för prediktionsändamål. Om man utvärderar modellerna på samma stickprov som man skattat modellen på tenderar prediktionskvaliteten att verka bättre än vad den egentligen är (se t.ex. Efron, 1986). Man blir m.a.o. alltför optimistisk om modellens kvalitet. För att hantera detta problem kan man använda någon form av korsvalidering. Detta innebär att man delar upp stickprovet i två delar: ett ”träningsstickprov” och ett ”valideringsstickprov” och använder träningsstickprovet för att estimera modellen och ”valideringsstickprovet” för att bedöma modellens kvalitet i termer av dess förmåga att generera ”bra” prediktioner. En specifik variant av korsvalidering är den s.k. ”leave-one-out” där en observation åt gången utelämnas från det ursprungliga stickprovet för att skapa ett antal träningsstickprov (lika många som antalet observationer i det ursprungliga stickprovet, n)[1]. Då skattas modellen n st. gånger och vi får n st. observationer på prediktionsfelet. För en kvalitativ variabel som färdmedelsval är denna estimator av modellens sanna prediktionsfel väntevärdesriktig men den har en stor spridning jämfört med s.k. ”bootstrap” metoder (Efron & Tibshirani, 1997, Efron, 1986, Efron, 1983). ”Bootstrap” metoden innebär i korthet att man drar upprepade stickprov med återläggning från det stickprov man arbetar med (här är det resvaneunder-sökningen).

Den bootstrap metod Efron & Tibshirani (1997) föreslår (0,632+ estimatorn) baseras i huvudsak på en ”utjämnad” (jfr ”smoothed”) variant av korsvalidering där bootstrap används för ”utjämningen”. Den har visat sig fungera bra i ett antal olika studier (se t.ex. Efron & Tibshirani, 1997, Ambroise & McLachlan, 2002, Steyerberg m.fl., 2001, 2003, och Leeb, 2008). Därför används denna estimator av färdmedelsvalsmodellernas prediktionsfel för att välja modell. En fördel med att använda bootstrap-metoden här är också att den del av prediktionsfelet i registerdatamaterialet som beror på stickprovsvariation i RES enkelt kan simuleras för varje individ i registerdatamaterialet.

[1] Denna kallas ibland även för ”Jack-knife”.

Analyserna som presenteras i denna rapport baseras på den nationella resvaneundersökningen från 2005-2006. Men tillvägagångssättet för att ta fram en bedömning av resandet i JA bör även vara tillämpbart på regionala/lokala resvaneundersökningar. Fokus i denna rapport ligger på arbetsresor, d.v.s. en individs resa till och från arbetsstället. Skälet till detta är att SCBs registerdata innehåller information om var individen bor och var hon/han arbetar men inte om var han/hon handlar eller var hon/han hämtar och lämnar barn på skola/fritidshem/daghem. Men resor till och från arbetet står för en stor del av antalet resor och en del av de värderingar som används i GC-kalk är kopplade till arbetsresor.

Återstoden av rapporten är disponerad på följande sätt. Avsnitt 2 presenterar tillvägagångssätt och metoder som används för analyserna. Avsnitt 3 presenterar datamaterialen och avsnitt 4 presenterar resultat. Slutsatser och förslag på fortsatta förbättringar av de prediktioner som presenterats i rapporten återfinns i avsnitt 5.

2 Metod

I det första steget för att göra en bedömning av antalet arbetsresande som går resp. cyklar i en OD-relation tar man fram en bedömning av hur många individer som reser i den relationen. Om man inte har denna information sedan tidigare kan man t.ex. beställa den av SCB. Om OD-relationen t.ex. avser två stycken kvadratkilometer stora rutor (A och B) så kan informationen avse antalet personer som bor i A och arbetar i B. Denna information kan även kompletteras med socioekonomiska karaktäristika för dessa personer och i vilken kommun A och B är lokaliserade. I denna rapport används ett registerbaserat datamaterial från SCB för detta ändamål. I detta material observeras alltså inte individens färdmedelsval och därför behöver man göra en bedömning av hur många individer som använder resp. färdmedel.

För att genomföra bedömningen av hur stor andel av individerna som går eller cyklar används resvaneundersökningen (RES). Bedömningen kan vara väldigt enkel, t.ex. färdmedelsandelar i RES för arbetsresor i det län (eller kommun om stickprovsstorleken tillåter det) där A och B är lokaliserade, eller i form av en modell som relaterar färdmedelsval till reseavståndet, län, samt socioekonomiska karaktäristika. En restriktion i detta sammanhang är att de variabler som ingår i modellen måste finnas tillgängliga i registerdatamaterialet

Givet denna restriktion är frågan vilka variabler som ska användas för att prediktera färdmedelsvalet. För att välja ut de variabler som ska ingå i modellen för resp. färdmedel används i det följande en procedur där en variabel i taget läggs till modellen på basis av vilket p-värde den har och om en i modellen inkluderad variabel understiger ett visst p-värde så utelämnas den från modellen (jfr ”forward selection”). Denna procedur leder alltså succesivt fram till ett antal variabler som används för att prediktera färdmedelsvalet. Risken är dock att man får en modell som är överanpassad till de data som använts för att estimera modellen. För att undvika detta används ett mått på hur väl modellen predikterar färdmedelsval i en del av stickprovet som inte har använts för att estimera modellen. Detta mått ligger till grund för val av modell. Ett par andra mått som beskriver anpassningen av modellen till datamaterialet presenteras också som komplement till måttet som baseras på modellens prediktionsförmåga. Därefter används den valda modellen för att prediktera färdmedelsval för de individer som bor i A och arbetar i B. I det följande kallas det sätt på vilket RES används för ”modell” och bedömningen av antalet resande med olika färdmedel eller motsvarande andel kallas för ”prediktion”.[2]

Återstoden av detta avsnitt är upplagt enligt följande. Först presenteras vilka modeller som estimeras på RES (avsnitt 2.1) och hur variabler för dessa modeller valts ut. Därefter beskrivs utvärderingen av vilken modell som är ”bra” (avsnitt 2.2). Slutligen beskrivs hur den utvalda modellen används på registerdatamaterialet från SCB för att generera prediktioner över antalet personer som går resp. cyklar till jobbet (avsnitt 2.3).

[2] I en del av litteraturen används ibland ”regel” istället för ”modell”.

2.1 Modeller

Alla modeller estimeras som en multinomial logit modell där fyra färdmedelsval beaktas: gång, cykel, kollektiv färdmedel samt motoriserat färdmedel (i huvudsak bil). Modellerna kan alltså beskrivas på följande sätt.

$$p_{im} = \frac{exp(\boldsymbol{z}_i'\boldsymbol{\delta}_m)}{\sum_{m=1}^{M} exp(\boldsymbol{z}_i'\boldsymbol{\delta}_m)} \quad (1)$$

där p_{im} är sannolikheten att individ i (i=1, 2, …, N) väljer färdmedel m (m=1, 2, 3, 4). Där de ”förklarande” variablerna i $\boldsymbol{z}_i$ varierar beroende på modell. Då modellen har estimerats har motoriserat färdmedel använts som referensalternativ.

De ”förklarande” variabler som har använts i denna uppsats är: avståndet för resan (mer specifikt användes logaritmen för avståndet), det kvartal då resan genomfördes, indikatorvariabler för det län i vilket individen bor samt individens ålder, inkomst, indikatorvariabel för kön och indikatorvariabler för individens utbildningsnivå. Dessutom inkluderas en indikatorvariabel som beskriver tillgång till en privatägd bil i hushållet och en indikatorvariabel för om individen räknar med att göra avdrag för bil i deklarationen. Dessutom används kommungenomsnitt för ett tillgänglighetsmått till jobb. Det har beräknats på registerdatamaterialet och avser: (i) antal jobb mellan 0 och 5 kilometer ifrån individens bostad, (ii) antalet jobb mellan 5 och 25 kilometer från individens bostad, (iii) antalet jobb mellan 25 och 50 kilometer från individens bostad samt (iv) antalet jobb mellan 50 och 100 kilometer ifrån individens bostad.

Argumenten för att använda dessa variabler i detta sammanhang är följande. Avståndet mellan bostaden och arbetsplatsen är sannolikt en starkt avgörande faktor för om man går eller cyklar istället för ta bilen eller åka kollektivt. Informationen om vilket kvartal resan genomfördes fångar upp säsongsmässiga variationer i valet att gå eller cykla; ju kallare och snöigare det är ju mindre troligt verkar det att en individ väljer att gå eller cykla. Informationen om vilket län individen bor i hanterar bl.a. klimatologiska skillnader mellan olika delar av landet vilka kan vara väsentliga för valet att gå och cykla. Den hanterar också på ett *grovt* sätt både variationer i trängsel i vägnätet i olika delar av landet samt utbud av kollektivtrafik. Socioekonomiska karaktäristika som ålder, inkomst, kön och utbildningsnivå kan också spela en viss roll för färdmedelsval. Tillgång till bil i hushållet och planer på att göra avdrag för bilresor till jobbet är sannolikt också viktiga för valet att använda bil som färdmedel till jobbet. Här kan vi notera att planer på att göra avdrag för bilresor motsvaras av faktiska avdrag för bilresor i registerdatamaterialet.

Motivet för att använda tillgänglighetsmått till jobb i kommunen där man bor är att detta kan fånga upp mellankommunal variation i förtätning av bebyggelse, jobb och service vilken kan vara relevant för färdmedelsval. Liss och Isacsson (2014) visar att denna information korrelerar med individers bilinnehav och bilanvändning vilket överensstämmer med internationell forskning om hur bebyggelsetäthet påverkar

bilanvändning (se t.ex. Bento m.fl., 2005, Brownstone & Golob, 2009, samt Newman & Kenworthy, 1989). Rietveld & Daniel (2004) visar f.ö. att skillnader i kommuners cykelpolicy har betydelse för andelen cyklande i kommunen. Men här har inte sådana policyvariabler funnits tillgängliga.

