Estimation and Evaluation of Discrete Choice Models[1]

Samuel Lindgren – VTI
Magnus Johansson – VTI
Inge Vierth – VTI
Rune Karlsson – VTI
Jaap Baak – Significance
Gerard de Jong – Significance

This study estimates and evaluates logit models of firms' choice of transport chain and shipment size to address questions relating to the implementation of a stochastic logistics module in the freight model Samgods. We use the 2016 Swedish Commodity Flow Survey combined with the new Samgods commodity classification and updated cost parameters. We find that the multinomial logit (MNL) model predicts observed frequencies well, both for in- and out-of-sample observations. Predicted tonnes are close to observed levels for most commodities but far off for a few others. Mispredictions can be handled by adjusting predicted shipment sizes. We find that a nested logit model does slightly better in predicting tonnes than the MNL model. A comparison model that always selects the low-cost alternative is very useful for predicting transport chains but predicts shipment sizes larger than observed. Our robustness checks show that the estimated coefficients do not systematically predict worse when applied to transport cost data that differ from those used in estimation. We also find that a coarser classification of shipment sizes leads to less precise predictions while expanding the choice set to differentiate between container and non-container options only includes alternatives with a low probability of being selected. We end by discussing our results' implications for implementing a stochastic logistics module.

Keywords: Freight, Choice model, Samgods

[1] This project was financed by the Swedish Transport Administration. We thank Jonas Westin and Petter Hill for helpful comments.

Summary in Swedish

Bakgrund och syfte

År 2013 började Trafikverket utveckla en stokastisk logistikmodul för den nationella godsmodellen Samgods. Logistikmodulen simulerar valet av trafikslag och sändningsstorlek för sändningar mellan företag i modellen. Tidigare har valet modellerats på ett deterministisk vis, där allt gods mellan två företag antas fraktas med transportlösningen som har lägst kostnad. En av nackdelarna med detta sätt är att det kan leda till osannolikt stora ändringar av användningen av de olika trafikslagen i modellen när faktorer som påverkar kostnader förändras. Detta är ett problem inte minst eftersom Samgods används för policyanalys. I en stokastisk logistikmodul å andra sidan baseras valet på en logit-formel som inte ger lika stora förändringar i val av trafikslag när kostnader ändras.

Den stokastisk logistikmodulen utgår precis som den deterministiska från en matris med godstransportefterfrågan i ton mellan produktions- och konsumtionszonerna i Samgods. Efterfrågan fördelas sedan på flöden mellan företag i tre olika storleksklasser och på flöden mellan industrier i olika zoner. Den stokastiska logistikmodulen bestämmer valet av transportkedja och sändningsstorlek för dessa flöden. Den använder sig av 16 varugrupper och modellerar valet av 14 stycken transportkedjetyper och 16 kategorier av sändningsstorlekar.

Trafikverket har tagit flera steg för att utveckla en stokastisk logistikmodul men fortfarande kvarstår flera frågor som rör i) behovet av att använda uppdaterade datakällor vid estimering, ii) olika logit-modellers lämplighet, iii) hur känsliga logit-modellerna är för förändringar i estimeringsprocessen och iv) hur väl dessa logit-modeller prognostiserar valet av trafikslag och sändningsstorlek.

Syftet med denna studie är att analysera dessa frågor. Vi använder oss av varuflödesundersökningen (VFU) från 2016 tillsammans med den senaste varugruppsindelningen och kostnadsparametrar för att skatta logit-modeller över företagens val av transportkedja och sändningsstorlek. Vi jämför olika modellspecifikationer med varandra och jämför modellernas prediktioner med de observerade valen i varuflödesundersökningen. Vi undersöker sedan hur våra resultat påverkas av förändringar i datamaterialet och antalet transportalternativ som företagen antas kunna välja.

Data

Vårt huvudsakliga datamaterial består av den svenska varuflödesundersökningen från 2016. Den innehåller information om trafikslag, sändningsstorlek, värde, vikt, lasttyp, varugrupp samt mottagande och avgående ort för fem miljoner sändningar. Vi översätter valet av trafikslag i VFU:n till valet av en av de 14 kedjetyperna som används i den stokastiska logistikmodulen.

Den andra datakällan utgörs av beräknade transportkostnader för de kombinationer av kedjor och sändningsstorlekar som är tillgängliga för sändningarna i varuflödesundersökningen. Vi beräknar dessa kostnader genom att först koppla avgång- och ankomstort för varje sänding

till respektive Samgodszon och sedan använda den deterministiska logistikmodulen i Samgods för beräkning av bästa ruttval och tillhörande transportkostnader. Vi kopplar kedjorna som finns förprogrammerade i Samgods till de 14 kedjetyperna i den stokastiska logistikmodulen. Slutresultatet är ett datamaterial med information om fem miljoner sändningar (såsom varugrupp, lasttyp och värde) och kostnader för de transportalternativ som finns tillgängliga för varje sändning.

Metod

I vår huvudanalys använder vi oss av varuflödesundersökningen och transportkostnaderna för att skatta logit-modeller över val av transportkedja och sändningsstorlek. Baserat på de skattade koefficienterna gör vi sedan prediktioner av antal sändningar för varje transportkedja och sändningsstorlek samt mängden ton för varje transportkedja och jämför dem med de observerade värdena i varuflödesundersökningen. I en fördjupande analys undersöker vi hur ändringar i våra inputdata, kategorier av sändningsstorlekar och antalet alternativ påverkar våra huvudresultat. Vi analyserar också om våra resultat förändras om vi använder oss av andra typer av logit-modeller.

Vår huvudanalys utgörs av skattningar av multinomiala logit-modeller (MNL). I dessa modeller krävs en funktion som beskriver den mätbara nyttan, V_{st}, som varje företag antas erhålla vid valet av en trafikkedja t och sändningsstorlek s:

$$V_{st} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

där x-värdena betecknar observerbara variabler såsom transportkostnader och transporttid. Funktionen innehåller även koefficienter β som beskriver hur nyttan av ett alternativ påverkas av att en observerbar variabel förändras. Dessa koefficienter skattar vi i studien. När nyttofunktionen för varje alternativ är skattad kan sannolikheten att välja ett av dessa alternativ beräknas genom följande logit-formel:

$$P_i(t,s) = \frac{\exp(V_{st})}{\sum_{s',t'} \exp(V_{s't'})}$$

Denna formel använder vi för att göra våra prediktioner. Som jämförelsemodell definierar vi en beslutsregel som säger att för varje sändning kommer det transportalternativ med lägst transportkostnad C att användas med sannolikhet ett:

$$P_i(t,s) = \begin{cases} 1 \; om \;\; C_{its} < C_{ikl} \;\; \forall k, \forall l \\ 0 \; annars \end{cases}$$

Denna modell har likheter med beslutsregeln hos den deterministiska logistikmodulen i Samgods som antingen allokerar allt eller inget av godset till varje transportalternativ.

Huvudresultat

Genom att jämföra prediktioner med observerade val i varuflödesundersökningen visar vi att vår jämförelsemodell med allt-eller-inget-allokering ger exakta prediktioner när det kommer till antalet sändningar för de 14 kedjorna. Detta visar att principen om lägsta

transportkostnader förklarar vissa mönster på transportmarknaden väldigt väl. Å andra sidan ger jämförelsemodellen prediktioner av sändningsstorlek som är mycket högre än de som observeras i varuflödesundersökningen .

Logit-modellen ger prediktioner av antalet sändningar per kedja och sändningsstorlek som är väldigt nära de observerade värdena. Även prediktionerna av mängden ton är precisa för de flesta av varugrupperna vilket visar att logit-modellen kan överträffa jämförelsemodellen. Logit-modellens prediktioner avviker dock kraftigt från observerade nivåer för ett fåtal av varugrupperna. Detta verkar inte bero på att en särskild kedja är svår att göra prediktioner för utan snarare är det kombinationen av kedja och varugrupp som är svårfångad.

Fördjupande analys

I flera ytterligare analyser fördjupar vi oss i våra huvudresultat. Vi visar att svårigheten att göra prediktioner för några av varugrupperna kan hanteras genom att inkludera en justeringsfaktor för att skala upp eller ner mängden predikterade ton. Med andra ord är det svårare för logit-modellen att ge en god prognos av hur stora sändningarna är än vilka transportkedjor som kommer att användas.

I huvudanalysen använer vi oss av ett set av valbara alternativ bestående av 14 kedjor och 16 sändningsstorlekar. Vi visar att när vi expanderar setet till att bestå av 14 kedjor med containertransporter och 14 kedjor utan containertransporter (samt 16 stycken sändningsstorlekar) inkluderas nya alternativ som har liknande transporttid och avstånd men lägre transportkostnader i genomsnitt. Skillnaden i genomsnittskostnader minskar markant när vi fokuserar på alternativ som valdes i varuflödesundersökningen. Det tyder på att de alternativ som bara inkluderades i det expanderade setet hade låg sannolikhet att bli valda av företagen ändå.

Vi testar även att minska antalet valbara sändningsstorlekar från 16 till fem stycken och visar att det leder till att de valbara alternativen har lägre kostnader i genomsnitt jämfört med det ursprungliga antalet storlekar. Prediktionerna som baseras på det lägre antalet sändningsstorlekar är sämre, vilket visar på fördelen med att behålla en mer detaljerad uppdelning av storlekarna. Därtill jämför vi vår MNL-modell med en så kallad nästlad logit-modell som tillåter mer komplexa substitutionsmönster mellan olika transportalternativ. Vi finner att prognoserna baserade på den nästlade logit-modellen är något bättre.

Till sist visar vi att de koefficienter som vi skattat i huvudanalysen inte systematiskt ger sämre prediktioner när de appliceras på transportkostnader som skiljer sig något från de som använts vid skattningarna. Prediktionerna är mer exakta för vissa kedjor och mindre exakta än andra jämfört med våra huvudresultat. Det visar att de skattade koefficienterna är någorlunda robusta för mindre ändringar i inputdata.

Konsekvenser för implementering av en stokastisk logistikmodul i Samgods

Ett av våra huvudsakliga bidrag är att producera koefficienter som kan användas i logit-formeln i den stokastiska logistikmodulen. Dessa koefficienter visas i appendix A. När vi applicerar dessa på sändningarna i varuflödesundersökningen ges prediktioner av antalet

sändningar och ton som för de flesta varugrupper överensstämmer med de observerade värdena.

Våra resultat visar att justeringar troligtvis måste göras om dessa koefficienter ska användas i logit-formeln i den stokastiska logistikmodulen. I vår analys behövde vi justera sändningsstorleken upp och ner för de varugrupper där prediktionerna var långt ifrån de observerade värdena. Ett alternativt tillvägagångssätt är att inkludera logaritmen av skillnaden mellan observerade och predikterade värdet i modellspecifikationen för varje transportkedja. Denna ansats användes vid utvecklingen av en stokastisk logistikmodul för två varugrupper i ett tidigare projekt (Abate et al. 2016).

De valbara transportalternativen i vår analys och i den föreslagna stokastiska modulen utgörs av 14 kedjor och 16 sändningsstorlekar. När vi utökar de valbara transportalternativen till att inkludera både containertransporter och icke-containertransporter inkluderas bara alternativ som inte är särskilt relevanta för företagens val i varuflödesundersökningen. Däremot förbättras kvaliteten i prediktionerna vilket är en god anledning att se över hur många valbara alternativ som skall ingå. Men fördelarna av att öka antalet alternativ bör vägas mot den ökade bördan av att implementera och köra en sådan logistikmodul i Samgods.

Det finns flera tänkbara förändringar i definitionen av den stokastiska logistikmodulens komponenter, inte minst hur fordonstyperna i Samgods ska kopplas till varje transportkedja i modulen och vilka kedjor och sändningsstorlekar som skall ingå. Om sådana förändringar genomförs i framtiden är det möjligt att på olika sätt kombinera de redan skattade koefficienterna för olika kedjor och sändningsstorlekar utan att behöva skatta nya koefficienter.

De koefficienter vi har skattat framstår som relativt robusta för alternativa värden på transportkostnaderna. Mindre förändringar hos de faktorer som bestämmer dessa värden bör därför vara oproblematiska. Skulle större förändringar i inputdata (och transportkostnader) genomföras i Samgods är det möjligt att justera koefficienterna för att ta hänsyn till att variablernas skala har förändrats, utan att behöva skatta nya koefficienter. Koefficienterna kan justeras genom att multipliceras med en faktor lika med skillnaden mellan det gamla och nya medelvärdet för varje variabel.

Vi har i analysen genomfört skattningar och prediktioner för samtliga 16 varugrupper, men det är inte uppenbart att en stokastisk logistikmodul krävs för alla grupper. En anledning att använda en stokastisk modul är att förhindra osannolikt stora förändringar i efterfrågan på olika transportkedjor. Men sådana förändringar lär inte uppstå för varugrupper där en enstaka kedja utgör hela marknaden. Vi visar att det bara är för varugrupperna 1 (jordbruks-, skogsbruksprodukter) och 2 (kol, gas och olja) som en kedja har mer än 90 procent av mängden ton som transporteras.

En anledning till att använda den deterministiska ansatsen är svårigheten att modellera valet av transportalternativ. I vår huvudanalys visar vi att prediktionerna för varugrupperna 2, 3, 5, 7 och 14 är långt ifrån de observerade värdena. Samtidigt går det att justera sändningsstorlekarna för att prediktionerna ska vara relativt precisa igen. Baserat på detta

drar vi slutsatsen att frågor som rör implementeringen av en mixad logistikmodell (med både deterministisk och stokastisk modellering) väger tyngre vid beslutet om vilka varugrupper som ska modelleras på vilket sätt.

Vi finner att prediktionerna baserade på den nästlade logit-modellen är något bättre än de baserade på MNL-modellen. På så vis finns det fördelar med att basera den stokastiska logistikmodulen på en nästlad logit-modell, inte minst med tanke på att vi använde oss av en relativt enkel nästlad modellspecifikation och att det sannolikt finns förbättringspotential. Men den nästlade modellen ställer högre krav på beräkningsprestanda och enkelheten i MNL-modellen är ett starkt argument för att använda den i den stokastiska logistikmodulen.

Introduction

In 2013, the Swedish Transport Administration started the process to develop a stochastic logistics module for the national freight transport model Samgods. The logistics module includes the combined choice of shipment size and transport chain for firm-to-firm relations in the model. The choice has previously been modelled in a deterministic way, where all tonnes on a firm-to-firm relation is assigned to the transport solution with lowest cost.

One of the downsides with such selection rule is that it can lead to implausibly large swings in modal shares following changes in the variables determining costs. This is a concern not least because Samgods is used for policy analysis. The Swedish Transport Administration has taken several steps towards the development of a so-called stochastic logistics module, where the logistics choices are based on a logit-formula and coefficients are estimated on observed choices by shippers. This way of modelling choices reduces the risk of sharp changes in modal shares and has a stronger behavioral foundation.