Det finns förstås många andra faktorer som är viktiga för om man går eller cyklar till jobbet; t.ex. topografiska förutsättningar som avgör hur backigt det är. Sådana faktorer är dock icke-observerade i de datamaterial som används här och kan därför inte inkluderas i analysen.

För att välja ut vilka förklarande variabler som ska ingå i modellen har en procedur använts där en variabel i taget har inkluderats i modellen på basis av vilket p-värde den har. Dessutom har beslut om vilka variabler som ska behållas i modellen baserats på det p-värde de har haft efter det att andra variabler har inkluderats i modellen. Denna procedur brukar kallas ”forward selection” i litteraturen. Här har p-värdet för att inkludera en variabel i modellen satts till 5 procent och p-värdet för att behålla en inkluderad variabel i modellen har satts till 10 procent. Eftersom en multinomial logit-modell kan estimeras som en uppsättning binära logitmodeller (se t.ex. Allison, 1999, s.122-123) så har denna procedur genomförts separat för var och en av modellerna: gång och motoriserat färdmedel, cykel och motoriserat färdmedel, samt kollektivt färdmedel och motoriserat färdmedel. (Observera att motoriserat färdmedel är referensalternativ i den multinomiala logitmodellen i ekvation 1.) Detta innebär att effektiviteten i skattningarna i proceduren var lägre än om motsvarande multinomiala logit hade estimerats men ansatsen medgav samtidigt en viss flexibilitet att välja ut vilka variabler som var relevanta för resp. färdmedel. När modellerna utvärderades i termer av prediktionsförmåga användes dock motsvarande multinomiala logitmodell. För att tydliggöra detta genom ett exempel, proceduren innebar att avstånd var den första variabel som inkluderades i modellen för gång och motoriserat färdmedel. Avstånd var även den första variabel som inkluderades i modellen för cykel och motoriserat färdmedel, medan tillgång till bil i hushållet var den första variabel som inkluderas i modellen för kollektivt färdmedel och motoriserat färdmedel. Den första multinomiala logitmodellen inkluderade därför i det första steget variabeln avstånd för alternativen gång och cykel och variabeln tillgång till bil i hushållet för alternativet kollektiv färdmedel. Prediktionerna från denna modell utvärderades därefter med de mått som beskrivs i avsnitt 2.2. I nästa steg inkluderades de variabler som proceduren valde ut för resp. modell nummer 2 och den motsvarande multinomiala logitmodellen estimerades och utvärderades. Proceduren innebar att sammanlagt 18 olika modeller estimerades och utvärderades.

2.2 Utvärdering av modellerna

Det är välkänt att de genomsnittliga predikterade sannolikheterna för de olika färdmedelsvalen i en logitmodell exakt replikerar de i stickprovet observerade frekvenserna för resp. färdmedel. Detta säger dock inget om hur väl de predikterade sannolikheterna från den skattade modellen överensstämmer med de observerade frekvenserna för olika delar av stickprovet och än mindre för predikterade val på

individnivå. Om modellen, t.ex. är skattad på data för hela landet så är det inte säkert att den ger bra prediktioner i Skåne. Detta kan bero på slumpmässig variation från stickprovet men det kan också bero på att modellen inte lyckas beakta specifika och i datamaterialet icke-observerade förutsättningar i Skåne vilket kan leda till en bias i prediktionen. Då kanske man kan tycka att man bara ska använda data från Skåne för att estimera modellen antingen genom att bara välja ut observationer från detta län eller genom att genomföra en egen resvaneundersökning där. Det första alternativet innebär att man får för ett för litet stickprov om observationerna tas ifrån RES och det andra alternativet är kanske bra men grundproblemet kvarstår ändå om man behöver prediktera färdmedelsval för en specifik OD-relation i Skåne. Därför kan en av de färdmedelsvalsmodeller som beskrevs i avsnitt 2.1 vara användbar. Men frågan är vilken av dessa som är bäst för att prediktera antalet som går resp. cyklar till arbetet.

För att besvara denna fråga används tre olika mått. Det första av dessa baseras på modellens prediktionsfel som här estimeras med en metod som föreslagits av Efron & Tibshirani (1997). Detta mått på prediktionsfelet beräknas dels för hela stickprovet, dels för varje län för att undersöka om en och samma modell verkar vara ”bäst” för alla län eller om den ”bästa” modellen skiljer sig åt mellan olika län.

Den metod som Efron & Tibshirani (1997) föreslagit i detta sammanhang kan kort beskriva enligt följande. Låt y_i beteckna det faktiska valet och låt $r_x(t)$ beteckna det predikterade valet då stickprov $\boldsymbol{x}$ har använts för att estimera modellen och t avser ett specifikt värde på de ”förklarande” variablerna som används i modellen. Definiera först det synbara (jfr ”apparent”) prediktionsfelet:

$$\overline{err} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r_x(t)]$$

där

$$Q[y_i, r_x(t)] = \begin{cases} 0, om\ y = r \\ 1\ om\ y \neq r \end{cases}$$

och $\boldsymbol{x}$ avser det ”ursprungliga” stickprovet. Detta är alltså avvikelser mellan observerade och predikterade val för varje individ i stickprovet då det ursprungliga stickprovet används både för estimering och prediktion. Det synbara prediktionsfelet är en underskattning av det sanna prediktionsfelet. Underskattningen blir mer påtaglig då stickprovet är litet (se Efron, 1986). Definiera därefter felet för utelämna-en-bootstrap (jfr ”leave-one-out bootstrap”)

$$Err^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{E}_i$$

där

$$\hat{E}_i = \sum_b I_i^b Q_i^b \Big/ \sum_b I_i^b$$

$b = 1, 2, \ldots, B$ betecknar bootstrap-replikerat stickprov och

$$I_i^b = \begin{cases} 1, om\ individ\ i\ inte\ \text{är}\ med\ i\ bootstrap\ b \\ 0, om\ individ\ i\ \text{är}\ med\ i\ bootstrap\ b \end{cases}$$

och

$$Q_i^b = Q\left[y_i, r_{x^{*b}}(t)\right] = \begin{cases} 0, om\ y = r \\ 1\ om\ y \neq r \end{cases}$$

och $\boldsymbol{x}^{*b}$ är bootstrap-stickprov b som inte inkluderar individ i. Då sannolikheten att varje individ ingår minst en gång i ett bootstrap stickprov är omkring 0,632 så är $\text{Err}^{(1)}$ en överskattning av det sanna prediktionsfelet men med lägre varians än den estimator som baserar sig på en s.k. korsvalidering som baseras på att utelämna en observation i taget (se Efron & Tibshirani, 1997 för en förklaring). Därför föreslog Efron (1983) följande estimator av prediktionsfelet.

$$\widehat{Err}^{(,632)} = 0{,}368\overline{err} + 0{,}632\widehat{Err}^{(1)}$$

För att korrigera bias för prediktionsmodeller som är kraftigt överanpassade till data föreslog Efron & Tibshirani (1997) följande estimator av prediktionsfelet

$$\widehat{Err}^{(,632+)} = \widehat{Err}^{(,632)} + \left(\widehat{Err}^{(1)} - \overline{err}\right)\frac{0{,}368 * 0{,}632 * \hat{R}'}{1 - 0{,}368\hat{R}'} \qquad (2)$$

Där $\hat{R}'$ är en skattning av den relativa överanpassningsgraden. Denna senare estimator av prediktionsfelet används i denna uppsats. Varje bootstrap baseras på ett stratifierat stickprov med återläggning från RES, där strata är resp. län. Därigenom uppnås samma stickprovsstorlek för resp. län i varje bootstrap replikering b.

De andra två måtten baseras istället på hur bra anpassning modellen har till de data som använts för att estimera den med en viss ”bestraffning” för att använda för många variabler: Akaikes informations kriterium (AIC) och det s.k. Bayesianska informationskriteriet (BIC). Båda dessa baseras på ”log-likelihood-funktionens” värde

men det senare ”bestraffar” överanpassning av modellen till data hårdare än det förra. AIC och BIC beräknas med uttrycken:

$$AIC = -2lnL + 2q$$

$$BIC = -2lnL + qln(N)$$

där *lnL* är värdet på ”log-likelihood-funktionen”, *q* är antalet estimerade parametrar i modellen och *N* är storleken på stickprovet som använts för att estimera modellen. Dessa två mått är vanligt förekommande för att utvärdera en modells anpassning till ett datamaterial (se t.ex. Cameron & Trived, 2005, s. 278-279). Dessa två mått beräknas bara för hela stickprovet. Observera att inget av dessa mått baseras på hur bra prediktioner modellen ger. Eftersom modellen ska användas för prediktion av antalet personer som går resp. cyklar i registerdatamaterialet redovisas de framför allt som kompletterande information till det mått som beskrevs tidigare.

2.3 Tillämpning av modellen på registerdatamaterialet

Modellen med lägst prediktionsfel enligt Efron & Tibshiranis estimator används därefter på registerdatamaterialet för att få fram individbaserade prognoser. En fördel med den tidigare beskrivna bootstrap metoden är att den genererar *B* stycken skattade modeller. Detta innebär att man för varje individ kan generera *B* stycken prediktioner. Därigenom kan man för varje individ få en uppskattning av stickprovsvariationens betydelse för prediktionen genom att beräkna standardavvikelsen för resp. individs prediktion. Detta ger alltså ett mått på den stickprovsrelaterade osäkerheten i prediktionen vilket kan vara informativt vid tillämpningen av prediktionen i GC-kalk. Mer specifikt estimeras först sannolikheter för resp. färdmedel enligt följande:

$$\hat{p}_{im}^{b} = \frac{exp\left(\boldsymbol{z}_i'\widehat{\boldsymbol{\delta}}_m^b\right)}{\sum_{m=1}^{M} exp\left(\boldsymbol{z}_i'\widehat{\boldsymbol{\delta}}_m^b\right)} \tag{5}$$

där $\hat{p}_{im}^{b}$ är den prognostiserade sannolikheten att individ i med karaktäristika $\mathbf{z}_i$ väljer färdmedel *m* betingat på de skattade parametrarna $\hat{\boldsymbol{\delta}}_m^b$ från bootstrap-stickprov *b*. Därefter predikteras individens färdmedelsval på basis av det alternativ som har högst värde på $\hat{p}_{im}^{b}$. Osäkerheten i de predikterade sannolikheterna och i de predikterade valen estimeras därefter med standardavvikelserna för $\hat{p}_{im}^{b}$ och det predikterade valet. Då registerdatamaterialet avser årsdata används genomsnitten för antal observationer i RES som observeras resp. kvartal i de modeller som inkluderar variabler som beskriver vilket kvartal resan genomfördes.

3 Data

I denna rapport används den nationella resvaneundersökningen 2005-2006 (RES) för att estimera färdmedelsvalsmodellerna (SIKA, 2007). I denna studie används data för mätdagens huvudresa och urvalet är begränsat till arbetsresor för individer i åldrarna 20-64 år eftersom registerdatamaterialet avser befolkningen i åldern 20-64 år och deras arbetsplatser. Det huvudsakliga ärendet för resan är begränsat till resor mellan bostad och arbetsplats vilka antingen startar vid individens folkbokföringsadress eller vid hans/hennes huvudarbetsplats.

Färdmedel avser här ”huvudsakligt färdmedel” och för att reducera brus i modellen har följande restriktioner tillämpats. Endast resor som består av en delresa och som avser en enkel resa inkluderas i stickprovet för analyserna. Det huvudsakliga färdmedlet kategoriseras enligt följande i denna rapport: gång, cykel, kollektivtrafik, och ”motoriserat”. ”Kollektivtrafik” avser i huvudsak resor med tåg, tunnelbana, buss eller spårvagn. ”Motoriserat” avser i huvudsak resor med bil, passagerare i bil, motorcykel, moped, taxi, lastbil och färdtjänst. Resor med färdmedel som t.ex. båt och flyg inkluderas inte i analysen, framför allt p.g.a. att antalet observationer med sådana färdmedel är relativt få.

De variabler som används från RES skall i möjligaste mån matchas av motsvarande variabler i registerdatamaterialet. I det följande används följande information: individens ålder, kön, inkomst, utbildningsnivå (sex olika nivåer), tillgång till (privatägd) bil i hushållet, bostadskommun och bostadslän samt om individen räknar med avdrag för arbetsresor i deklarationen. Dessutom används reseavståndet mellan bostad och arbetsplats samt vilket kvartal som resan genomfördes. Den senare informationen finns inte i registerdatamaterialet men har ändå bedömts vara viktig då väderförhållanden antagligen spelar en stor roll för val att gå eller cykla till jobbet. När färdmedelsvalsmodellen sedan tillämpas i registerdatamaterialet sätts värdet för resp. kvartal till motsvarande genomsnittliga värde i RES (se tabell 1). Tillgång till (privatägd) bil i hushållet har definierats som skillnaden mellan antalet bilar i hushållet och antalet tjänstebilar. Till informationen om bostadskommun kopplas information om kommungenomsnittet för det tillgänglighetsmått som beskrevs i avsnitt 2.1.

Registerdatamaterialet består av samtliga sysselsatta individer i Sverige år 2005. Individerna är kopplade till sina resp. huvudsakliga arbetsställen. Bostäder och arbetsställen är koordinatsatta vilket gör att (det Euklidiska) avståndet mellan dessa kan beräknas. Koordinaterna är definierade för ett rutnät om 250 kvadratmeter stora rutor i tätorter och för ett rutnät om 1000 kvadratmeter stora rutor utanför tätort. Datamaterialet inkluderar ett stort antal socio-ekonomiska karaktäristika. Här används dock bara ett antal variabler som även finns i RES (se föregående stycke).

Följande restriktioner har använts på registerdatamaterialet. Bara sysselsatta individer i åldrarna 20-64 år ingår. Bara individer med fullständig information på de tidigare beskrivna variablerna ingår. Endast individer med ett reseavstånd som är kortare än 25 mil (enkel resa) ingår. Skälet för att exkludera individer med längre reseavstånd är att de

kan ha dubbelt boende med en övernattningslägenhet i närheten av arbetsplatsen. Längre reseavstånd kan även representera olika former av felklassificeringar av bostad och arbetsplats. Det är förstås väldigt få arbetsresor som genomförs med färdmedlen gång och cykel för avstånd som överstiger 15-20 km (enkel resa). Motivet för att trots detta inkludera avstånd däröver är att få stabilare skattningar av hur sannolikheterna för gång resp. cykel minskar med avståndet.

Tabell 1 presenterar beskrivande statistik för det urval av observationer från RES som används samt motsvarande information för registerdatamaterialet. Här ser vi bl.a. att 12-13 procent av individerna väljer att gå resp. cykla till jobbet och att det genomsnittliga reseavståndet är något kortare i RES (ca. 14 km) än i SCBs registerdatamaterial (ca. 16 km). Andelen män verkar vara något högre i RES än i registerdatamaterialet vilket även tycks vara fallet med den genomsnittliga åldern och den genomsnittliga inkomsten. Men andelarna för de olika utbildningsnivåerna är ungefär desamma i RES som i registerdatamaterialet. Andelen med tillgång till bil i hushållet är dock högre i RES än i registerdatamaterialet samtidigt som andelen som räknar med/har gjort bilavdrag är något högre i registerdatamaterialet. Detta kan tyda på vissa skillnader i definitionen för dessa variabler. Bilavdrag avser t.ex. en intention i RES men avser faktiskt avdrag i registerdatamaterialet. Antalet jobb på avstånd mellan 50 och 100 kilometer verkar vara något högre i RES medan övriga mått på tätheten i kommunen där individen bor inte skiljer sig så mycket åt mellan RES och registerdatamaterialet. Andelen resor till jobbet verkar vara något lägre det tredje kvartalet än övriga kvartal.

Tabell 2 beskriver variation i färdmedelsval mellan de olika länen för urvalet från RES tillsammans med länsvisa genomsnitt för några av de övriga variablerna från tabell 1. (Motivet för att inte redovisa genomsnitt för samtliga variabler från tabell 1 är utrymmesskäl.) Här ser vi att Skåne har lägst andel som går till jobbet och att motsvarande andel är högst i Västerbotten. Andelen som cyklar är högst i Kronobergs län och länet med lägst andel cyklister är Västernorrland. Kollektivtrafikens andel är högst i Stockholms län och lägst i Dalarna. Lägst andel resenärer som använder något motoriserat färdmedel (i huvudsak bil) återfinns i Stockholm och högst andel i Västernorrland. Den genomsnittliga individen har längst reseavstånd i Uppsala och kortast i Jönköping. Vi ser också från tabell 2 att andelen män i stickprovet är högst i Blekinge och lägst i Värmland och att den genomsnittliga åldern varierar mellan knapp 46 år i Värmland och drygt 41 år i Västerbotten. Högst genomsnittlig inkomst har individer i Stockholm och lägst i Jämtland. Biltillgången är högst i Norrbottens län och lägst i Stockholms län. Stickprovsstorleken är störst för Stockholm och lägst för Blekinge. Notera dock att antalet observationer från Gotland var så få att Gotlands och Kalmars län hanteras som ett län.

Tabell 1. Beskrivande statistik medelvärden, standardavvikelser.

Variabel	*RES 2005-2006*	*Registerdata*
Gång	0,124	-
Cykel	0,126	-
Kollektivt	0,161	-
Motoriserat	0,589	-
Avstånd (km)	14,242 (20,253)	16,179 (30,675)
Man	0,540	0,490
Ålder	43,377 (11,961)	42,254 (11,922)
Inkomst (tkr)	285,700 (133,879)	258,591 (177,420)
Utbildningsnivå 1	0,040	0,039
Utbildningsnivå 2	0,087	0,088
Utbildningsnivå 3	0,500	0,500
Utbildningsnivå 4	0,061	0,061
Utbildningsnivå 5	0,299	0,301
Utbildningsnivå 6	0,009	0,011
Biltillgång i hushållet	0,866	0,713
Bilavdrag	0,208	0,258
Antal jobb inom 5 km	42 704 (68659)	43 914 (68730)
Antal jobb inom 5-10 km	168 230 (237 914)	168529 (231377)
Antal jobb inom 10-50 km	103 891 (128 296)	100440 (122948)
Antal jobb inom 50-100 km	300 049 (271 813)	261383 (225603)
Kvartal 1	0,298	-
Kvartal 2	0,245	-
Kvartal 3	0,160	-
Kvartal 4	0,297	-
Antal observationer	9 801	3 244 714

Noter: Standardavvikelse inom parentes. Eftersom standardavvikelsen för en indikatorvariabel är lika med $\sqrt{p(1-p)}$ där p är genomsnittet för indikatorvariabeln anges bara genomsnitten för resp. indikatorvariabeln i tabellen av utrymmesskäl. Bilavdrag avser en intention i RES men avser vad som faktiskt gjorts i registerdatamaterialet. Antal jobb på olika avstånd från individens bostad avser genomsnitt för den kommun i vilken individen bor.