Despite earlier work on this topic there remains several questions regarding the production of logit-based coefficients that can be used in a stochastic logistics module. The purpose of this study is to address these questions by estimating and evaluating logit models, produce coefficient estimates that can be used in the implementation of a stochastic logistics module and analyze how our results are affected by changes in the main analysis, including the size of the choice set, weight class interval and new input data.

The study is outlined as follows. The following section gives a background to the proposed stochastic module and previous projects. The subsequent sections describe the data and data processing steps we take, outlays details of our estimation and evaluation method and the results. The final section concludes.

Background

Overview of the stochastic logistics module

The logistics module in Samgods simulates the logistics decisions (regarding shipment size, use of consolidation and distribution centres, mode- and vehicle/vessel type and loading unit type) at the firm-to-firm level. The joint choice of shipment size and transport chain has previously been modelled in a deterministic way, although changes implemented in version 1.2 of the model allow for not only one but two choices on each firm-to-firm relation. [2]

A proposal for how a stochastic logistics module in Samgods would work has been outlined in previous reports (Abate et al. 2014, 2016). As with the deterministic module, the starting point is the matrix of freight transport demand between producer and consumer zones (a PC-matrix). The matrix consists of estimated tonnes of goods to be transported from each origin to each destination in a given year. The flows in the PC-matrix are disaggregated into flows between three different size classes of firms (small, medium and large) and a group for volumes to be sent in large shipments between industries.

[2] Transport cost is set to follow a stochastic distribution so that realized transport cost may be above or below the expected average cost. This creates the possibility that the second-best chain will have lower realized cost than the best chain, although the latter has lower expected costs.

The proposed stochastic module also works at the firm-to-firm level to select the transport chain and shipment size. It operates on 16 commodities and have a choice set that consists of 14 chain types and 16 shipment size categories. Each choice alternative is made up of a combination of a chain type and a shipment size interval so there are 14 times 16 = 224 alternatives in total. Table 1 shows the dimensions of the proposed stochastic module.

No.	Chain type	No.	Shipment size
1	Truck	1	0-50 kg
2	Vessel	2	51-200 kg
3	Rail	3	201-800 kg
4	Truck-Vessel	4	801-3000 kg
5	Rail-Vessel	5	3001-7500 kg
6	Truck-Truck-Truck	6	7501-12500 kg
7	Truck-Rail-Truck	7	12501-20000 kg
8	Truck-Ferry-Truck	8	20001-30000 kg
9	Truck-Vessel-Truck	9	30001-35000 kg
10	Truck-Air-Truck	10	35001-40000 kg
11	Truck-Ferry-Rail-Truck	11	40001-45000 kg
12	Truck-Rail-Ferry-truck	12	45001-100000 kg
13	Truck-Vessel-Rail-Truck	13	100001-200000 kg
14	Truck-Rail-Vessel-Truck	14	200001-400000 kg
		15	400001-800000 kg
		16	800001+ kg

Table 1. Overview of choice set

The choice model in the stochastic logistics module is given by the logit-formula. It says that the probability that alternative k is selected for observation i among K different alternatives is

$$Pr_i(k) = \frac{\exp(V_{ik})}{\sum_j^K \exp(V_{jk})}$$

$$V_{ik} = \boldsymbol{X_{ik}\hat{\beta}}$$

where V_{ik} is called the representative utility of alternative k for observation i, that is specified to depend on the set of variables X_{ik} and coefficients $\hat{\beta}$**.** In the stochastic logistics module, each observation i would refer to a firm-to-firm flow of a particular commodity going from one Samgods zone to another. The number of different alternatives K would be 224 at a maximum.

The logit-probability has several desirable features: it is necessarily between 0 and 1 and the probabilities for all alternatives sum to 1. The association between the logit probability and representative utility is S-shaped, as shown in Figure 1. This means that if the representative utility is very low or very high compared to the alternatives, an increase in the utility of that option (e.g. due to transport cost reductions) changes its choice probability by very little. This feature is what reduces the swings in modal shares following changes in policy variables and makes the logit-formula suitable to use in the logistics module in Samgods.

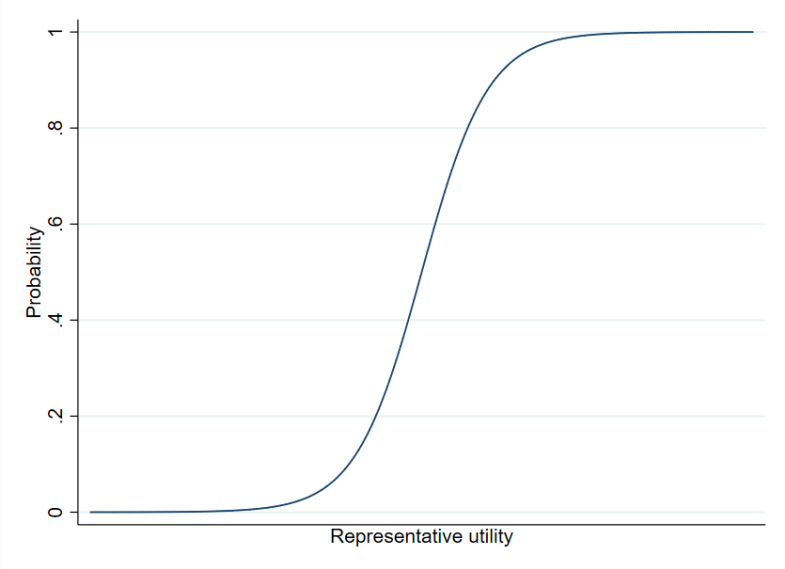


Figure 1. Graph of logit curve

The first step in the stochastic logistics module is to reduce a longlist of some 100 transport chains that are pre-defined in Samgods to the shortlist of the 14 chain types. The next step is to calculate for each firm-to-firm relation on each OD-pair the probability that each alternative (i.e. combination of chain type and shipment size) is chosen. This probability is given by the logit-formula. The stochastic module then sums the tonnes-weighted probabilities over all firm-to-firm relations to produce matrices containing the tonnes transported by mode between origin and destination zones in Samgods. In a memo accompanying this study, we provide a longer description of our suggested stochastic logistics module (Lindgren et al. 2019).

Previous literature

Implementation of the stochastic logistics module requires several decisions to be made regarding the logit formula. This includes selecting one of the many different discrete choice models that are based on the logit-formula; specifying how to model firms' choice of *which* mode to use and their decision of *how much* to ship; deciding which variables that should enter the representative utility and their functional form as well as estimating their coefficients.

Some of these questions have been analyzed in previous work. Abate et al. (2014) estimated and reviewed various econometric models of discrete choice that can be used to describe firms' choice of shipment size and mode. They compared three types of discrete choice models; one that only consider the mode choice decision, another of the joint choice of mode and discrete shipment size choice and a third of the joint choice of mode and continuous shipment size choice. Their conclusion was that converting shipment sizes into discrete categories gives other behavioral responses but might be preferable when using large datasets such as the Commodity Flow Survey. Thus, this is the approach we also take in this study.

Abate et al. (2016) estimated and implemented models of the joint choice of transport chain and discrete shipment size. The logit models were estimated separately by commodity group using the old commodity group classification. We follow their approach and estimate one logit model per commodity group. They implemented a stochastic logistics module for two of the

commodity types based on the estimated coefficients. The authors showed that cost and time elasticities produced by the stochastic module was closer to zero compared to elasticities based on the deterministic module.

There remain several questions regarding the production of logit-based coefficients that can be used in a stochastic logistics module. These broadly concern i) the need to base estimation on updated data sources that are internally consistent, ii) the suitability of different logit model specifications, iii) the sensitivity of logit model performance to changes in the estimation process and iv) how well a logit-based formula works for predicting the transport chain and shipment size choice.

This study analyses these questions. We use the 2016 Commodity Flow Survey for Sweden together with the new Samgods commodity classification and cost parameters that have been updated to the same year as the CFS to estimate discrete choice models of firms' joint choice of shipment size and transport chain. We test various model specifications, make predictions of modal shares using the best logit model and compare those against the observed choices in the CFS. The resulting coefficients can be readily applied in the logit-formula in a future stochastic logistics module. We then analyze how our results are affected by changes in the main analysis, including the size of the choice set, weight class interval and new input data.

Data

Commodity Flow Survey

The main data source used in this project is the Swedish Commodity Flow Survey from 2016. The data set contains records of some 5 million individual shipments to or from a company in Sweden, with information about the origin, destination, value, weight, cargo type and commodity class of the shipments.

The survey also contains information about the mode(s) used to transport each shipment. It distinguishes between road, rail, sea and air transport and includes an additional category for unknown modes. We translate the observed mode choice in the CFS to a choice of one of the 14 chains in the choice set of the proposed stochastic logistics module. The correspondence of these sets is shown in table 2.

Chain 1 is truck and corresponds to shipments in the CFS that has recorded using only road transportation. Chains 2 and 3 are direct transportation by vessel and sea and include CFS shipments using only the corresponding mode (or in combination with an unknown mode). Chain 4 (truck-vessel) and 5 (rail-vessel) consist of various combinations using either only road and sea transportation or rail and sea transportation (in combination with an unknown mode in some instances).

Chain 6 (truck-truck-truck) corresponds to CFS shipments recording road in combination with an unknown mode. The information in the CFS makes it difficult to differentiate between chain 1 (truck) and 6 (truck-truck-truck). The latter chain consists of a leg with a small truck, followed by consolidation in a larger truck and ends with a smaller truck on the last leg. This chain cannot reliably be taken from the CFS as the survey does not distinguish between vehicle size. In practice, the CFS combinations corresponding to chain 6 are rarely recorded in the CFS.

NO.	Shortlist chaintype	CFS combinations included (V=road, J=rail, S=sea, L=air, X=unknown)
1	Truck	V
2	Vessel	S, SX,
3	Rail	J, JX
4	Truck-Vessel	SV, SVS, VS, VSX
5	Rail-Vessel	JS, JSJ, JSX, SJ, XSJ
6	Truck-Truck-Truck	VXV, VX, XV,
7	Truck-Rail-Truck	JV, VJ, VJV, VJX
8	Truck-Ferry-Truck	
9	Truck-Vessel-Truck	VSV, VSVSV
10	Truck-Air-Truck	All chains containing L
11	Truck-Ferry-Rail-Truck	
12	Truck-Rail-Ferry-truck	
13	Truck-Vessel-Rail-Truck	SJV, SVJV, VSJ, VSVJ, VSJV, XSJV
14	Truck-Rail-Vessel-Truck	VJS, VJSJ, VJSJV, VJSV, VJSX, VJVS, VJVSV, VJX, JSJV, JSV, JVS

Table 2. Correspondence shortlist and CFS combinations

Chains 7 (truck-rail-truck) and 9 (truck-vessel-truck) is made up by shipments using road in combination with rail or vessel respectively. Chains 13 (truck-vessel-rail-truck) and 14 (truck-rail-vessel-truck) both consist of chains where road, rail and sea transportation are utilized. We differentiate between the two chains by allocating shipments that use sea before rail transportation to chain 13 and the opposite to chain 14. We link all shipments using air transportation to chain 10 (truck-air-truck).

Unfortunately, the CFS does not allow the identification of ferries from vessels. We decided to regard all records of sea transportation as a vessel so the chains containing ferries (8, 11, 12) will not be coded as chosen in the data. This is motivated by vessels being the more common mode than ferries and that identification of ferries based on e.g., shipment origin and destination involves a degree of arbitrariness.

Figure 2 shows the tonnes modal shares of the chains for each commodity based on the shipments in the CFS 2016. Overall, there is substantial variation in chain types shares, although direct road transportation (chain 1) is the most common choice for most commodity groups. There are only two groups where a single chain constitutes more than 90 percent of the market. Direct road transportation dominates in commodity group 1 (agriculture and forestry products (excluding round wood) and fishing) while vessels (chain 2) has virtually all of the market for transportation of group 2 (coal, crude petroleum and natural gas).

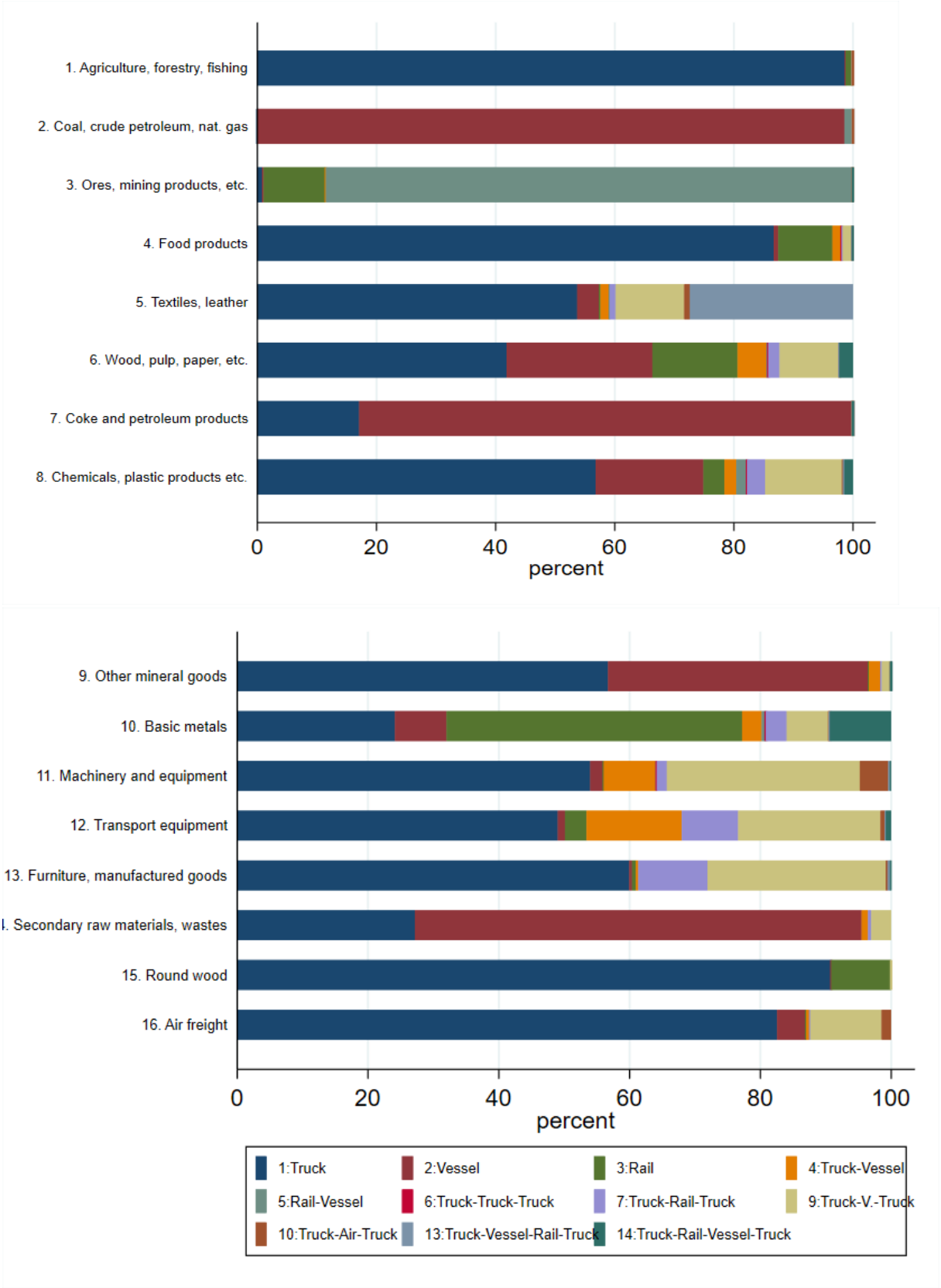


Figure 2. Chain shares in tonnes, based on CFS 2016

Samgods-based data

For our logit model estimations, we will use information about the cost, time and distance of chosen and non-chosen alternatives facing each shipment in the CFS. We create these variables using the existing logistics module in Samgods as follows. First, a regular model run is performed with the deterministic model to determine the consolidation factors that will be used in the cost calculations. These consolidation factors will be fixed during the steps that will be described below. Second, we identify the Samgods zone of origin and destination for each shipment in the CFS 2016. Samgods zones corresponds to municipalities in Sweden, larger regions in the rest of Scandinavia and some European countries and to countries or country-combinations in the rest of the World. We geocode the origin and destination using the sender's and receiver's postal address or using the country code when the postal address is unavailable or unreliable.