Tabell 2. Medelvärden och antal observationer inom resp. län RES

län	*gång*	*cykel*	*koll.*	*motor.*	*avstånd*	*man*	*ålder*	*inkomst*	*biltillgång*	*#obs*
1	0,115	0,063	**0,373**	**0,449**	14,666	0,518	42,600	**329**	**0,759**	**2450**
3	0,103	0,130	0,229	0,539	**24,193**	0,521	43,700	279	0,872	516
4	0,168	0,128	0,069	0,635	16,885	0,570	44,467	272	0,862	537
5	0,142	0,154	0,102	0,602	12,330	0,590	43,322	266	0,904	332
6	0,163	0,110	0,090	0,637	**9,531**	0,507	44,341	260	0,949	355
7	0,096	**0,218**	0,064	0,622	14,438	0,596	42,590	283	0,904	156
9	0,105	0,177	0,068	0,650	14,958	0,600	44,664	269	0,927	220
10	0,156	0,119	0,055	0,670	15,414	**0,651**	41,752	276	0,908	**109**
12	**0,068**	0,175	0,125	0,632	16,173	0,520	42,748	279	0,878	790
13	0,118	0,158	0,092	0,632	15,629	0,518	42,333	273	0,917	228
14	0,111	0,137	0,108	0,644	14,666	0,578	43,086	275	0,865	1218
17	0,117	0,141	0,047	0,695	12,343	**0,477**	**45,852**	250	0,922	256
18	0,169	0,201	0,053	0,577	9,779	0,485	43,434	267	0,909	551
19	0,103	0,152	0,071	0,674	13,156	0,595	44,047	291	0,919	533
20	0,096	0,156	**0,041**	0,706	12,197	0,619	45,601	265	0,931	218
21	0,156	0,143	0,046	0,654	10,895	0,544	44,172	259	0,936	518
22	0,175	**0,053**	0,063	**0,709**	11,143	0,534	45,728	277	0,981	206
23	0,165	0,120	0,068	0,647	12,601	0,541	42,872	**239**	0,947	133
24	**0,176**	0,188	0,061	0,576	10,610	0,531	**41,159**	268	0,882	245
25	0,139	0,100	0,061	0,700	12,028	0,478	43,987	267	**0,970**	230

Noter: Värden markerade med fet stil indikerar det lägsta värdet och högsta värdet i resp. kolumn. Län 1=Stockholm, Län 3=Uppsala, Län 4=Södermanland, Län 5=Östergötland, Län 6= Jönköping, Län 7=Kronoberg, Län 9 = Kalmar och Gotland (här sammanslagna), Län 10 = Blekinge, Län 12=Skåne, Län 13= Halland, Län 14 = Västra Götaland, Län 17= Värmland, Län 18=Örebro, Län 19=Västmanland, Län 20=Dalarna, Län 21=Gävleborg, Län 22=Västernorrland, Län 23=Jämtland, Län 24=Västerbotten, Län 25=Norrbotten. Avstånd i kilometer och inkomst i tusentals kronor.

Tabell 3 beskriver hur färdmedelsandelarna ser ut för olika reseavstånd. Här ser vi att reseavståndet har en stor betydelse för valet att gå eller cykla till jobbet vilket tyder på att avstånd är en viktig faktor för att prediktera valet att gå och cykla. Det finns i princip ingen som går om avståndet överstiger 10 km och väldigt få väljer att cykla då avståndet överstiger 15 kilometer. För avstånd upp till 1 kilometer väljer de flesta att gå medan andelen som går faller starkt ju högre avståndet är. För alla avståndsintervall som överstiger 1 kilometer dominerar andelen som väljer att åka bil. Högst andel för cykel återfinns i intervallet 1-2 kilometer men även på dessa relativt korta avstånd dominerar alltså andelen som väljer bil framför cykel.

Tabell 3. Färdmedelsandelar för olika intervall på reseavståndet

Avståndsintervall	*Gång*	*Cykel*	*Kollektivt*	*Motoriserat*	*# obs.*
0-1 km	0,624	0,250	0,004	0,122	1186
1-2 km	0,230	0,342	0,021	0,406	1043
2-5 km	0,098	0,211	0,103	0,588	2028
5-10 km	0,016	0,079	0,227	0,678	1640
10-15 km	0,005	0,024	0,244	0,726	961
15-20 km	0,001	0,007	0,212	0,780	751
20-25 km	0,000	0,000	0,244	0,755	544
25-30 km	0,005	0,000	0,240	0,756	438
Över 30 km	0,000	0,001	0,281	0,718	1210

4 Resultat

Tabell 4 sammanfattar information från den procedur som användes för att välja ut variabler i modellen. Den visar vilka variabler som ingått i var och en av de 18 olika versioner av den multinomiala logit modellen vars prediktionsförmåga har undersökts. Här ser vi t.ex. att avstånd ingår i alla 18 modeller för gång och cykel. Men avstånd ingår i modellen för kollektivt färdmedel först fr.o.m. modell 4. Vi ser också att den sista variabel som inkluderas för gång är kvartal 4 och att den sista variabel som inkluderas för cykel och kollektivt färdmedel är län 3 (Uppsala) resp. län 14 (Västra Götaland). Tabellen visar också att 13 variabler ingår för att prediktera gång och att motsvarande antal är 18 för cykel och 14 för kollektivt färdmedel.

I tabell 5 presenteras de estimerade prediktionsfelen (se avsnitt 2.2) för resp. modell avseende färdmedlen gång och cykel. Överlag är prediktionsfelen lägre för gång än för cykel vilket indikerar att det är svårare att prediktera att en individ tar cykel till jobbet än att en individ går till jobbet. Dessutom redovisas de ”log-likelihood” baserade måtten AIC och BIC för resp. modell. Vi ser här att det lägsta prediktionsfelet för gång uppnås med modellerna 14 och 16 (0,087). Det lägsta prediktionsfelet för cykel uppnås med modellerna 13-16 och modell 18 (0,129). Minimum för AIC och BIC uppnås för modell 18. Eftersom prediktionsfelet för gång i modell 18 (0,088) är nära minimum för gång väljs modell 18 för att generera prediktioner för gång och cykel i registerdatamaterialet. Dessutom tyder värdena på AIC och BIC att denna modell är ”bäst” av de 18 olika modellerna.

För att ytterligare undersöka prediktionskvaliteten i modell 18 presenterar Tabell 6 resultat för prediktionsfelet i olika län för färdmedlen gång och cykel. Dessutom redovisas det lägsta (minimum) prediktionsfelet bland de 18 modellerna i resp. län och information om för vilken eller vilka modeller minimum uppnåddes. Resultaten visar att den ”bästa” prediktionsmodellen för resp. färdmedel varierar mellan olika län. Men vi ser också från tabell 6 att skillnaderna i estimerade prediktionsfel i de flesta jämförelser är liten. För gång överstiger prediktionsfelet i modell 18 minimum med över en procentenhet i sex stycken län. Motsvarande siffra för cykel är nio stycken län även om skillnaderna i de flesta fall fortfarande inte är stor. Ett undantag kan dock noteras, i län 7 (Kronoberg) är avvikelsen ca. 5 procentenheter för både gång och cykel vilket tyder på att det finns en modell som skulle ge bättre prediktioner i detta län. Vi ser också från tabellen att minimum uppnås med modell 6 för båda färdmedlen i Kronobergs län.

Tabell 4. Modeller där resp. variabel inkluderats

Variabel	*Gång*	*Cykel*	*Kollektiv färdmedel*
Avstånd	1-18	1-18	4-18
Man	4-18	-	5-18
Inkomst	-	12-18	9-18
Utbildningsnivå 4	-	-	10-18
Utbildningsnivå 5	6-18	3-18	8-18
Utbildningsnivå 6	-	15-18	13-18
Tillgång till bil	2-18	2-18	1-18
Bilavdrag	5-18	4-18	3-18
Antal jobb inom 5 km	3-18	9-18	6-18
Antal jobb 5-10 km	-	14-18	2-18
Län 3	-	18	7-18
Län 4	7-18	-	12-18
Län 6	-	16-18	-
Län 7	-	17-18	-
Län 9	8-18	-	-
Län 12	9-18	10-18	-
Län 14	-	-	14-18
Län 17	10-18	-	-
Län 18	-	11-18	-
Län 22	-	5-18	-
Län 25	-	13-18	-
Kvartal 2	12-18	7-18	11-18
Kvartal 3	11-18	6-18	-
Kvartal 4	13-18	8-18	-

Noter: 1-18 betyder att variabeln var med i samtliga 18 modeller, 2-18 betyder att variabeln inkluderats i alla modeller utom den första osv. Även om en variabel skulle kunna uteslutas från modellen då andra variabler inkluderats genom att dess p-värde skulle överstiga 10 procent så inträffade detta aldrig. Därför har ingen variabel uteslutits från modellen efter det att den en gång har inkluderats. De variabler som fanns med i tabell 1 men som inte finns med i denna tabell valdes aldrig ut för att inkluderas i modellen av den procedur som beskrivs i texten. Ett ”-” betyder att parametern för den variabeln har satts till noll vilket alltså innebär att den variabeln inte ingår i modellen för det färdmedlet. Motivet är att den variabeln inte är signifikant förklarande för val av det färdmedlet. För länskoder se t.ex. noterna till tabell 2.

Tabell 5. Estimerade prediktionsfel (0,632+ estimatorn) för gång och cykel, samt AIC och BIC för resp. multinomial logit modell

Modell	*Prediktionsfel gång*	*Prediktionsfel cykel*	*AIC*	*BIC*
1	0,097	0,140	16 172,00	16 183,95
2	0,099	0,134	15 165,84	15 183,76
3	0,095	0,137	14 803,05	14 826,95
4	0,094	0,137	14 321,12	14 350,99
5	0,094	0,138	14 147,40	14 183,24
6	0,091	0,137	13 947,65	13 989,47
7	0,091	0,134	13 844,41	13 892,20
8	0,090	0,136	13 772,80	13 826,56
9	0,089	0,135	13 737,69	13 799,42
10	0,089	0,134	13 721,98	13 789,68
11	0,088	0,132	13 699,66	13 773,34
12	0,088	0,130	13 683,95	13 763,60
13	0,088	**0,129**	13 673,15	13 758,77
14	**0,087**	**0,129**	13 665,53	13 755,14
15	0,088	**0,129**	13 665,76	13 757,36
16	**0,087**	**0,129**	13 657,19	13 750,78
17	0,088	0,130	13 653,63	13 749,21
18	0,088	**0,129**	**13 649,97**	**13 747,54**

Not: Prediktionsfelen har estimerats med Efron & Tibshiranis metod som beskrivs i texten. Siffror markerade med fet stil anger det lägsta värdet i kolumnen.

Tabell 6. Estimerade prediktionsfel (0,632+ estimatorn) för modell 18 efter län samt det lägsta prediktionsfelet (minimum) bland de 18 modellerna – färdmedlen gång och cykel

	Gång			*Cykel*		
Län	*Modell 18*	*Minimum*	*Modeller m. minimum*	*Modell 18*	*Minimum*	*Modeller m. minimum*
1	0,076	0,074	6-8	0,070	0,067	2
3	0,072	0,072	18	0,124	0,121	13-17
4	0,105	0,101	3	0,133	0,133	18
5	0,093	0,086	8	**0,140**	**0,125**	1
6	**0,121**	**0,094**	5	0,117	0,116	6
7	**0,111**	**0,060**	6	**0,224**	**0,176**	3, 6 och 7
9	0,053	0,051	13-17	**0,186**	**0,164**	6
10	**0,059**	**0,037**	1-3	**0,148**	**0,119**	1
12	0,046	0,044	9 och 10	0,145	0,142	13
13	0,064	0,064	13-18	0,112	0,111	8 och 16
14	0,082	0,080	16 och 17	0,146	0,143	3
17	0,085	0,080	10	**0,142**	**0,130**	9
18	0,102	0,097	12	0,202	0,195	12
19	0,106	0,100	1 och 3	**0,186**	**0,162**	3
20	0,088	0,078	3	**0,173**	**0,152**	2
21	0,118	0,114	7	0,151	0,149	12
22	**0,143**	**0,117**	3	0,053	0,053	5-18
23	0,061	0,060	13-15	**0,102**	**0,088**	7
24	**0,162**	**0,148**	6	**0,201**	**0,186**	2
25	**0,121**	**0,107**	12	0,115	0,110	2

Noter: Minimum indikerar det lägsta värdet på prediktionsfelet för de 18 testade modellerna och därmed den ”bästa” modellen för det länet och det färdmedlet. Prediktionsfel markerade med fet stil indikerar att skillnaden mellan prediktionsfelet i modell 18 och det lägsta prediktionsfelet för någon av testade 18 modellerna överstiger en procentenhet. Län 1=Stockholm, Län 3=Uppsala, Län 4=Södermanland, Län 5=Östergötland, Län 6= Jönköping, Län 7=Kronoberg, Län 9 = Kalmar och Gotland (här sammanslagna), Län 10 = Blekinge, Län 12=Skåne, Län 13= Halland, Län 14 = Västra Götaland, Län 17= Värmland, Län 18=Örebro, Län 19=Västmanland, Län 20=Dalarna, Län 21=Gävleborg, Län 22=Västernorrland, Län 23=Jämtland, Län 24=Västerbotten, Län 25=Norrbotten.

För att få en bild av vilken vikt den valda modellen lägger på de olika variablerna i modellen presenteras i tabellerna 7a och 7b de skattade parametrarna i modellen för alternativen gång och cykel relativt alternativet ”motoriserat”. Notera att tabellerna 7a och 7b är estimerade i samma multinomiala logitmodell även om användandet av två tabeller för redovisningen skulle kunna ge intryck av att så inte är fallet. Negativa värden i kolumnen ”estimat” innebär att högre värden på variabeln sänker sannolikheten att färdmedlet valdes istället för motoriserat färdmedel (betingat på värdena för de andra variablerna i modellen).

Alltså, jämfört med att välja ett motoriserat färdmedel för resan till jobbet ser vi följande ang. gång och cykel i tabellerna 7a och 7b. För det första, ju längre en individ har till jobbet ju lägre är sannolikheten att han/hon går/cyklar. För det andra, sannolikheten att män väljer att gå är lägre än för kvinnor och sannolikheten att cykla minskar med högre inkomst. För det tredje, sannolikheten att gå eller cykla till jobbet är lägre för individer med tillgång till bil i hushållet eller som planerar att göra avdrag för bilresor till arbetet. För det fjärde, sannolikheten att individer med utbildningsnivå 5 går eller cyklar till jobbet är högre än för individer med utbildningsnivå 1. Dessutom tenderar individer med utbildningsnivå 6 att cykla mer än individer med utbildningsnivå 1. Denna parameter är dock inte signifikant skild ifrån noll på tio procentsnivån vilket kan verka märkligt då den procedur som använts för att välja ut variabler till modellen bara skulle behålla variabler som var signifikanta på tio procentsnivån. Detta beror på att proceduren är baserad på motsvarande binära logitmodell och parametrarna samt standardfelen i denna modell är lite annorlunda än de som ges för färdmedlet i den motsvarande multinomiala logitmodellen. För det femte, ju fler jobb som finns nära den genomsnittliga individen i kommunen (ett mått på tätheten i kommunen) desto högre är sannolikheten att individen går eller cyklar till jobbet. Ju högre tätheten på avståndet 5-10 km ju lägre är dock sannolikheten att individen cyklar till jobbet. Vi ser också att sannolikheten att en individ går istället för att ta ett motoriserat färdmedel till jobbet är högre i Södermanland än i Stockholm medan motsvarande sannolikhet är lägre i Kalmar och Gotland, Skåne och Värmland. Sannolikheten att cykla till jobbet istället för att ta ett motoriserat färdmedel är högre i Uppsala, Kronoberg, Skåne och Örebro än i Stockholm. Den är dock lägre i Jönköping, Västernorrland och Norrbotten än i Stockholm. Något oväntat är sannolikheten att gå till jobbet jämfört med att ta ett motoriserat färdmedel högre i kvartal 1 än under övriga kvartal medan motsvarande sannolikhet för att cykla är lägre i kvartal 1 än under övriga kvartal.

Tabell 7a. Modell 18 gång (Referensfärdmedel: Motoriserat)

Parameter	*Estimat*	*Standardfel*
Intercept	2,5042***	0,1612
Avstånd	-2,4403***	0,0637
Man	-0,6247***	0,0864
Inkomst	0	.
Tillgång till bil	-1,6342***	0,1281
Bilavdrag	-1,4808***	0,385
Utbildningsnivå 4	0	.
Utbildningsnivå 5	0,6988***	0,1003
Utbildningsnivå 6	0	.
Antal jobb 0-5 km (100 000)	0,6926***	0,0703
Antal jobb 5-10 km (100 000)	0	.
Län 3	0	.
Län 4	0,1065**	0,0431
Län 6	0	.
Län 7	0	.
Län 9	-0,0914***	0,0331
Län 12	-0,0485***	0,0166
Län 14	0	.
Län 17	-0,0321**	0,016
Län 18	0	.
Län 22	0	.
Län 25	0	.
Kvartal 2	-0,3693***	0,124
Kvartal 3	-0,5974***	0,1483
Kvartal 4	-0,3382***	0,1153

Noter: Tabellen redovisar parametrarna för gång i den multinomiala logitmodell som beskrivs i texten. Siffran 0 innebär att parametern har varit begränsad till det värdet men att motsvarande parameter för annat färdmedel inte har varit det. Variabler som beskrivs i texten men som inte ingår i tabellen har inte valts ut för något färdmedel av den procedur som beskrivs i texten. Referenskategori för kön är kvinnor, för län är det Stockholm och för kvartal det första kvartalet. Avstånd avser den naturliga logaritmen av avståndet. Län 3=Uppsala, Län 4=Södermanland, Län 6= Jönköping, Län 7=Kronoberg, Län 9 = Kalmar och Gotland (här sammanslagna), Län 12=Skåne, Län 14 = Västra Götaland, Län 17= Värmland, Län 18=Örebro, Län 22=Västernorrland, Län 25=Norrbotten. *** betyder att parametern är signifikant skild ifrån noll på en procentsnivån och ** att den är signifikant skild från noll på fem procentsnivån.