We then create a synthetic production-consumption-matrix for the Samgods zones based on the observations in the CFS. We subsequently run a special version of the deterministic logistics module where the cost-minimizing option per chain type is determined. This procedure is run at a fixed annual shipment frequency of 1 and produces the best chain for each of the chain types that are available for the particular origin- and destination-pair being considered. For each of these best chains the route (and corresponding transport time, distance and costs) and vehicle type being used are determined. This step therefore effectively determines the set of available chains for each shipment in the CFS.

This step requires the use of cost parameters as input. The cost parameters that are currently available in Samgods refer to the base years 2012 or 2014 while we which them to be related to the same year as the CFS 2016 for internal consistency. We therefore update the cost parameters to their 2016 level using price indices. Table 3 shows the indices used to adjust cost parameters.

Cost category	Cost parameter	Mode	Index (source)	Adj. factor
Vehicle parameters	All	Road	Swedish SPPI for road freight transportation[a] (SCB)	1,009
		Rail	Swedish SPPI for land-based transportation[b] (SCB)	1,025
		Sea	Swedish SPPI for maritime freight transportation[c] (SCB)	1,041
		Air	Swedish CPI (SCB)	1,009
Cargo	Product value	All	Swedish CPI (SCB)	1,009
	Inventory costs	All	Service price index, group H: transport and handling services (SCB)	0,996
	Order costs	All	Service price index, group H: transport and handling services (SCB)	0,996
Kiel Canal	Link-based tax	Sea	CPI EU28 (OECD)	1,023
Tax per country rail and road	Road and rail tax	Road	CPI EU28 (OECD)	1,023
		Rail	CPI EU28 (OECD)	1,023
Toll links	Toll link	Road	Swedish CPI (SCB)	1,009
		Rail	Swedish CPI (SCB)	1,009
Pilot fees	Pilot fees	Sea	Swedish CPI (SCB)	1,009
Tax link	Tax link	Sea	No update	-

Note: [a] SPPI product group 49.41 ,[b]SPPI product group 49., [c] SPPI product group 50.2

Table 3. Overview of cost parameters and indices

Vehicle cost parameters are adjusted according to the Swedish Producer Price Index for Services (SPPI) and consumer price index (CPI). Cargo-specific costs are adjusted according to various price indices. The development of product values is assumed to be well-approximated by that of the general consumption bundle. More detailed indices are available (e.g. producer price index by product linked to sector (SPIN)) but their high-resolution makes them sensitive to price shocks. Inventory and order costs are approximated by price of services for transport and handling., Swedish and international taxes and fees are adjusted by the Swedish and EU-level CPIs respectively.

The cost parameters covering product value, inventory costs and order costs are specific to a single commodity group. The existing cost parameters follows the old commodity classification in Samgods that consists of 35 groups, while our analysis is based on the 16 commodity groups used in Samgods version 1.2 (Vierth et al. 2017). We solve this by linking the commodities and their costs from the old Samgods classification to the commodities in the new classification based on the correspondence of sub-groups within each commodity.

Within the Commodity Flow Survey only the size of the observed shipment is available. The annual volume on a specific firm-to-firm relation is not available from the Commodity Flow Survey. For this reason, it is not possible to calculate the order and storage costs for the observed transport flows. Therefore, the stochastic model is estimated on the transport costs only, order and storage costs are not included in the estimation.

The transport costs are calculated for all combinations of the 16 shipment sizes and sub-modes available in Samgods. We link the sub-modes to the 14 chain types in the choice set that the proposed stochastic logistics module will operate on by creating a correspondence key. Matching between the set of 14 chain types and the sub-modes is non-trivial because the former is specified in terms of modes (truck, vessel, rail and air) while the sub-mode combinations are given at the Samgods vehicle type level (e.g. feeder train, wagonload train).

NO.	Shortlist chain type	Principle for inclusion
1	Truck	No transfers
2	Vessel	No transfers to other modes
3	Rail	No transfers to other modes
4	Truck-Vessel	Vessel taken to be all sea modes apart from road and rail ferry
5	Rail-Vessel	Vessel taken to be all sea modes apart from road and rail ferry
6	Truck-Truck-Truck	Trucks in combination
7	Truck-Rail-Truck	Any combination containing only truck and rail
8	Truck-Ferry-Truck	Ferry taken to be road ferry (sub-modeNr P)
9	Truck-Vessel-Truck	Vessel taken to be all sea modes apart from road and rail ferry
10	Truck-Air-Truck	All chains including air
11	Truck-Ferry-Rail-Truck	Includes one of only two chain types containing rail ferry (HQH)
12	Truck-Rail-Ferry-truck	Includes one of only two chain types containing rail ferry (GHQH)
13	Truck-Vessel-Rail-Truck	Order of rail/vessel important, vessel taken to be all sea transports apart from road and rail ferry
14	Truck-Rail-Vessel-Truck	Order of rail/vessel important, vessel taken to be all sea transports apart from road and rail ferry

Table 4. Correspondence shortlist and Samgods transport chains

Table 4 shows the principle for defining a sub-mode in Samgods as a particular chain in the choice set. This means for instance that chain 1 (truck) consists of sub-modes that only uses direct road (i.e., sub-modes A, B, X, C, S and c). Chain 2 (vessel) consists of sub-modes that only use sea transportation without transfers to other modes (sub-modes J, KL, LK and M) and so forth. The full correspondence key between chains and sub-modes is shown in Appendix B.

When we calculate the costs representative for each chain type and more than one available sub-mode can be linked to a chain, the sub-mode with the lowest cost is set to represent that chain. In this way, for each shipment in the CFS there is information of the cost, time and distance for all combinations of the 16 shipment sizes and 14 chains (or at least for all available chains).

The procedure used to generate the costs involves several assumptions about the input data, not least the cost parameters. In the analysis we test our results against an alternative cost variable to assess the robustness of our procedure.

Methodology

In our main analysis we use the CFS in combination with estimated transport cost, distance and time to estimate discrete choice models of the transport chain and shipment size choice. The choice set consists of the 14 chain types and 16 shipment size categories included in the choice set of the proposed logistics module.

Based on the available set of variables, we try different model specifications (variable combinations and functional form) and choose the best model to make predictions. We end our main analysis by comparing these predictions to observed levels in the CFS and quantify the deviation of predictions to observed output. We then extend the analysis to investigate whether variations in input data, weight class interval and size of choice set alter the main results. We also examine if our findings are changed if we use choice models with other assumptions about choice patterns.

Models

Our main model is the multinomial logit (MNL). It assumes that the representative utility of choosing the combination of shipment size s and transport chain t for shipment i is

$$V_{ist} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where β_k is the coefficient on variable x_k. Exactly which variables to be used will be determined in our analysis. Under some distributional assumption about the unobserved component of the utility (Train 2003), the probability that transport chain t and shipment size s is used for shipment i is given by:

$$P_i(t,s) = \frac{\exp(V_{ist})}{\sum_{s',t'} \exp(V_{is't'})}$$

This expression is the basis for which our predictions will be made and the formula that enters the stochastic logistics module. As a benchmark to compare our logit model predictions to, we define a mock-deterministic model that for each decision-maker selects the alternative with lowest transport costs C:

$$P_i(t,s) = \begin{cases} 1 \; if \; C_{its} < C_{ikl} \;\; \forall k, \forall l \\ 0 \; Otherwise \end{cases}$$

This decision rule has the same all-or-nothing assignment found in the deterministic logistics module, although the one used here is arguably coarser as it lacks feedback loops and does not account for rail capacity constraints.

Estimation

Estimation follows the standard procedure for logit models. The probability that transport chain t and shipment size s was actually chosen for shipment i can be expressed as

$$\prod_{t,s} P_i(t,s)^{I_{its}}$$

where $I_{its} = 1$ if shipment i is transported using chain t and size s and 0 otherwise, and the multiplication is over all chain and size-combinations available for shipment i. The probability of all of the N shipments in the sample being transported by the chain and size actually chosen is given by the likelihood

$$L(\beta) = \prod_{i=1}^{N} \prod_{t,s} P_i(t,s)^{I_{its}}$$

which depends on the vector of coefficients β through the logit-formula $P_i(t,s)$. The log-likelihood is then:

$$LL(\beta) = \sum_{i=1}^{N} \sum_{t,s} I_{its} \times \ln P_i(t,s)$$

and the estimator of β is the value the maximizes this function. To handle the large sample size associated with 5 million shipments and up to 224 alternatives for each shipment, we randomly sample 10 alternatives (including the chosen one) for each shipment.[3] The estimation is run on a 20 percent random sample of each commodity group separately. Running commodity-specific models facilitates the estimation procedure and produces commodity-specific coefficients directly.

Variable Selection and Model Fit Evaluation

In principle, any variable set can be included in the utility specification in the logit model. The included variables can in turn be specified in various functional forms, including linear, log-linear and splines. They could also be set to be specific to one or more of the choice alternatives.

[3] Sampling non-chosen alternatives with equal probability as is done here provides consistent estimates of the coefficients (Train 2003).

In this project, we first restrict the model specifications to be estimated and evaluated based on two factors. First, all variables included in the model must also be possible to make available in the Samgods model. Finding a model with high predictive accuracy is of limited use if the coefficient cannot be applied in the logistics module of Samgods due to lack of data. Second, we restrict attention to the variables available in the CFS 2016 and the alternative-specific cost, time and distance variables estimated in Samgods. The table below shows the variables and their functional form tested in the utility specification for some or all commodity groups.

Variable	Unit	Functional form tested
Transport cost	SEK/tonne	Linear, scaled by shipment value, log,
Transport time	Hours	Linear
Value density	SEK/kilogram	Linear
Chain type	Dummy (0/1)	Linear
Shipment size	Dummy (0/1)	Linear

Table 5. Variables and their functional form tested in utility specification

Our choice of variables to include is based on the following procedure. First, we select the number of chain and shipment size dummies to include for every commodity group. This choice is based on the accuracy with which the coefficients were estimated and whether the category was observed in the estimation sample. Coefficients for chain and shipment sizes that are seldom chosen are difficult to estimate with precisions and these variables are therefore left out of the model. Second, we select the number of cost variables to include in the model. We let cost enter linearly and separately for each chain and keep the cost variable if the coefficient has the expected negative sign. In the third step we estimate the coefficient on a common transport time variable and keep the variable if the coefficient has the expected negative sign. This procedure implies that the best model specification can vary between commodity groups.[4]

Model Prediction Evaluation

We use the coefficients from our best logit models to make predictions both for the observations used for estimation (i.e. in-sample predictions) and those not used for estimation (out-of-sample predictions). We make separate predictions for each commodity group. In-sample observations constitute 20 percent of all observations, which translates to between 840 and 1.6 million observations depending on the size of the commodity group.[5]

Our prediction targets are the number of shipments per transport chain and shipment size and the tonnes by each transport chain and in total. These are standard measures of modal shares, used in Samgods and straightforward to calculate in the validation data (CFS). The observed frequency by transport chain t in the CFS is given by

[4] An alternative way of selecting model specification is to use some measure of model fit, such as the log-likelihood and Akaike/Bayesian Information Criteria. We tested this approach but found that the included variables sometimes were very imprecisely measured and had coefficients with the unexpected sign. We argue that such coefficients would be of little use in Samgods.

[5] Appendix A shows the number of observations used in estimation for each commodity group.

$$N(t) = \sum_{i=1}^{N} I_{it}$$

where I_{it} is an indicator equal to 1 if shipment i uses transport chain t and 0 otherwise. Similarly, the observed frequency by shipment size s in the CFS is

$$N(s) = \sum_{i=1}^{N} I_{is}$$

where I_{is} is an indicator equal to 1 if shipment size s is chosen and 0 otherwise. Our predicted transport chain frequency is obtained as follows. For each shipment, the logit and mock-deterministic model predicts the probability that the combination of transport chain t and shipment size s is chosen: $P_i(t,s)$. These formulas are shown above. Summing over the 16 shipment sizes alternatives within a shipment gives the probability that transport chain t is used for shipment i:

$$P_i(t) = \sum_{s=1}^{16} P_i(t,s)$$

Summing over all N observations gives the predicted frequency using transport chain t:

$$\widehat{N}(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P_i(t,s)$$

Analogously, our predicted frequency using shipment size s is obtained by summing the probabilities $P_i(t,s)$ over the 14 transport chains and N observations

$$\widehat{N}(s) = \sum_{i=1}^{N} \sum_{t=1}^{14} P_i(t,s)$$

The observed tonnes by transport chain t in the CFS is given by

$$Q(t) = \sum_{i=1}^{N} I_{it} \times w_i$$

where I_{it} again is the indicator equal to 1 if shipment i uses transport chain t and 0 otherwise, and w_i is the observed shipment weight. Our prediction of tonnes by each transport chain is obtained as follows. We first sum over the shipment sizes to get the probability that transport chain t is used for shipment i

$$P_i(t) = \sum_{s=1}^{16} P_i(t,s)$$

We introduce tonnes by multiplying the probability that the combination of chain t and size s is chosen by the weight of shipment size s. Because each size refers to a weight interval, we use the midpoint of each interval as the actual weight. (For instance, the weight used in interval 20-30 tonnes is 25 tonnes). Letting $\bar{s}$ be the average weight in interval s, s=1, 2, …, 16, the predicted tonnes by transport chain t of shipment i is

$$\hat{Q}_i(t) = \sum_{s=1}^{16} P_i(t,s) \times \bar{s}$$

Finally, we sum over all N shipments to get the predicted tonnes by transport chain t:

$$\hat{Q}(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P_i(t,s) \times \bar{s}$$

The way our predictions are calculated will have an impact on the final results. As will be shown, the choice of weight used to predict shipment size can be adjusted to reach the desired tonne level targets for commodities for which the initial prediction was poor.

We present our results graphically. We also evaluate the predictive performance of the logit models by calculating the root mean squared errors (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum \varepsilon_{tc}^2}$$

where $\varepsilon_{tc} = \hat{Q}_c(t) - Q_c(t)$ is the deviation of the predicted tonnes from observed tonnes for chain t and commodity c. We calculate the RMSE for each commodity and for each transport chain. The RMSE is a standard measure frequently used to measure the difference between observed and predicted values. It is always non-negative and the closer the RMSE is to zero, the more accurate is the prediction.

Main Results

Model specification

The best model specification was for each commodity group c to set representative utility equal to the following:

$$V_{st} = \sum_{t=1}^{T_c} \delta_t d_t + \sum_{s=1}^{S_c} \mu_s d_s + \sum_{t=1}^{\tau_c} \beta_t (Cost_{st} \times d_t) + \theta Time_{st}$$

where d_t is a dummy equal to 1 for transport chain t and 0 otherwise, d_s is a dummy equal to 1 for shipment size s and 0 otherwise, $Cost_{st}$ and $Time_{st}$ are the transport cost and time of chain-size combination st respectively. T_c and S_c is the number of chain constant and shipment size constants included for commodity c. τ_c is the number of chain-specific cost coefficients for commodity c.

This means that a separate coefficient was estimated on each transport chain constant, each shipment size constant and for the cost variable for each chain. The coefficient on transport time was set to be the same for all alternatives and entered the specification only for one commodity group. Note that none of the coefficients depend on the attribute of the shipment, except of course on the commodity group. Transport cost and time enters linearly in the utility specification as using a logarithmic specification of cost did not improve the fit in the cases we tried. Table 6 summarizes the specification for each of the 16 models.

Commodity	# chain constants	# size constants	# Cost-by-chain	Transport time	# parameters
1	1	15	No	No	16
2	3	2	No (common cost)	No	6
3	6	15	4	No	25
4	6	15	6	No	27
5	5	1	2	No	7
6	7	15	5	No	27
7	1	14	1	No	16
8	7	15	5	No	27
9	4	15	2	No	21
10	7	15	6	No	28
11	7	9	0	No	16
12	6	15	6	No	27
13	3	11	No	No	14
14	2	12	2	No	16
15	1	15	1	No	17
16	2	8	1	Yes	12

Table 6. Summary of model specifications

The complete estimation results for the best logit models are shown in appendix A. Direct road transportation (chain 1) is preferred in almost all commodity groups, holding all other attributes constant. The highest weight category is preferred relative the lowest for shipments of commodity group 2 (coal, petroleum, natural gas), 3, (ores and mining products) and 15 (round wood) whereas the opposite is true for the other groups. The coefficient on transport cost differs both between transport chains and commodity groups, showing that utility weight put on cost indeed is different across chains and commodities. The coefficient on transport time enters the utility function only for commodity group 16 (air freight).

Model prediction

We next turn to our analysis of the predictive performance of our models. Figure 3 shows the predicted and observed frequency for each of the 14 chain types. The commodity group is indicated at the top of every figure. For each group, the left window shows the predictions for observations not used in the estimation and the right window for in-sample observations. The scale of the y-axis is the same within commodity groups but differs between them. Blue bars show the frequencies observed in the CFS, red bars show the frequencies predicted by the logit model and green bars the predictions from the mock-deterministic model.

Chain 1 (direct road transportation) dominates the market for almost all commodities while chain 9 (truck-vessel-truck) is also frequently used for many commodities. In-sample frequencies are lower than out-of-sample frequencies since in-sample observations constitute some 20 percent of all observations within each commodity group.

In-sample predictions by the logit model are very close to observed levels. This is due to the inclusion of constants for chains and shipment size, which leads to close to perfect predictions of observed frequencies. In-sample predictions by the mock-deterministic model is also close

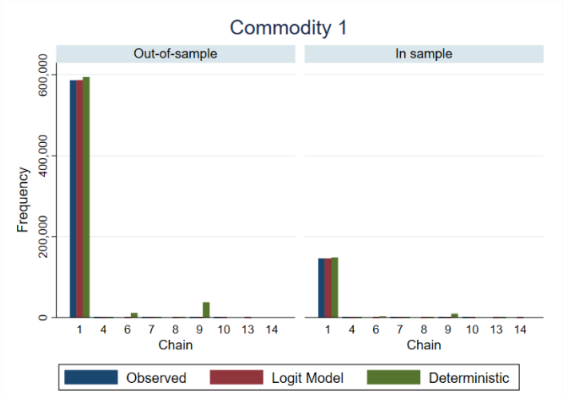
to observed levels, indicating that the alternative offering the lowest cost per tonne tends to be chosen in practice.

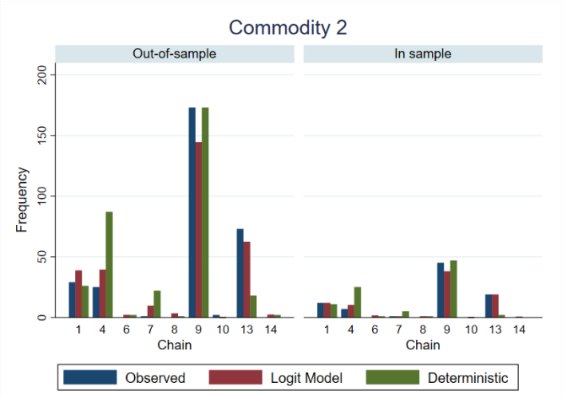
Out-of-sample predictions by the logit and mock-deterministic models are also close to observed levels but do not perform as well as the in-sample predictions. This is likely to be a result of modal shares being different in-sample compared to out-of-sample, perhaps due to sampling variation. The overall pattern is that both the logit model and the mock-deterministic model do well in predicting transport chain frequencies.

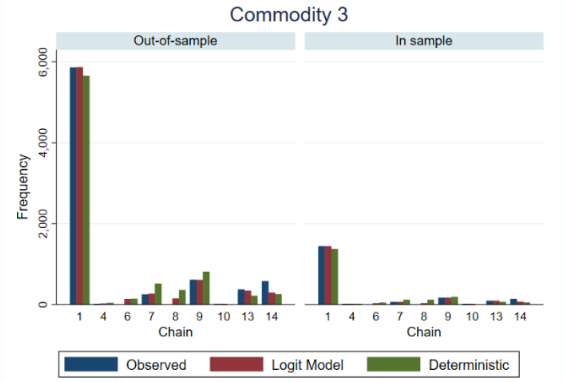
Figure 4 shows the results for predicted and observed frequencies calculated for each of the 16 shipment sizes. There is a clear variety in observed shipment sizes across commodity groups. A natural benchmark is weight class 10 or lower, which corresponds to below-truck-weight of 40 tonnes. For some commodity groups (4, 5, 11, 13 and 16), almost all shipments fall between 0-50 kg (weight class 1).[6] Groups 3, 6, 8, 9, 12 and 14 have more even distribution of sizes. Commodity group 2 stand out as having a shipment size distribution heavily skewed to the right.

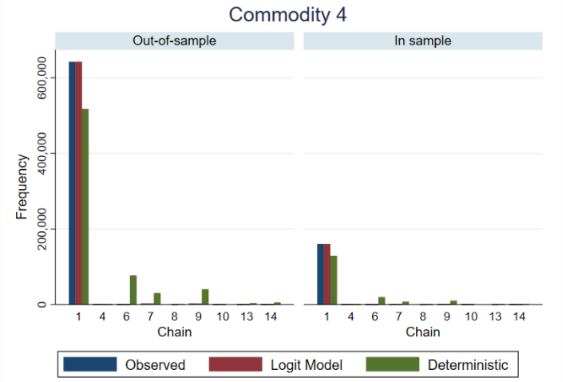
The performance of the logit and mock-deterministic models are similar for in-sample predictions compared to out-of-sample predictions, probably owing to the similarity in patterns between samples. However, there is a clear difference in the predictive performance of the logit model compared to the mock-deterministic model. The frequencies predicted by logit model are very close to the observed ones, again most likely because of the inclusion of shipment size constants. The mock-deterministic model tends to predict large shipment sizes and is therefore accurate for commodity group 2 where this is the case, and off for groups that tend to use lighter shipments (4, 5, 11, 13 and 16). This is consistent with transport cost per tonne on average being lower for larger shipment sizes.

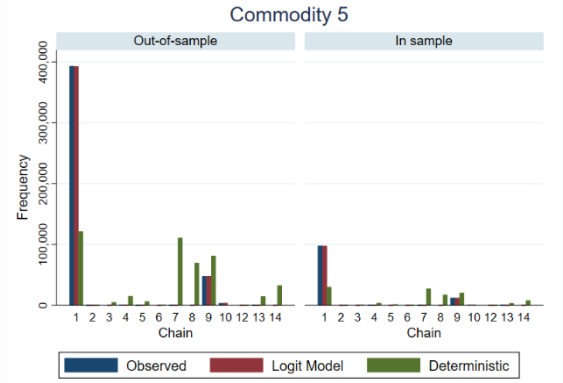
[6] Commodity group number and names are 1 (Products of agriculture, hunting, and forestry; fish and other fishing products. Not timber); 2 (Coal and lignite; crude petroleum and natural gas); 3 (Metal ores and other mining and quarrying products; peat); 4 (Food products, beverages and tobacco); 5 (Textiles and textile products; leather and leather products); 6 (Wood and products of wood and cork (except furniture); pulp, paper and paper products; printed matter and recorded media); 7 (Coke and refined petroleum products); 8 (Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel); 9 (Other non-metallic mineral products); 10 (Basic metals; fabricated metal products, except machinery and equipment); 11 (Machinery and equipment; medical, precision and optical instruments); 12 (Transport equipment); 13 (Furniture; other manufactured goods); 14 (Secondary raw materials; municipal wastes and other wastes); 15 (Timber); 16 (Air freight (fractions of some of the commodity groups). A list of commodities and their number is found in Appendix C.

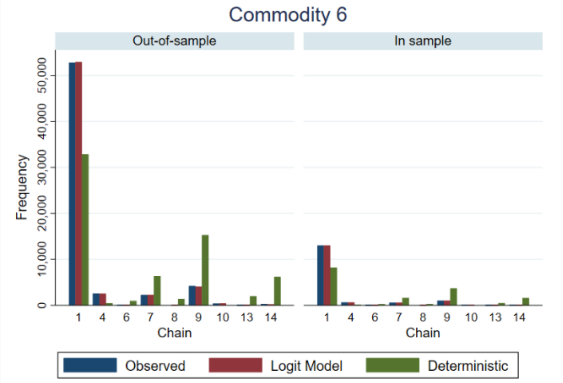

Commodity 1
Out-of-sample
In sample
Frequency
0
200,000
400,000
600,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic

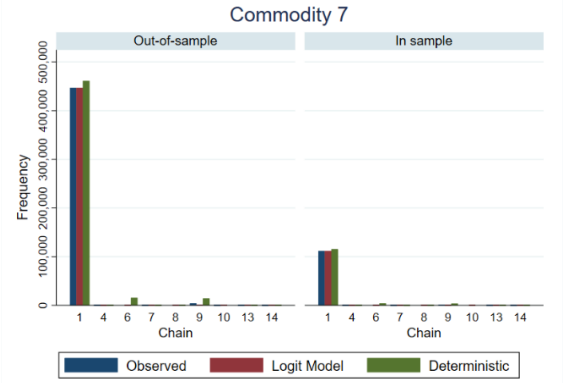

Commodity 2
Out-of-sample
In sample
Frequency
0
50
100
150
200
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic

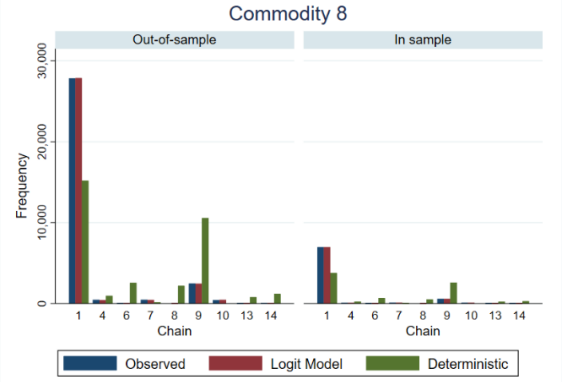

Commodity 3
Out-of-sample
In sample
Frequency
0
2,000
4,000
6,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic


Commodity 4
Out-of-sample
In sample
Frequency
0
200,000
400,000
600,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic


Commodity 5
Out-of-sample
In sample
Frequency
0
100,000
200,000
300,000
400,000
1 2 3 4 5 6 7 8 9 10 12 13 14
Chain
Observed
Logit Model
Deterministic


Commodity 6
Out-of-sample
In sample
Frequency
0
10,000
20,000
30,000
40,000
50,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic


Commodity 7
Out-of-sample
In sample
Frequency
0
100,000
200,000
300,000
400,000
500,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic


Commodity 8
Out-of-sample
In sample
Frequency
0
10,000
20,000
30,000
1 4 6 7 8 9 10 13 14
Chain
Observed
Logit Model
Deterministic