Tabell 7b. Modell 18 cykel (Referensfärdmedel: Motoriserat)

Parameter	*Estimat*	*Standardfel*
Intercept	1,0858***	0,1608
Avstånd	-1,3255***	0,0464
Man	0	.
Inkomst	-1,0162***	0,309
Tillgång till bil	-1,3179***	0,1139
Bilavdrag	-1,4950***	0,2266
Utbildningsnivå 4	0	.
Utbildningsnivå 5	0,8852***	0,0829
Utbildningsnivå 6	0,4589	0,3733
Antal jobb 0-5 km (100 000)	0,2604***	0,0847
Antal jobb 5-10 km (100 000)	-0,0512**	0,0249
Län 3	0,1291**	0,0532
Län 4	0	.
Län 6	-0,0865***	0,0309
Län 7	0,0830**	0,0324
Län 9	0	.
Län 12	0,0342***	0,0106
Län 14	0	.
Län 17	0	.
Län 18	0,0227***	0,00714
Län 22	-0,0594***	0,0148
Län 25	-0,0236**	0,00968
Kvartal 2	0,8573***	0,1067
Kvartal 3	1,0388***	0,1162
Kvartal 4	0,5681***	0,1046

Noter: Tabellen redovisar parametrarna för gång i den multinomiala logitmodell som beskrivs i texten. Siffran 0 innebär att parametern har varit begränsad till det värdet men att motsvarande parameter för annat färdmedel inte har varit det. Variabler som beskrivs i texten men som inte ingår i tabellen har inte valts ut för något färdmedel av den procedur som beskrivs i texten. Referenskategori för kön är kvinnor, för län är det Stockholm och för kvartal det första kvartalet. Avstånd avser den naturliga logaritmen av avståndet. Län 3=Uppsala, Län 4=Södermanland, Län 6= Jönköping, Län 7=Kronoberg, Län 9 = Kalmar och Gotland (här sammanslagna), Län 12=Skåne, Län 14 = Västra Götaland, Län 17= Värmland, Län 18=Örebro, Län 22=Västernorrland, Län 25=Norrbotten. *** betyder att parametern är signifikant skild ifrån noll på en procentsnivån och ** att den är signifikant skild från noll på fem procentsnivån.

I tabellerna 8a och 8b presenteras prediktioner för färdmedelsvalen gång och cykel i resp. län för modell 18 och hur dessa överensstämmer med observerade val i RES. Här ser vi att modellen fungerar relativt väl för att korrekt prediktera individer som inte går resp. inte cyklar. I Skånes län predikterar modellen felaktigt att en individ går då han eller hon inte gör det i lite drygt 1 procent av fallen. Den högsta felprocenten för individer som inte går hittar vi för Västernorrlands län där knappt 13 procent predikteras gå när de i själva verket inte gör det. För cykel ser vi att modellen perfekt predikterar de som inte cyklar i Västernorrlands län. Den högsta felprocenten för de som inte cyklar hittar vi Kronobergs län där knappt 14 procent av de som inte cyklar predikteras välja cykel som färdmedel. Däremot fungerar modellen sämre för att prediktera individer som faktiskt går resp. cyklar. För Kronobergs län ser vi t.ex. att modellen ger en felaktig prediktion för drygt 73 procent av de som går. I Västernorrlands län lyckas dock modellen korrekt prediktera drygt 80 procent av de som går. Modellen verkar fungera ännu sämre för att prediktera de som faktiskt cyklar. Här ser vi att den felaktigt predikterar samtliga cyklister som icke-cyklister i Jönköpings, Västernorrlands och Norrbottens län. Den fungerar bäst i Skåne där ca. 46 procent av cyklisterna felaktigt predikteras som icke-cyklister.

Svårigheten att korrekt prediktera cyklister kan bero på att andelen cyklister aldrig är högst på något avståndsintervall (se tabell 3). Dessutom kanske cykel är ett färdmedel som inte används varje dag utan då t.ex. vädret eller omständigheterna i övrigt tillåter individen att välja detta alternativ. Detta kan även i viss mån även gälla alternativet att gå till jobbet. Då informationen i RES avser färdmedelsvalet en arbetsdag någon gång under året är det inte säkert optimalt att försöka använda predikterade val för cykel som om en individ skulle välja det alternativet varje dag under året. Istället kanske det är mer relevant att basera prognoser för antalet personer som cyklar resp. går i en viss relation i registerdatamaterialet på anpassade (predikterade) sannolikheter för resp. färdmedelsval. Därför redovisas i det följande den predikterade sannolikheten att välja gång resp. cykel för individerna i registerdatamaterialet istället för dessa individers predikterade val.

Tabell 8a. Frekvenser för korrekta och felaktiga prediktioner efter observerat val – *gång* i modell 18

	Observerat: Går inte		*Observerat: Går*	
län	*Korrekt*	*Fel*	*Fel*	*Korrekt*
Alla	0,952	0,048	0,361	0,639
Stockholm	0,956	0,044	0,326	0,674
Uppsala	0,978	0,022	0,491	0,509
Södermanland	0,924	0,076	0,256	0,744
Östergötland	0,961	0,039	0,426	0,574
Jönköping	0,912	0,088	0,293	0,707
Kronoberg	0,95	0,05	0,733	0,267
Kalmar & Gotland	0,975	0,025	0,217	0,783
Blekinge	0,978	0,022	0,235	0,765
Skåne	0,989	0,011	0,481	0,519
Halland	0,97	0,03	0,296	0,704
Västra Götaland	0,953	0,047	0,356	0,644
Värmland	0,965	0,035	0,433	0,567
Örebro	0,945	0,055	0,344	0,656
Västmanland	0,935	0,065	0,418	0,582
Dalarna	0,954	0,046	0,476	0,524
Gävleborg	0,954	0,046	0,494	0,506
Västernorrland	0,871	0,129	0,194	0,806
Jämtland	0,982	0,018	0,273	0,727
Västerbotten	0,876	0,124	0,349	0,651
Norrbotten	0,924	0,076	0,375	0,625

Tabell 8b. Frekvenser för korrekta och felaktiga prediktioner efter observerat val – *cykel* i modell 18

	Observerat: Cyklar inte		*Observerat: Cyklar*	
län	*Korrekt*	*Fel*	*Fel*	*Korrekt*
Alla	0,962	0,038	0,744	0,256
Stockholm	0,987	0,013	0,916	0,084
Uppsala	0,942	0,058	0,507	0,493
Södermanland	0,974	0,026	0,855	0,145
Östergötland	0,968	0,032	0,745	0,255
Jönköping	0,994	0,006	1,000	0,000
Kronoberg	0,861	0,139	0,559	0,441
Kalmar & Gotland	0,923	0,077	0,590	0,410
Blekinge	0,948	0,052	0,846	0,154
Skåne	0,931	0,069	0,464	0,536
Halland	0,979	0,021	0,583	0,417
Västra Götaland	0,958	0,042	0,814	0,186
Värmland	0,941	0,059	0,667	0,333
Örebro	0,916	0,084	0,676	0,324
Västmanland	0,945	0,055	0,877	0,123
Dalarna	0,957	0,043	0,824	0,176
Gävleborg	0,957	0,043	0,757	0,243
Västernorrland	1,000	0,000	1,000	0,000
Jämtland	0,974	0,026	0,625	0,375
Västerbotten	0,950	0,050	0,848	0,152
Norrbotten	0,990	0,010	1,000	0,000

I tabell 9 redovisas genomsnitt av predikterade sannolikheter för individerna i registerdatamaterialet dels för hela landet, dels efter resp. län. Dessa kan alltså betraktas som predikterade färdmedelsandelar för resp. län. Tabellen redovisar också antalet individer i registerdatamaterialet både för hela landet och för resp. län. För var och en av alla 3 244 714 individer i registerdatamaterialet estimeras alltså 100 unika sannolikheter för att välja gång resp. cykel. Alla prediktioner baseras på den tidigare utvalda modellen men bootstrap-metoden innebär att vi kan få 100 unika estimat av denna modell. Detta innebär i sin tur att vi kan estimera standardavvikelsen i den predikterade sannolikheten för varje individ. Detta ger oss ett mått på osäkerheten i prediktionen som beror på stickprovsvariationen i RES. Detta redovisas i tabellen som genomsnittliga individuella standardavvikelser för den predikterade sannolikheten. De länsvisa genomsnittliga predikterade sannolikheterna kan jämföras med motsvarande färdmedelsandel i tabell 2. Den informationen replikeras i tabell 9 för att underlätta jämförelser mellan registerdatamaterialet och RES i detta avseende. Vi ser att högsta och lägsta värden för resp. ”färdmedelsandel” överensstämmer relativt väl mellan registerdatamaterialet och RES. Andelen för gång är lägst i Skåne både enligt prediktionerna i registerdatamaterialet och i RES. Den lägsta (högsta) andelen cyklister återfinns i Västernorrlands län (Kronobergs län) både enligt de predikterade sannolikheterna i registerdatamaterialet och enligt observerade färdmedelsval i RES. Däremot finns den högsta färdmedelsandelen för gång enligt bedömningen i registerdatamaterialet i Jämtland när den enligt RES återfinns i Västerbotten. Överlag verkar storleksordningen på färdmedelsvalandelarna i registerdatamaterialet överensstämma relativt väl med motsvarande andel i RES. Detta gäller i synnerhet om man även beaktar den genomsnittliga individuella osäkerheten (standardavvikelsen) i bedömningen.

Samtidigt ska man komma ihåg att RES är ett stickprov vilket innebär att det finns en statistisk osäkerhet i färdmedelsvalsandelen även där. Därför presenteras i tabell 10 95-procentiga konfidensintervall för färdmedelsandelen i RES för gång resp. cykel, för hela landet och för resp. län. Dessa intervall jämförs med motsvarande punktskattning i registerdatamaterialet och i tabellen indikeras om punktskattningen ligger i eller utanför motsvarande konfidensintervall. Här ser vi, för hela landet, att andelen som går till jobbet enligt prediktionerna i registerdatamaterialet är något högre än den övre gränsen för det 95-procentiga konfidensintervallet i RES (15 procent jämfört med 13,1 procent). För cykel är prediktionen för hela landet i registerdatamaterialet marginellt högre än den övre gränsen för det 95-procentiga konfidensintervallet i RES (13,4 procent jämfört med 13,3 procent). Vad gäller motsvarande jämförelser i resp. län ser vi att andelen för gång i registerdatamaterialet hamnar utanför det 95-procentiga konfidensintervallet i RES i sex stycken län. I samtliga dessa fall ligger färdmedelsandelen för gång i registerdatamaterialet över den övre gränsen i intervallet. För cykel är dock den predikterade färdmedelsandelen endast utanför det 95-procentiga konfidensintervallet i ett län och då är den över den övre gränsen i intervallet.