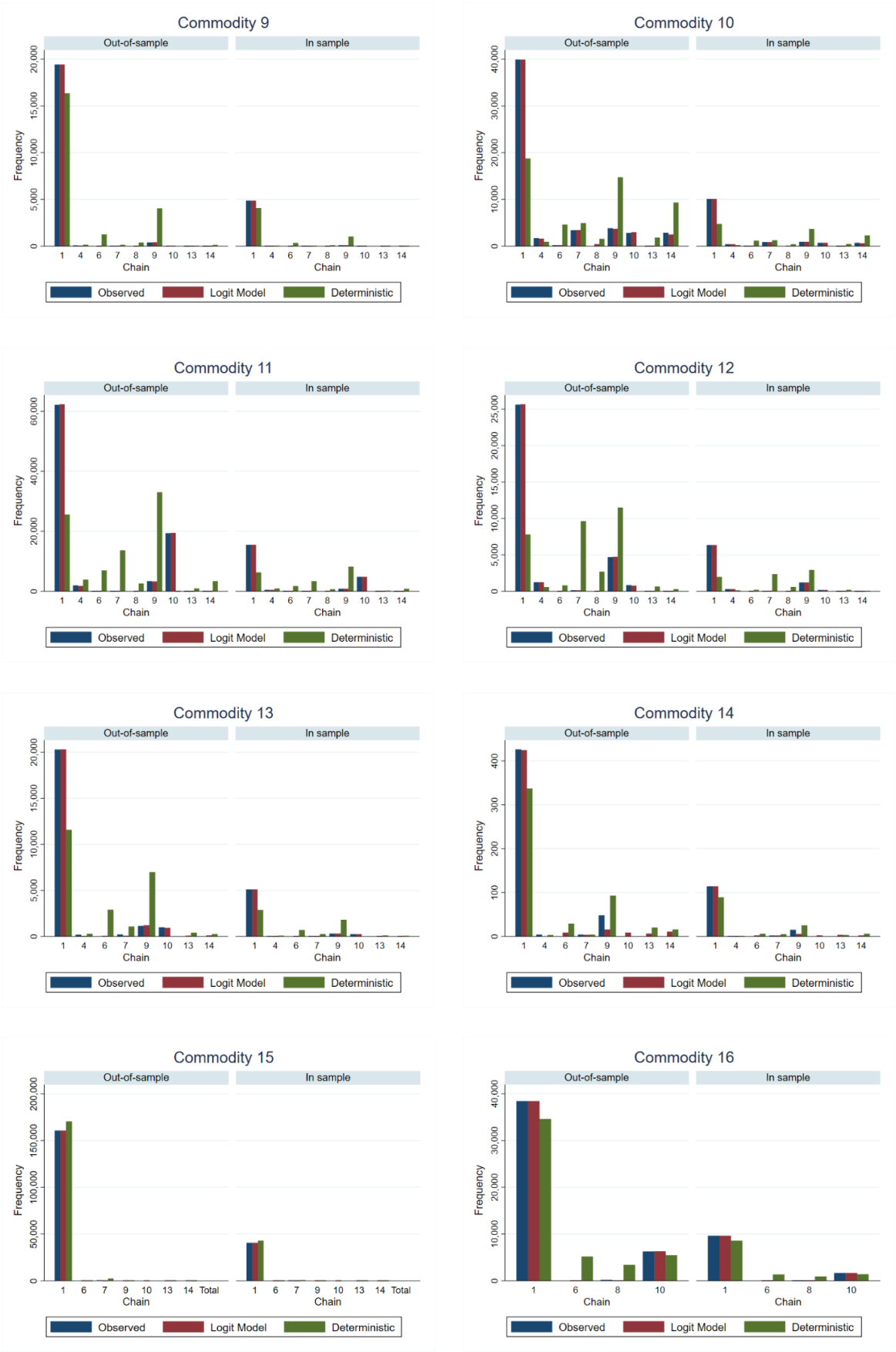
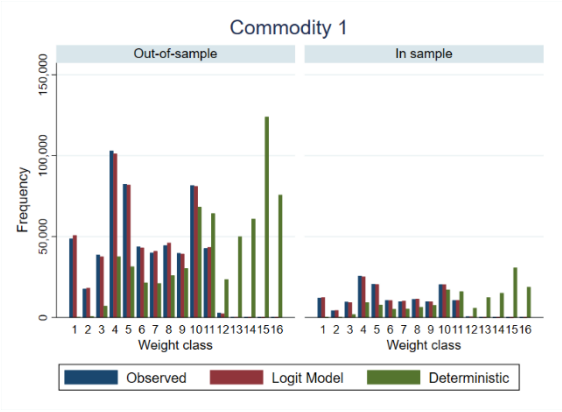
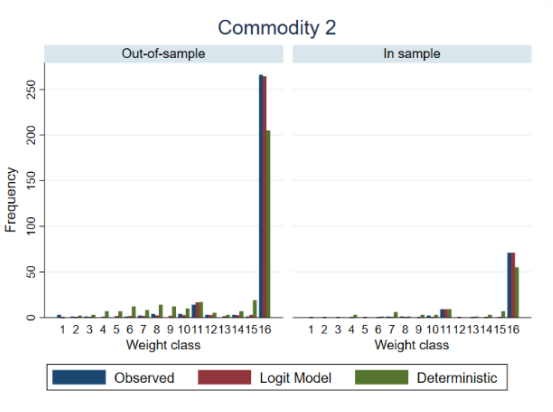
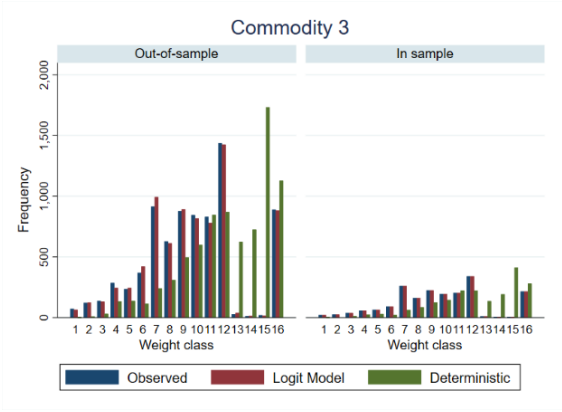
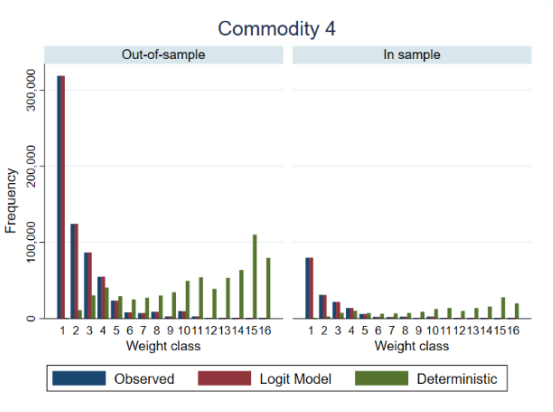


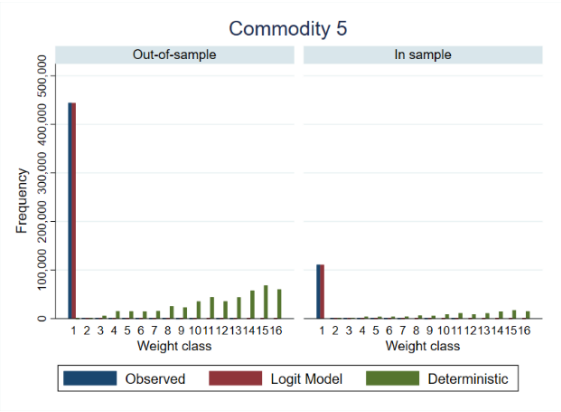
Figure 3. Observed and predicted frequency, by chain, commodity and sample group

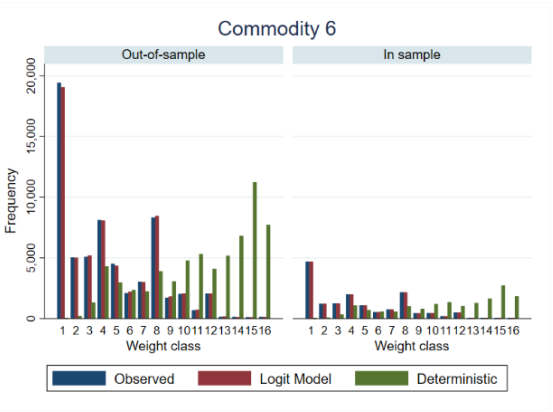

Commodity 1
Out-of-sample
In sample
Frequency
0
50,000
100,000
150,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic

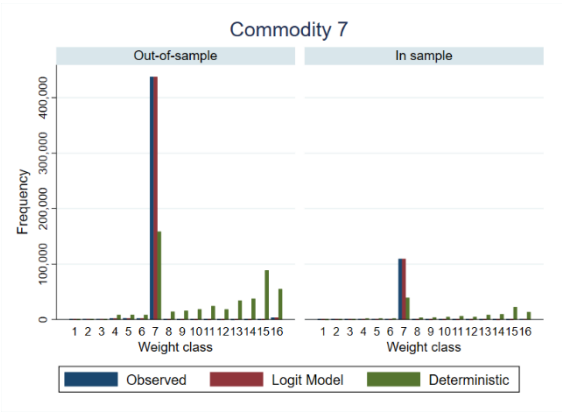

Commodity 2
Out-of-sample
In sample
Frequency
0
50
100
150
200
250
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic

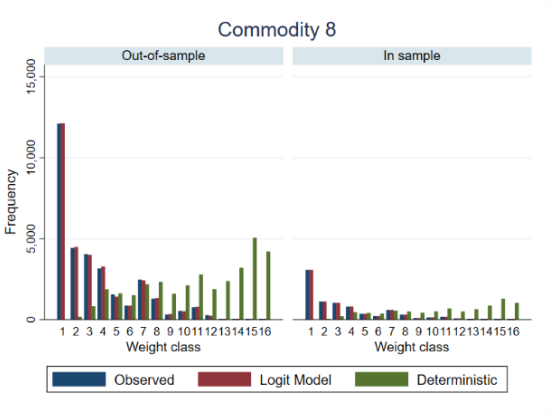

Commodity 3
Out-of-sample
In sample
Frequency
0
500
1,000
1,500
2,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic


Commodity 4
Out-of-sample
In sample
Frequency
0
100,000
200,000
300,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic


Commodity 5
Out-of-sample
In sample
Frequency
0
100,000
200,000
300,000
400,000
500,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic


Commodity 6
Out-of-sample
In sample
Frequency
0
5,000
10,000
15,000
20,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic


Commodity 7
Out-of-sample
In sample
Frequency
0
100,000
200,000
300,000
400,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic


Commodity 8
Out-of-sample
In sample
Frequency
0
5,000
10,000
15,000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Weight class
Observed
Logit Model
Deterministic

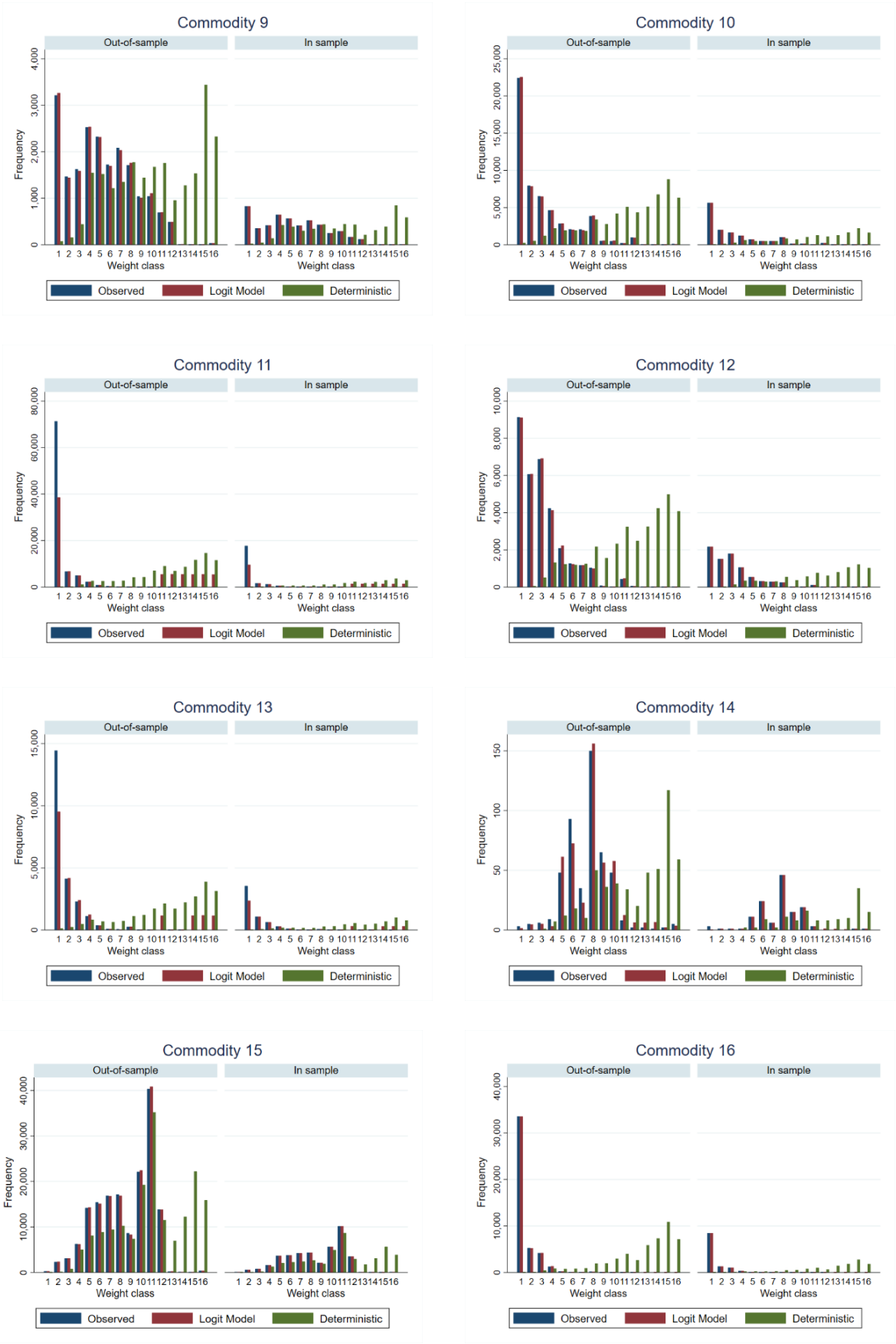


Figure 4. Observed and predicted frequency, by weight class, commodity and sample group

We now turn to our predictions of tonnes for each chain. Figure 6 shows the predicted and observed tonnes for each chain and in total. Blue bars show the tonne levels as observed in the CFS and red bars show the levels predicted by the logit model. The left window shows the predictions for out-of-sample observations and the right window for observations used in the estimation.

We have also calculated predictions based on our mock-deterministic model. However, this model severely over-predicts the tonne levels for the transport chains. This makes a visual analysis of the accuracy of our logit model very difficult. Figure 5 illustrates this point. The left graph contains the observed levels and the predictions both from the logit and mock-deterministic model (in green) for commodity group 8. It is difficult to assess the accuracy of the logit model. The right figure omits the prediction from the mock-deterministic which aids the visual interpretation of our results. In the remainder of the analysis we discuss our logit model results.

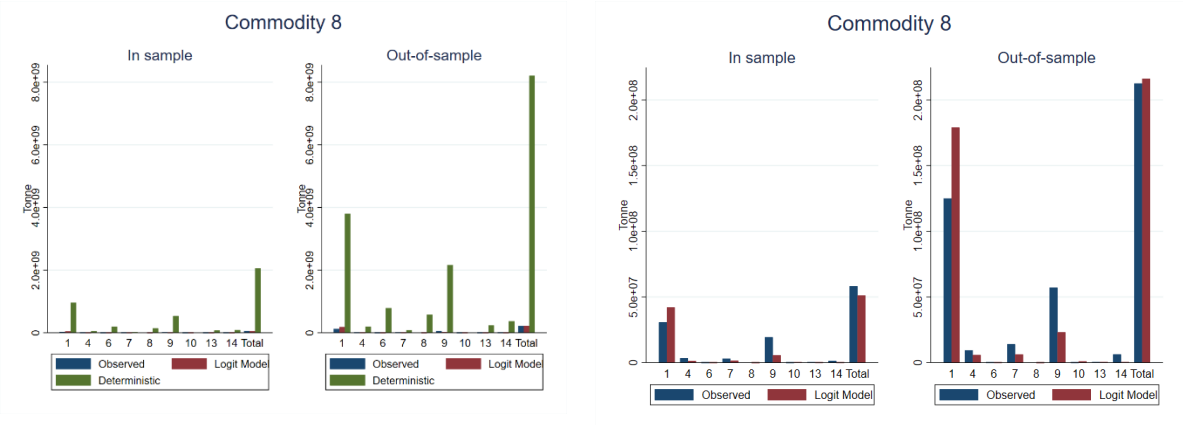
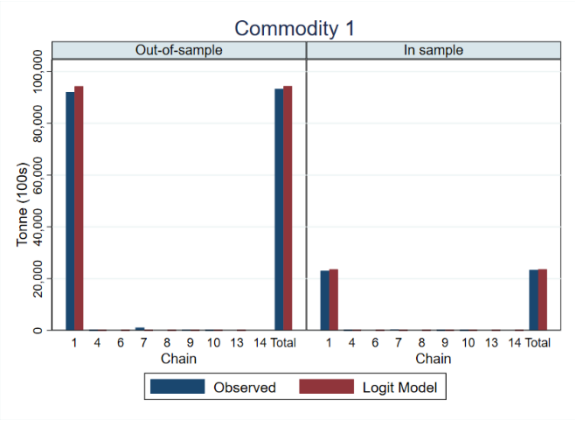


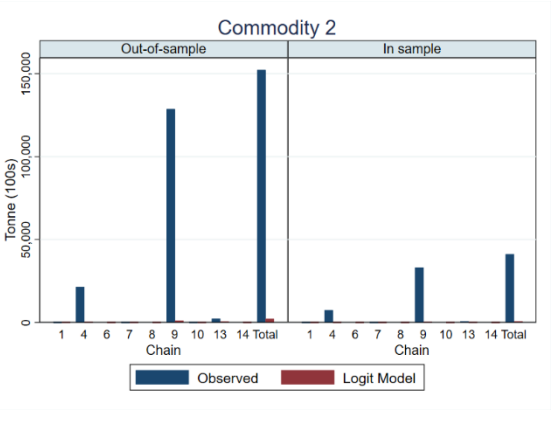
Figure 5. Predictions against observed levels, with and without deterministic model

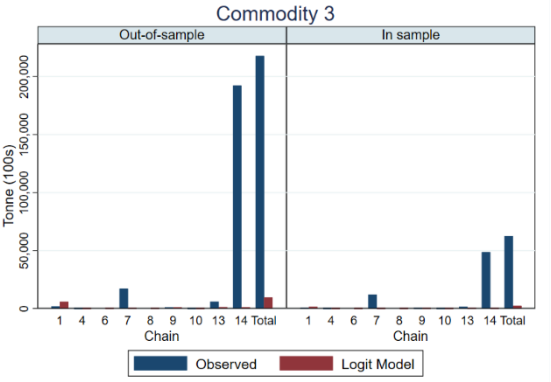
Just as for the frequencies, several commodity groups are dominated by one or two chain types. In most cases this is direct road transportation (chain type 1). Chain type 9 (truck-vessel-truck) also tends to have high shares of the market.

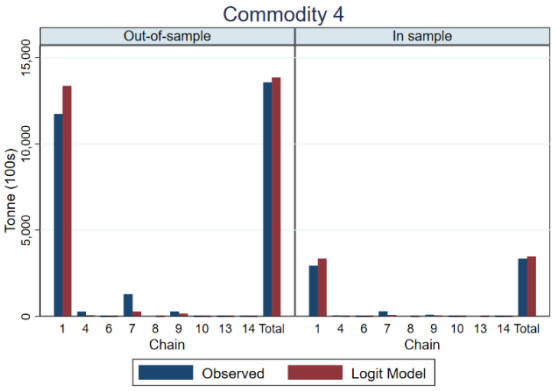
Prediction errors for individual chain types within commodity groups are similar for in-sample and out-of-sample observations. This is a by-product of modal shares being similar for observations used in estimation compared to the rest of the observations. The exception is commodity group 14 for which the modal shares are very different for the estimation sample compared to the rest of the observations. This makes the out-of-sample forecasts inaccurate for this group. The difference in modal shares is likely due to the very low sample size of commodity group 14 which makes sampling variability large.

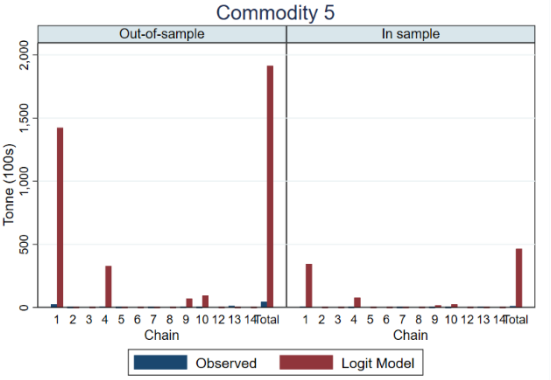
In-sample predictions are relatively accurate for commodity groups 1, 4, 6, 8, 9, 10, 12 and 14-16. The MNL model captures both the shares and the tonne levels of the chain types for these groups. Predictions for commodity groups 2, 3, 5 and 7 are inaccurate — the model overpredicts for commodity 5 and underpredicts for the other groups.

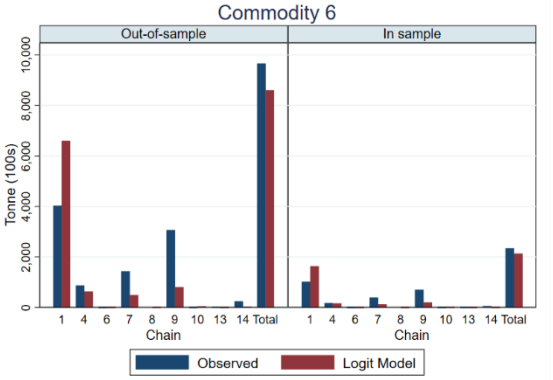

Commodity 1
Out-of-sample
In sample
Tonne (100s)
0
20,000
40,000
60,000
80,000
100,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model

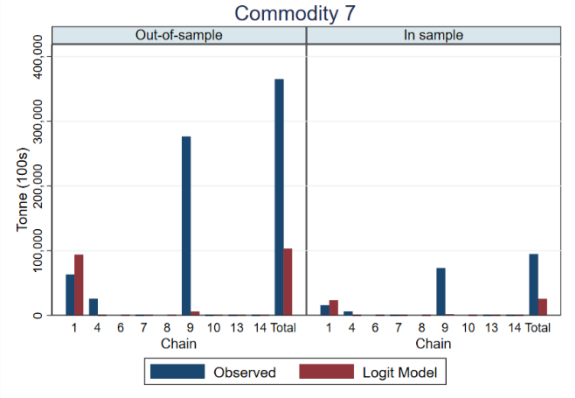

Commodity 2
Out-of-sample
In sample
Tonne (100s)
0
50,000
100,000
150,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model

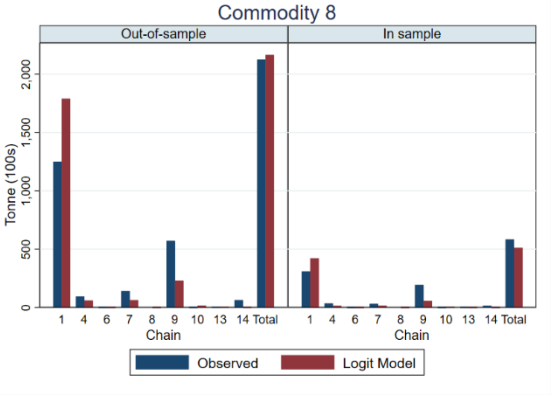

Commodity 3
Out-of-sample
In sample
Tonne (100s)
0
50,000
100,000
150,000
200,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model


Commodity 4
Out-of-sample
In sample
Tonne (100s)
0
5,000
10,000
15,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model


Commodity 5
Out-of-sample
In sample
Tonne (100s)
0
500
1,000
1,500
2,000
1 2 3 4 5 6 7 8 9 10 12 13 14 Total
Chain
Observed
Logit Model


Commodity 6
Out-of-sample
In sample
Tonne (100s)
0
2,000
4,000
6,000
8,000
10,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model


Commodity 7
Out-of-sample
In sample
Tonne (100s)
0
100,000
200,000
300,000
400,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model


Commodity 8
Out-of-sample
In sample
Tonne (100s)
0
500
1,000
1,500
2,000
1 4 6 7 8 9 10 13 14 Total
Chain
Observed
Logit Model

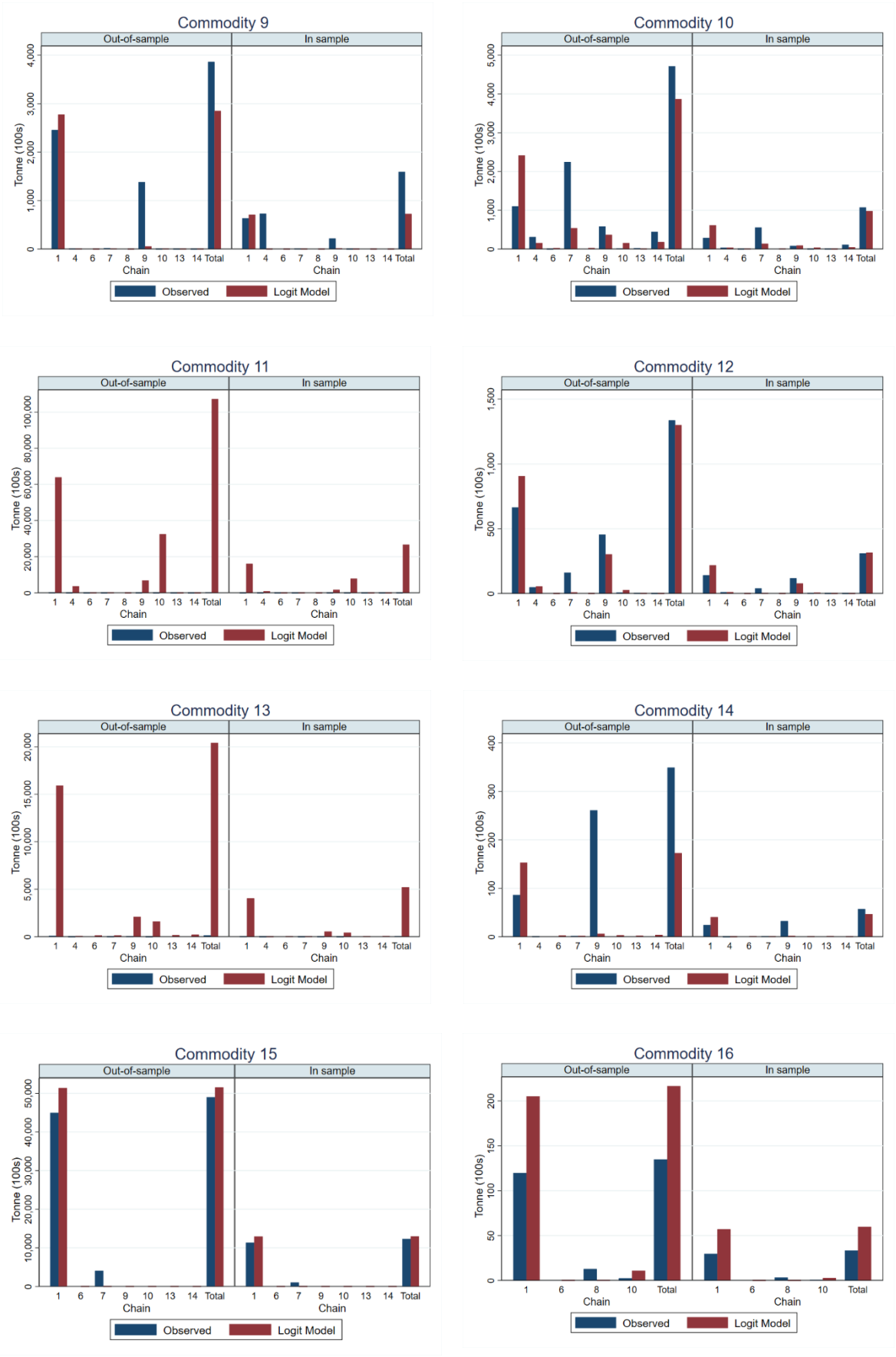


Figure 6. Observed and predicted tonnes by commodity and chain type

To summarize the accuracy of predicted tonnes by our logit model we calculate the root mean squared errors for in-sample and out-of-sample predictions. Figure 7 shows RMSE for each chain (i.e. calculated over commodities) to the left and for each commodity (i.e. calculated over chains) to the right.

The logit model predicts poorly for chains 1 (truck), 9 (truck-vessel-truck) and 14 (truck-rail-vessel-truck) and is off target for the total levels as well. The other chains have relatively low prediction deviations. Commodity groups 2, 3, 7 and 11 stand out as being difficult to make predictions for. These groups contain large deviations for chains 1, 9 and 14, as can be seen in figure 6. The inaccuracy for these categories depends partly on the fact that there are large volumes for transport chains in these commodity groups and that RMSE tends to increase with observed volumes.

One concern with our models is that the inclusion of the many constants for chains and shipment size leads to over-fitting. This would imply very good predictions in-sample but not out-of-sample. Figure 7 shows that in-sample predictions are more accurate than out-of-sample predictions but there is not a very large difference between in- and out- of sample predictions as measured by the RMSE.

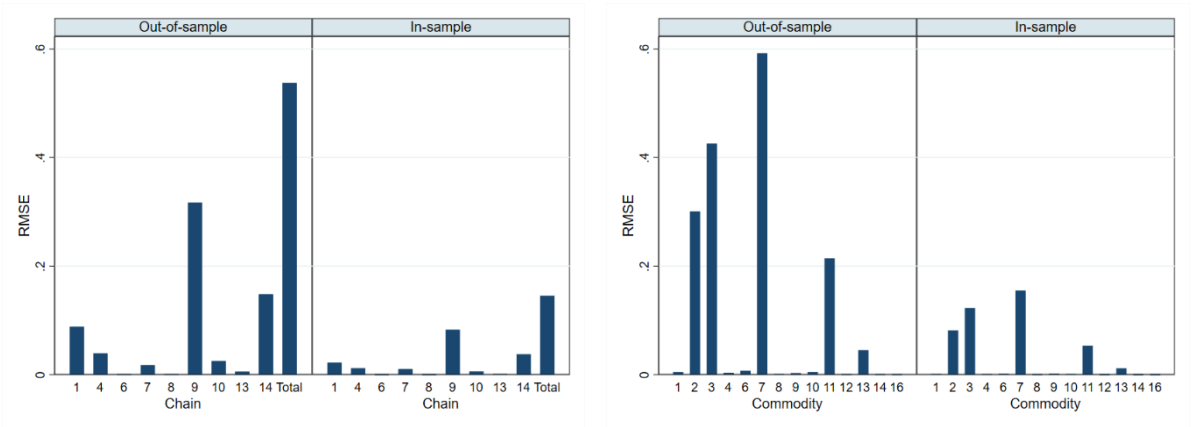


Figure 7. Root mean squared error by chain and commodity

Additional results

This section presents the results from a set of additional analyses made to investigate how changes in the main analysis impact the results.