Tabell 9. Individuella predikterade sannolikheter för att välja gång resp. cykel och tillhörande genomsnittliga *individuella* standardavvikelser i registerdatamaterialet, genomsnitt per län samt motsvarande färdmedelsandelar i RES.

län	*# Obs*	*Gång genomsnitt*	*Gång Std.av.*	*Gång RES*	*Cykel Genomsnitt*	*Cykel Std.av.*	*Cykel RES*
Alla	3 244 714	0,150	0,011	0,124	0,134	0,014	0,126
1	764 529	0,150	0,010	0,115	0,080	0,010	0,063
3	117 938	0,112	0,010	0,103	0,152	0,019	0,130
4	92 359	0,189	0,017	0,168	0,115	0,012	0,128
5	148 289	0,160	0,009	0,142	0,149	0,011	0,154
6	120 681	0,198	0,011	0,163	0,100	0,017	0,110
7	65 618	0,149	0,015	0,096	**0,228**	0,034	**0,218**
9	96 479	0,122	0,018	0,105	0,183	0,017	0,177
10	50 369	0,150	0,008	0,156	0,138	0,010	0,119
12	394 017	**0,100**	0,012	**0,068**	0,198	0,018	0,175
13	98 224	0,133	0,007	0,118	0,122	0,009	0,158
14	526 596	0,154	0,009	0,111	0,135	0,011	0,137
17	91 501	0,125	0,019	0,117	0,165	0,016	0,141
18	99 742	0,147	0,011	0,169	0,197	0,020	0,201
19	93 702	0,164	0,009	0,103	0,151	0,011	0,152
20	96 940	0,157	0,008	0,096	0,138	0,010	0,156
21	97 173	0,166	0,009	0,156	0,146	0,010	0,143
22	84 441	**0,206**	0,012	0,175	**0,052**	0,016	**0,053**
23	42 478	0,203	0,010	0,165	0,156	0,011	0,120
24	87 562	0,185	0,010	**0,176**	0,165	0,012	0,188
25	76 076	0,179	0,023	0,139	0,104	0,022	0,100

Not: Genomsnittliga sannolikheter och andelar som är markerade med fetstil indikerar det högsta resp. lägsta värde i resp. kolumn. För länskoder se t.ex. noter till tabell 2. #Obs avser registerdatamaterialet.

Tabell 10. Konfidensintervall (95%) för färdmedelsandelar i RES efter län och alternativen gång och cykel jämfört med motsvarande prediktion i registerdatamaterialet

län	*Gång låg*	*Gång hög*	*Register*	*i eller utanför (u)*	*Cykel låg*	*Cykel hög*	*Register*	*i eller utanför (u)*
Alla	0,117	0,131	0,150	u	0,119	0,133	0,134	u
1	0,102	0,128	0,150	u	0,053	0,073	0,080	u
3	0,077	0,129	0,112	i	0,101	0,159	0,152	i
4	0,136	0,200	0,189	i	0,100	0,156	0,115	i
5	0,104	0,180	0,160	i	0,115	0,193	0,149	i
6	0,125	0,201	0,198	i	0,077	0,143	0,100	i
7	0,050	0,142	0,149	u	0,153	0,283	0,228	i
9	0,064	0,146	0,122	i	0,127	0,227	0,183	i
10	0,088	0,224	0,15	i	0,058	0,180	0,138	i
12	0,050	0,086	0,100	u	0,149	0,201	0,198	i
13	0,076	0,16	0,133	i	0,111	0,205	0,122	i
14	0,093	0,129	0,154	u	0,118	0,156	0,135	i
17	0,078	0,156	0,125	i	0,098	0,184	0,165	i
18	0,138	0,200	0,147	i	0,168	0,234	0,197	i
19	0,077	0,129	0,164	u	0,122	0,182	0,151	i
20	0,057	0,135	0,157	u	0,108	0,204	0,138	i
21	0,125	0,187	0,166	i	0,113	0,173	0,146	i
22	0,123	0,227	0,206	i	0,022	0,084	0,052	i
23	0,102	0,228	0,203	i	0,065	0,175	0,156	i
24	0,128	0,224	0,185	i	0,139	0,237	0,165	i
25	0,094	0,184	0,179	i	0,061	0,139	0,104	i

Noter: ”Gång låg” (”cykel låg”) och ”gång hög” (”cykel hög”) är den undre resp. övre gränsen i intervallet för gång (cykel).”Register” avser punkskattningen i registerdata och ”i eller utanför” anger om denna ligger i eller utanför (u) det 95 procentiga konfidensintervallet. För länskoder se noter till tabell 2.

För att ytterligare undersöka hur väl prediktionerna i registerdatamaterialet överensstämmer med observerade färdmedelsvalsandelar i RES, redovisas i tabell 11 gång och cykels färdmedelsandelar för de tre största kommunerna Stockholm, Göteborg och Malmö. Huvudskälet för att inte ta fler kommuner är att urvalen i RES blir väldigt små för andra enskilda kommuner vilket ger en stor osäkerhet i bedömningen. Ett viktigt motiv för att presentera jämförelsen i tabell 11 är att modellen som estimerats på RES inkluderar indikatorvariabler för vissa län vilket innebär att det inte är så konstigt att man får en hyfsad överensstämmelse mellan färdmedelsandelar på länsnivå i registerdatamaterialet och i RES. Jämförelsen i tabell 11 följer alltså samma upplägg som i tabell 9 men avser nu tre kommuner. Predikterade färdmedelsandelar för gång i registerdatamaterialet verkar något högre än motsvarande färdmedelsandel i RES. Detta gäller alla tre kommunerna. Predikterade färdmedelsandelar för cykel i registerdatamaterialet verkar dock överensstämma bättre med motsvarande färdmedelsandel i RES. Undantaget är Stockholm där den är något högre än färdmedelsandelen i RES.

Tabell 11. Individuella predikterade sannolikheter för att välja gång resp. cykel och tillhörande genomsnittliga individuella standardavvikelser (inom parentes) i registerdatamaterialet, genomsnitt per kommun samt motsvarande färdmedelsandelar i RES, *Stockholm, Göteborg och Malmö*

Kommun	*# Obs*	*Gång*		*Gång*	*Cykel*		*Cykel*
		Genom-snitt	*Stand. Avvik.*	*RES*	*Genom-snitt*	*Stand. Avvik.*	*RES*
Stockholm	319 908	0,213	0,015	0,161	0,095	0,012	0,073
Malmö	86 783	0,134	0,017	0,065	0,274	0,025	0,279
Göteborg	173 542	0,173	0,011	0,126	0,158	0,014	0,165

Not: Antalet observationer från RES som ligger till grund för den skattade färdmedelsandelen i resp. kommun är: 1004 i Stockholm, 154 i Malmö och 382 i Göteborg. RES-baserade konfidensintervallen (95%) för gång är: Stockholm (0,139-0,184), Malmö (0,026-0,104) och Göteborg (0,092-0,159). RES-baserade konfidensintervallen (95%) för cykel är: Stockholm (0,057-0,089), Malmö (0,208-0,351) och Göteborg (0,127-0,202).

Men huvudsyftet för att generera prognoser med den ansats som illustreras i denna uppsats är inte att presentera färdmedelsandelar på läns- och kommunnivå. De jämförelser som hittills har gjorts (i tabellerna 9-11) mellan de genomsnittliga predikterade sannolikheterna för att välja gång resp. cykel har bara syftat till att ge en bild av hur modellen fungerar relativt de i RES observerade färdmedelsandelarna. Ytterligare nedbrytning av RES för detta ändamål bedöms dock inte vara särskilt informativt. För att illustrera hur prediktioner på det geografiskt högupplösta registerdatamaterialet kan användas för att bedöma antalet som går resp. cyklar i ett JA för GC-kalk redovisar tabell 12 istället genomsnittliga predikterade sannolikheter för gång resp. cykel för tre rutor i Stockholms kommun. Här ser vi t.ex. att det bor 333 sysselsatta individer i den ruta vars nedre vänstra hörn ligger i nord-sydlig koordinat 657400 och ost-västlig koordinat 1633000. För dessa individer är den genomsnittliga

färdmedelsandelen för gång 0,082 och 0,057 för cykel. På motsvarande sätt ser vi att färdmedelsandelarna för gång i de andra rutorna i tabellen är 0,096 resp. 0,085 samt 0,078 resp. 0,063 för cykel. Det skulle förstås även gå att ta fram en komplett OD-matris för alla kombinationer av rutidentitet för bostad och rutidentitet för arbetsplats (OD-relationer) eller en karta för kommunen som illustrerar antalet personer som bedöms gå resp. cykla i resp. relation.