Effect of size of choice set

The choice set in the proposed stochastic logistics module and our analysis consists of 14 chain types and 16 shipment sizes. These dimensions have been developed and analyzed in previous projects where they gave reasonable results (Windisch et al. 2010; Abate et al. 2014, 2016).

One concern is that aggregating transport chains into 14 types will mask important differences in choices patterns between the individual chains. One particular issue is that the choice of transportation solution for shipments using container transportation will be different from that for shipments in non-container cargos. This concern is partly motivated by the differences in price-elasticities for different train types in Samgods.

We test the size of choice set by comparing two different sets; the original choice set with 14 chains and an expanded choice set with 28 chains that differentiates between container and non-container options. Both sets have 16 shipment sizes. We compare the cost, time and distance variables generated by these two sets as well as the coefficients and predictions from logit models estimated on the sets.

We restrict our analysis to shipments of a single commodity group. We select commodity 6 (Wood, pulp, paper) as it has one of the highest shares of container usage (20%) in the CFS 2016. Differentiating between container and non-container choice is not likely to make a difference for shipments that only use one or the other load unit.

The method used to produce the expanded choice set follows that of the original one. Matching between observed choices in the CFS 2016 and the short list of chains is as before with the only exception that we now take into account the recorded load unit for the shipments in the CFS. A shipment is considered being transported in a container if it is registered as load unit 21-24 or 3 (containers, swap bodies and other exchangeable loading units) in the CFS. The matching rate is somewhat reduced, from 78,000 shipments to 75,000 shipments, when we also have to account for load unit. We again sample ten alternatives (including the chosen one) for each shipment.

Table 7 shows summary statistics for transport cost, time and distance produced under the different choice sets. Columns 1 and 2 show the statistics for all alternatives. The averages, standard deviation and min-max values are similar and sometimes identical for time and distance. Average costs are lower in the choice set that differentiates between container and non-container options. Columns 3 and 4 show statistics only for alternatives recorded as chosen in the CFS. These are even more similar than those for columns 1 and 2, suggesting that expanding the choice set only leads to more alternatives that have higher costs on average (and are not likely to be chosen).

		(1) Expanded set, all alt.	(2) Original set, all alt.	(3) Expanded set, chosen alt.	(4) Original set, chosen alt.
Cost (SEK/tonne)	Mean	3,840	5,363	1,903	1,802
	St. dev	19,799	24,603	15,646	15,358
	Min	66	66	69	69
	Max	640,711	640,711	578,559	578,559
Time (hrs)	Mean	43	41	39	45
	St. dev	147	142	141	164
	Min	0	0	0	0
	Max	2,128	2,084	1,843	1,212
Distance (KM)	Mean	1,509	1,486	1,363	1,377
	St. dev	3,642	3,367	3,897	3,836
	Min	1	1	1	1
	Max	26,426	26,426	26,070	25,740

Table 7. Comparison of choice sets for commodity 6

We estimate logit models on each of the choice sets using identical model specifications and compare the coefficients. Ideally, we would like the coefficients to be identical so that our selected choice set in the original analysis does not affect the coefficients entering the logistics module.

Table 8 shows the result. Column 1 displays the coefficients (and column 2 the standard errors) for the sample using the expanded choice set. Columns 3 shows the coefficients (and column 4 the standard errors) based on the original choice set. Column 5 presents the difference in coefficients for each variable and column 6 tests if they have the same sign.

	Expanded choice set		Original choice set		Difference (1)-(3)	Sign (1) = Sign (3)
	(1)	(2)	(3)	(4)	(5)	(6)
	Coeff.	S.E.	Coeff.	S.E.		
Chain 1	7.98***	(0.16)	7.61***	(0.14)	0.37	Yes
Chain 4	4.41***	(0.19)	6.36***	(0.30)	-1.95	Yes
Chain 6	0.45	(0.30)	-0.42	(0.29)	0.87	No
Chain 7	4.39***	(0.16)	4.18***	(0.15)	0.21	Yes
Chain 9	3.22***	(0.16)	4.59***	(0.15)	-1.37	Yes
Chain 10	0.94***	(0.16)	0.60***	(0.16)	0.34	Yes
Chain 13	-0.59	(1.06)	-0.75	(0.96)	0.16	Yes
Size 2	-1.95***	(0.053)	-1.78***	(0.052)	-0.17	Yes
Size 3	-2.11***	(0.053)	-1.83***	(0.053)	-0.28	Yes
Size 4	-1.68***	(0.050)	-1.53***	(0.049)	-0.15	Yes
Size 5	-2.28***	(0.055)	-2.05***	(0.055)	-0.23	Yes
Size 6	-2.88***	(0.065)	-2.57***	(0.064)	-0.31	Yes
Size 7	-2.60***	(0.060)	-2.27***	(0.058)	-0.33	Yes
Size 8	-1.79***	(0.053)	-1.49***	(0.052)	-0.3	Yes
Size 9	-3.26***	(0.072)	-2.99***	(0.071)	-0.27	Yes
Size 10	-3.18***	(0.071)	-2.91***	(0.070)	-0.27	Yes
Size 11	-4.26***	(0.095)	-3.99***	(0.094)	-0.27	Yes
Size 12	-3.01***	(0.069)	-2.69***	(0.067)	-0.32	Yes
Size 13	-5.64***	(0.16)	-5.39***	(0.17)	-0.25	Yes
Size 14	-6.11***	(0.19)	-5.83***	(0.20)	-0.28	Yes
Size 15	-6.22***	(0.21)	-6.06***	(0.21)	-0.16	Yes
Size 16	-6.15***	(0.20)	-5.87***	(0.20)	-0.28	Yes
Chain 1 x cost	-0.0012***	(0.000035)	-0.0012***	(0.000036)	0	Yes
Chain 4 x cost	-0.00040***	(0.000038)	-0.0014***	(0.00011)	0.001	Yes
Chain 7x cost	-0.00049***	(0.000046)	-0.00059***	(0.000054)	0.0001	Yes
Chain 9 x cost	-0.00048***	(0.000052)	-0.00099***	(0.000068)	0.00051	Yes
Chain 13 x cost	-0.00058	(0.00068)	-0.00081	(0.00077)	0.00023	Yes
N	148,810				154,600	
LL	-12830.3				-12414.6	
Parameters	27				27	

Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8. Coefficient comparison for commodity 6

The model shows that direct road transportation (chain 1) and road in combination with vessel (chain 4 and 9), rail (chain 7) or air (chain 10) has a statistically significant positive impact on utility. The smallest shipment size tends to be preferred over larger. Reassuringly, for almost every variable, the sign of the coefficient from one estimation is the same as the one from the other estimation. Magnitudes are also similar across coefficients suggesting that whether the choice set is expanded or not matters little for the final coefficients. The cost coefficients for the expanded choice set are somewhat larger in size, possibly as a response to lower average costs.

For the implementation of a stochastic logistics module, we are interested in the predictive performance of these two sets of coefficients. We therefore use the same prediction approach as in the main analysis for these two sets. The only difference is that we now make predictions over the expanded set of choices. Graphical results are shown in figure 8 and suggest that the predictive performance is improved when using the expanded choice set.

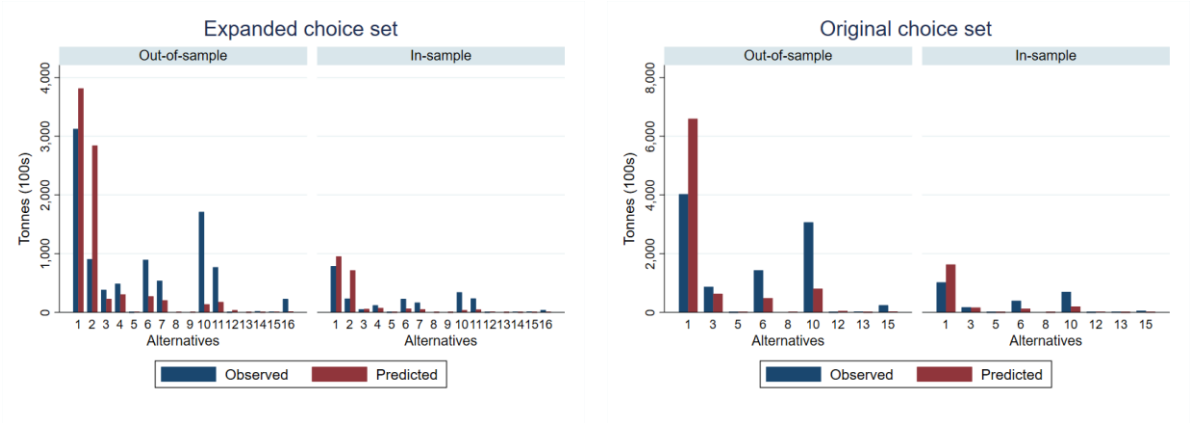


Figure 8. Comparing predictions for expanded and original choice sets

Effect of changing weight class interval

The choice set in our analysis consists of 16 shipment sizes. We test the effect of changing weight classes by first defining an alternative interval based on the existing one. We divide the weights into multiples of a truck-load (40 tonnes). This classification follows existing truck capacity limits and is easily compared to the original weight class interval.

We test the choice of weight class interval by comparing two different sets; the original choice set with 16 shipment size and the reduced set with 5 shipment sizes. Both sets have 14 chain types. Table 9 summarizes these sets. We compare the cost, time and distance variables generated by these two sets as well as the coefficients and predictions from logit models estimated on the sets. We again focus solely on commodity group 6. Matching rate is slightly higher with the broader weight class interval, 99.1% compared to 97% for the original weight class interval.

	Original weight class interval		Reduced weight class interval
1	0-50 kg	1	0-10 000 kg
2	51-200 kg		
3	201-800 kg		
4	801-3000 kg		
5	3001-7500 kg		
6	7501-12500 kg	2	10 0001 kg-20 000 kg
7	12501-20000 kg		
8	20001-30000 kg	3	20 001-40 000 kg
9	30001-35000 kg		
10	35001-40000 kg		
11	40001-45000 kg	4	40 001-400 000 kg
12	45001-100000 kg		
13	100001-200000 kg		
14	200001-400000 kg		
15	400001-800000 kg	5	400 001+ kg
16	800001+ kg		

Table 9. Overview of original and reduced weight class interval

Table 10 shows summary statistics for transport cost, time and distance produced under the different sets. Columns 1 and 2 shows the statistics for all alternatives. Transport time and distance are virtually unaffected by the merging of interval from 16 to 5 categories. But average and maximum costs are lower with this interval. The same holds true for chosen alternatives only in columns 3 and 4. One explanation is that the composition of chains (and thus transport time and distance) remains the same, while the shipment sizes for the available alternatives increase and bring down average cost per tonne. Altogether, this indicates that a coarser interval produces lower transport cost overall and for chosen alternatives.

		(1) Original interval, all alternatives	(2) Reduced interval, all alternatives	(3) Original interval, chosen alt.	(4) Reduced interval, chosen alt.
Cost (SEK/tonne)	Mean	4,922	3,437	1,802	1,181
	St. dev	19,693	13,799	15,358	10,829
	Min	66	66	69	68
	Max	640,711	561,814	578,559	499,662
Time (hrs)	Mean	30	30	45	44
	St. dev	98	96	164	161
	Min	0	0	0	0
	Max	7,907	1,212	1,212	1,212
Distance (KM)	Mean	1,241	1,240	1,377	1,368
	St. dev	2,355	2,355	3,836	3,766
	Min	1	1	1	1
	Max	26,426	26,426	25,740	25,740
Observations		8,668,576	2,708,930	78,019	81,022

Table 10. Cost, time and distance by weight class interval

We estimate logit models on each set using identical model specifications and compare the coefficients. Table 11 shows the results. The coefficients exhibit the same patterns overall, although some of the cost coefficients based on the reduced interval have a positive sign. We again perform predictions for the commodity group based on the updated weight class

interval. The results are shown in figure 9. The reduced weight class interval overpredicts the road transportation. These shipments are very small but predicted as too large. This shows the need to keep a more fine-grained weight interval in the logit application

	Reduced interval		Original interval		Difference (1)-(3)	Sign (1) = Sign (3)
	(1)	(2)	(3)	(4)	(5)	(6)
	Coeff.	S.E.	Coeff.	S.E.		
Chain 1	7.37***	(0.071)	7.91***	(0.072)	-0.54	Yes
Chain 4	3.10***	(0.14)	7.50***	(0.15)	-4.4	Yes
Chain 6	0.32**	(0.12)	0.20	(0.12)	0.12	Yes
Chain 7	4.16***	(0.089)	4.64***	(0.075)	-0.48	Yes
Chain 9	2.07***	(0.082)	4.07***	(0.078)	-2	Yes
Chain 10	0.50***	(0.071)	0.51***	(0.074)	-0.01	Yes
Chain 13	-2.29***	(0.59)	-0.78	(0.40)	-1.51	Yes
Size 2			-2.04***	(0.030)		
Size 3			-2.25***	(0.031)		
Size 4			-1.83***	(0.029)		
Size 5			-2.41***	(0.032)		
Weight (10-20t)	-2.27***	(0.022)				
Size 6			-3.18***	(0.038)		
Size 7			-2.68***	(0.034)		
Weight (20-40t)	-1.14***	(0.016)				
Size 8			-1.82***	(0.031)		
Size 9			-3.52***	(0.041)		
Size 10			-3.48***	(0.040)		
Weight (40-400t)	-2.50***	(0.024)				
Size 11			-4.49***	(0.053)		
Size 12			-2.93***	(0.038)		
Size 13			-5.92***	(0.089)		
Size 14			-5.91***	(0.091)		
Weight (+400t)	-5.10***	(0.066)				
Size 15			-6.26***	(0.10)		
Size 16			-6.00***	(0.091)		
Chain 1 x cost	-0.00094***	(0.000025)	-0.0011***	(0.000019)	0.00016	Yes
Chain 4 x cost	0.00025***	(0.000076)	-0.0018***	(0.000053)	0.00205	No
Chain 7x cost	-0.00043***	(0.000088)	-0.00064***	(0.000026)	0.00021	Yes
Chain 9 x cost	0.00094***	(0.000076)	-0.0013***	(0.000049)	0.00224	No
Chain 13 x cost	0.00043	(0.00052)	-0.00053*	(0.00023)	0.00096	No
N	663080		663080			
ll	-35637.7		-39873.8			
k	16		27			

Standard errors in parentheses * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11. Coefficient comparison by weight class interval

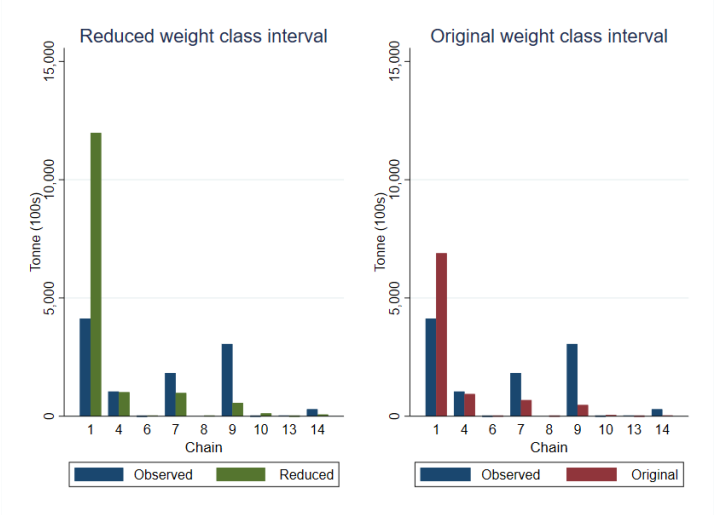


Figure 9. Comparison of weight class interval

Effect of changing weight used to predict shipment size

Our prediction of tonnes per transport chain has so far been based on the formula

$$\hat{Q}_i(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P_i(t,s) \times \bar{s}$$

where $\bar{s}$ is the average weight in interval s, s=1, 2, …, 16. Here we investigate alternative ways a weight class interval is coded as a weight. Specifically, we add a factor F to the formula used to predict tonnes levels:

$$\hat{Q}(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P_i(t,s) \times (\bar{s} \times F)$$

We focus on commodity groups 2,3 and 5 since the tonnes were poorly predicted for these groups. We set the factor F to obtain more accurate levels. We use a scaling factor of 80 for commodity 2, 25 for commodity group 3 and 1/320 for commodity group 5. For example, this means that for commodity 2, the predicted weight associated with interval 20-30 kg is rescaled from 25 kg to 2000 kg.