Tabell 12. Individuella predikterade sannolikheter för att välja gång resp. cykel och tillhörande genomsnittliga *individuella* standardavvikelser (inom parentes) i registerdatamaterialet, genomsnitt per ruta, *några utvalda rutor i Stockholm*

X-Koordinat	*Y-koordinat*	*# Obs*	*Gång*		*Cykel*	
			Genom-snitt	*Stand. Avvik.*	*Genom-snitt*	*Stand. Avvik.*
6574000	1633000	333	0,082	0,006	0,057	0,007
6571000	1633000	35	0,096	0,009	0,078	0,009
6575250	1632500	281	0,085	0,008	0,063	0,008

Den osäkerhet (standardavvikelsen) som beror på stickprovsvariation i RES kan användas för att anpassa prediktionen för JA som görs i registerdatamaterialet till specifika förutsättningar för investeringar som analyseras i GC-kalk. Detta beror på att man kan representera prognoserna för de enskilda individerna som intervall istället för som ett specifikt värde som i tabell 12. Om man t.ex. vet att cykelalternativet i en specifik OD-relation i huvudsak går i blandtrafik så kan man mot bakgrund av resultaten i t.ex. Björklund och Isacsson (2013) välja lägre värden för individernas predikterade sannolikhet att ta cykel än om det finns en gång- och cykelbana i relationen. För att illustrera detta beräknas först den predikterade sannolikheten för gång +/- standardavvikelsen multiplicerat med 2 för varje individ och motsvarande intervall för individens sannolikhet att välja cykel. Därefter beräknas genomsnitten för dessa höga och låga värden för samtliga individer i resp. ruta. Resultaten av detta presenteras i tabell 13 där samma rutor som i tabell 12 har använts. För gång ligger det högre värdet i tabellens exempel ca. 35-50 procent högre än motsvarande lägre värde. För cykel ligger det högre värdet ca. 60-70 procent över det lägre värdet. Det kan vara intressant att relatera detta till det policyexperiment som Björklund och Isacsson (2013) redovisar då all tid som spenderas i blandtrafik, i cykelfält på vägen och cykelbanor vid sidan av vägen överförs till cykelbanor som inte ligger i anslutning till vägen. De rapporterar att andelen som cyklar med de estimerade modellerna i den uppsatsen skulle öka med mellan 20 och 34 procent beroende på hur stor andelen cyklister är i utgångsläget. Ökningar i den storleksordningen ryms alltså inom intervallen i tabell 13. Det finns med andra ord god marginal att väga in specifik information om t.ex. infrastrukturen för den/de OD-relationer man utreder och anpassa prediktionen till en slutlig bedömning för resandet i JA.

Tabell 13. Individuella predikterade sannolikheter för att välja gång resp. cykel +/- 2*standardavvikelsen för varje individ i registerdatamaterialet, genomsnitt per ruta, *några utvalda rutor i Stockholm*

X-Koordinat	*Y-koordinat*	*# Obs*	*Gång*		*Cykel*	
			*-2*stand. avvik.*	*+2*stand. avvik.*	*-2*stand. avvik.*	*+2*stand. avvik.*
6574000	1633000	333	0,069	0,094	0,042	0,072
6571000	1633000	35	0,077	0,115	0,059	0,096
6575250	1632500	281	0,070	0,101	0,047	0,079

5 Diskussion och slutsatser

Denna rapport har illustrerat hur RES mer specifikt kan användas tillsammans med högupplösta geografiska registerdata över befolkning och arbetsställen för att göra en bedömning av antalet arbetsresor som genomförs med alternativen gång och cykel före en investering genomförs (JA i GC-kalk). Det visade sig vara svårt att göra rimliga prediktioner av valet att cykla för individer som faktiskt hade valt att cykla. Detta gällde även i viss mån för alternativet gång. Därför föreslogs att predikterade sannolikheter skulle användas istället. Argumentet för detta är framförallt att RES är en urvalsundersökning där den enskilda observationen avser en resa en viss dag. Gång och cykel kan dock för många individer vara alternativ som väljs vissa dagar då väder och andra specifika förutsättningar medger att det alternativet väljs. Därmed finns det en viss individuell stokastisk variation under året vad gäller en individs färdmedelsval för resan till jobbet. Därför kan det vara rimligt att utgå ifrån predikterade sannolikheter för att en individ väljer gång eller cykel i registerdatamaterialet snarare än motsvarande predikterade val.

Genomsnittliga predikterade sannolikheter för färdmedelsvalen gång och cykel i registerdatamaterialet överensstämde relativt väl med observerade färdmedelsandelar i RES då det senare materialet delades upp efter län. För de tre största kommunerna så verkade prediktionerna dock vara något höga i förhållande till RES. Syftet med ansatsen som presenterats här är dock inte att estimera färdmedelsandelar för gång och cykel på läns- eller kommunnivå. Poängen är istället att ansatsen möjliggör prediktioner i specifika OD-relationer på så låg geografisk nivå som 250 kvadratmeter stora rutor. Då bostäder och arbetsplatser är geokodade på detta sätt kan alltså kompletta OD-matriser tas fram för små geografiska områden i vilka man kan fylla i antalet som beräknas gå eller cykla. Motivet för att sträva efter hög geografisk upplösning är att alternativen gång och cykel framförallt används på relativt korta avstånd.

För varje individ genererade också den ansats som användes här standardavvikelser för den predikterade sannolikheten att välja gång resp. cykel. Detta mått på osäkerheten i prediktionen kan användas för att anpassa prognosen i JA till specifika förutsättningar som användaren av GC-kalk känner till, t.ex. hur cykelinfrastrukturen ser ut i det specifika område som utreds och räkningar av antalet cyklister och fotgängare. Prognosen för resandet i utredningsalternativet kan därefter beräknas med hjälp av de samband som rapporterats i bl.a. Björklund och Isacsson (2013).

De modeller som presenterats här har inte kunnat inkludera information om hur nätverken för gång- och cykelinfrastrukturen ser ut då rikstäckande information om dessa nätverk saknats då analyserna i denna rapport genomfördes. Modellerna i denna rapport har istället i huvudsak förlitat sig på avståndet för resan mellan bostad och arbetsplats samt information om vilket län individen bor i. Det vore sannolikt värdefullt att i framtida arbete med prognoser för JA i GC-kalk tydligare koppla prediktionsmodeller till information om gång- och cykelinfrastrukturen då sådan blir tillgänglig för att därigenom få bättre prediktioner.

Referenser

Algers, S. and Beser, M. (2000), "SAMPERS - The New Swedish National Travel Demand Forecasting Tool", Proceedings of the IATBR Conference.

Allison, P.D. (1999), "Logistic regression using the SAS® system: Theory and application", Cary, NC: SAS Institute Inc.

Ambroise, C. & McLachlan, G.J. (2002), Selection bias in gene extraction on the basis of microarray gene-expression data, *PNAS* 99(10), 6562-6566.

Bento, A. M., Cropper, M.L., Mobarak, A.M. and Vinha K. (2005), The effects of urban spatial structure on travel demand in the United States, *The Review of Economics and Statistics*, 87(3), 466–478.

Björklund, G. & Carlén, B. (2012). Värdering av restidsbesparingar vid cykelresor. VTI notat 26-2012. Linköping: Statens Väg- och Transportforskningsinstitut

Björklund, G. & Isacsson, G. (2013). Forecasting the impact of infrastructure on Swedish commuters' cycling behavior, Scandinavian working papers in Economics, Nr. 2013:36

Björklund, G., Mellin, A., & Odolinski, K. (2013). Fotgängares värderingar av gångvägar. VTI-rapport 806. Linköping: Statens Väg- och Transportforskningsinstitut

Björklund, G. & Mortazavi, R. (2013). Influences of infrastructure and attitudes to health on value of travel time savings in bicycle journeys. Scandinavian working papers in Economics, Nr. 2013:35

Börjesson, M. & Eliasson, J. (2012), The value of time and external benefits in bicycle appraisal, *Transportation Research* Part A 46, 673–683.

Brownstone, D. and Golob, T.F. (2009), The impact of residential density on vehicle usage and energy consumption, *Journal of Urban Economics* 65, 91–98

Efron, B. (1983), Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation, *Journal of the American Statistical Association*, 78(382), 316-331.

Efron, B. (1986), How Biased is the Apparent Error Rate of a Prediction Rule?, *Journal of the American Statistical Association*, 81(394), 461-470.

Efron, B. & Tibshirani (1997), Improvements on Cross-Validation: The 632+ Bootstrap Method, *Journal of the American Statistical Association*, 92(438), 548-560.

Leeb, H. (2008), Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process, *Bernoulli* 14(3), 661-690.

Liss, V. & Isacsson, G. (2014), The Relationship Between Population Density, Driving Distances And Greenhouse Gas Emissions In Swedish Cities, Manuscript.

Newman, P.W.G. and Kenworthy, J.R. (1989), Gasoline Consumption and Cities, *Journal of the American Planning Association*, 55(1), 24–37.

Naturvårdsverket (2005), Den samhällsekonomiska nyttan av cykeltrafikåtgärder – Förbättring av beslutsunderlag, Rapport 5456.

Rietveld, P. & Daniel, V. (2004): Determinants of bicycle use – do municipal policies matter? *Transportation Research* Part A 38, 531–550.

Steyerberg, E.W., Harrell Jr, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C. (René), Vergouwe, Y., & Habbema, J.D.F. (2001), Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis, *Journal of Clinical Epidemiology* 54, 774–781.

Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobee, D.E., Moons, K.G.M. (2003), Internal and external validation of predictive models: A simulation study of bias and precision in small samples, *Journal of Clinical Epidemiology* 56, 774–781.

SIKA (2007), *RES 2005–2006 Den nationella resvaneundersökningen*, SIKA Statistik Kommunikationsmönster, 2007:19

Trafikverket (2012), GC-kalk – Manual och bakomliggande formler version 1.0.

Wardman, M., Tight, M. & Page, M. (2007). Factors influencing the propensity to cycle to work, *Transportation Research* Part A, 41, 339-350.

WSP (2007), Utvecklingsplan för att möjliggöra samhällsekonomiska kalkyler av cykelåtgärder, WSP Analys & Strategi rapport 2007:15.