The effect of rescaling the weights are shown in figure 10. The original predictions are shown on the left side and the predictions based on the adjusted weight interval on the right side. The adjusted predictions show a high degree of consistency with the observed levels. This shows that the inaccuracy of the initial predictions for commodity groups 2,3 and 5 is not due to incorrect modal shares, these are fairly accurately predicted, but that the shipment sizes are poorly predicted.

This result shows that it is more difficult to predict how much firms will ship than which chain they will use. But it is not obvious that the weight class interval used in the analysis is inappropriate and should be adjusted since the existing one works well for the other commodity groups.

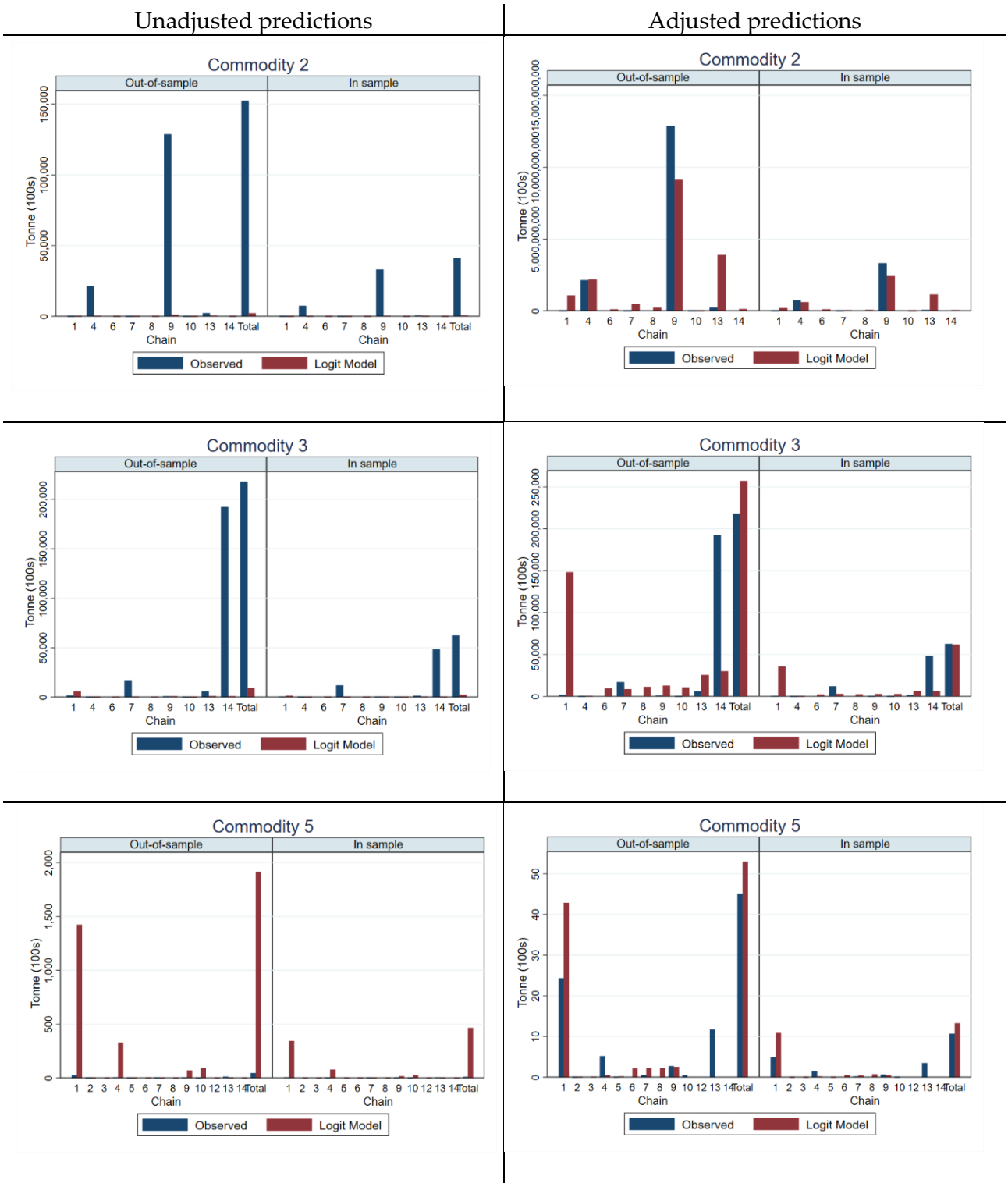


Figure 10. Comparison of original and adjusted predictions by commodity

Nested Logit Model Results

The multinomial logit model that was used in the main analysis implies that an improvement for one alternative in, say, transport cost, will reduce the probabilities for all the other alternative by the same percentage. To be specific, for two options j and k, the ratio of their probabilities is

$$\frac{P(j)}{P(k)} = \frac{e^{(X_j\widehat{\boldsymbol{\beta}})}/\sum_l e^{(X_l\widehat{\boldsymbol{\beta}})}}{e^{(X_k\widehat{\boldsymbol{\beta}})}/\sum_{\boldsymbol{l}} e^{(X_l\widehat{\boldsymbol{\beta}})}} = \frac{e^{(X_j\widehat{\boldsymbol{\beta}})}}{e^{(X_k\widehat{\boldsymbol{\beta}})}}$$

The relative probability of choosing j over k does not depend on any other alternatives. The nested logit relaxes this assumption and may therefore produce more realistic choice patterns. The model partitions the alternatives into nests in a way that i) the ratio of choice probabilities within a nest are independent of the attributes or existence of all other alternatives and ii) the ratio of choice probabilities for alternatives in different nests depend on the attributes of other alternatives in the two nests.

The nested logit model implies that the probability of shipment i is being transported by chain t and shipment size s within nest N is given by:

$$P_{iN}(t,s) = (P_i(t,s)|N) \times P_i(N)$$

where $(P_i(t,s)|N)$ is the probability of choosing alternative t,s, given that nest N is chosen and $P_i(N)$ is the probability of choosing an alternative in nest N. Because the nested logit is a generalization of the multinomial logit, it will mechanically fit any set of estimation data at least as good as the MNL in terms of log-likelihood. However, it need not predict as well as the logit model, e.g. in the case of over-fitting the model.

We estimate the nested logit model above and compare the coefficients with the baseline MNL. We select to estimate the model on commodity group 14 and estimate the model on all observations (as opposed to a 20% sub-sample, like before). We specify one nest for direct road transport (using any of the shipment sizes) and another for using multimodal transportation. We let the probability of using direct transportation, $P_i(N)$, depend on the ratio of shipment value to shipment weight. The probability of using transport chain t,s, given the choice of nest, $(P_i(t,s)|N)$, is specified as a function of chain type and size dummies.

Table 13 compares the coefficient estimates from our baseline MNL in column 1 and nested logit model in column 2. The coefficients determining the choice of shipment size and chain type are almost always of identical sign but vary in size. The nested logit is preferred according to the log-likelihood, shown in the bottom panel. But the information criteria, AIC and BIC, give inconclusive evidence as to which model is preferred.

Figure 11 shows the observed tonnes and the predicted tonnes by the MNL model and nested logit respectively. Neither model is able to predict the large amount of freight moved by chain type 9 (truck-vessel-truck), although the coefficient on this constant is positive and sizeable in both models. The nested logit does better in predicting the levels transported by chain type 1 (truck) — it deviates from the observed levels by 36 % compared to the 59 % for the baseline MNL. The nested model predicts worse than the baseline MNL for the other two chain types (truck-vessel and truck-rail-truck) but the tonnes lifted by these chains are very small and the deviations are not significant in terms of absolute levels.
Overall, this analysis illustrates that a nested logit specification can predict at least as well as the baseline MNL. It does come with greater computational requirements and the simplicity of the MNL is a strong argument for using it in our analysis.

	MNL		Nested		Difference (1)-(3)	Sign (1) = Sign (3)
$(P_i(t,s)\|N)$	(1)	(2)	(3)	(4)	(5)	(6)
	Coeff.	S.E.	Coeff.	S.E.		
Chain 1	21.65	(0.01)	27.87	(0.04)	-6,22	Yes
Chain 4	21.86	(0.00)	22.40	(0.01)	-0,54	Yes
Chain 6	0.0951	(0.00)	-3.647	(-0.00)	3,7421	No
Chain 7	17.74	(0.01)	22.81	(0.03)	-5,07	Yes
Chain 9	20.09	(0.01)	25.89	(0.03)	-5,8	Yes
Chain 10	-0.0102	(-0.00)	-4.759	(-0.00)	4,7488	Yes
Chain 13	0.0387	(0.00)	-3.968	(-0.00)	4,0067	No
Size 2	0.299	(0.55)	0.692	(1.55)	-0,393	Yes
Size 3	0.453	(0.88)	0.682	(1.55)	-0,229	Yes
Size 4	0.545	(1.17)	0.728	(1.76)	-0,183	Yes
Size 5	2.831***	(8.51)	2.527***	(5.08)	0,304	Yes
Size 6	3.601***	(11.21)	3.052***	(5.24)	0,549	Yes
Size 7	1.814***	(5.20)	1.735***	(4.14)	0,079	Yes
Size 8	3.271***	(10.51)	2.755***	(5.01)	0,516	Yes
Size 9	2.435***	(7.48)	2.075***	(4.47)	0,36	Yes
Size 10	2.482***	(7.45)	2.089***	(4.44)	0,393	Yes
Size 11	0.750	(1.68)	0.572	(1.29)	0,178	Yes
Size 12	-0.957	(-1.23)	-0.423	(-0.68)	-0,534	Yes
Size 13	-0.674	(-0.87)	-0.852	(-0.92)	0,178	Yes
Size 14	-1.315	(-1.25)	-1.230	(-1.10)	-0,085	Yes
$P_i(N)$						
Value/Weight			0.00225	(0.17)		
τ_{direct}			0.688***	(5.10)		
τ_{chain}			1.599***	(5.14)		
Observations	6140		6140			
Lok-likelihood	-511.8		-501.1			
AIC	1063.6		1048.2			
BIC	1198.1		1202.8			

Table 12. MNL and nested logit results

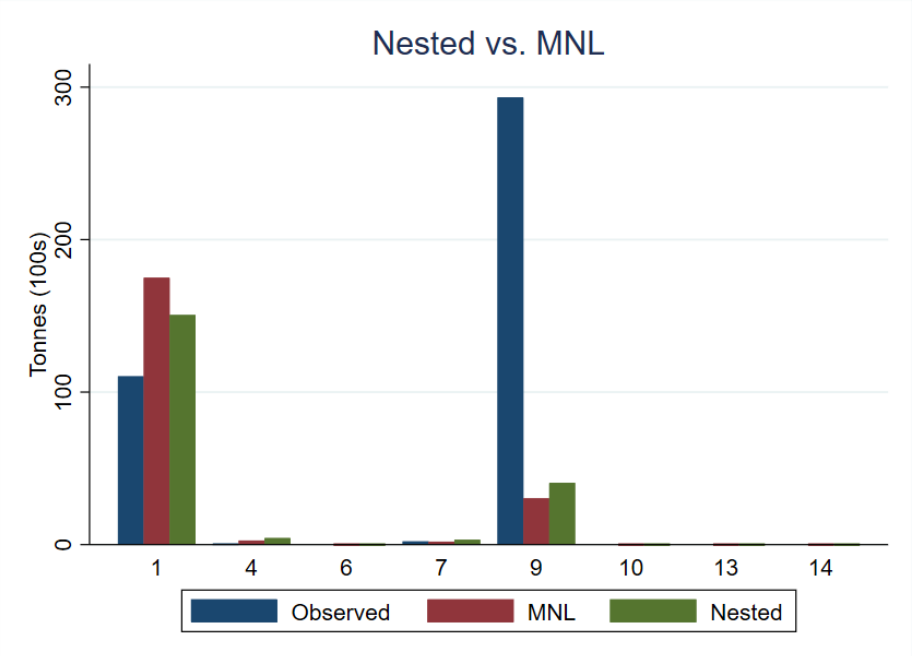


Figure 11. MNL and nested logit predictions against observed

New cost and time data cost data

A relevant question is how useful the coefficients produced in the main analysis are if changes are made to the process that generates the data. Cost, time and distance for the choice alternatives were clearly generated using specific values on the input parameters. Transport cost could therefore change if input values are updated or if alterations are made to the LOS-matrices. The estimated coefficients should ideally predict well even for smaller changes to the level of the cost variable.

We test the robustness of our estimated coefficients to changes in input data in the following manner. We first create an alternative cost function that produces alternative-specific transport costs that are similar but not identical to those used in the main analysis. We then use the estimated coefficients from the main analysis to make two sets of predictions — one based on the original cost variable and another based on the alternative cost variable. We end by comparing the predictions from the two variable sets against the observed levels. Large differences are concerning since it suggests that future changes in variables heavily influence the predictive accuracy of the coefficients.

Our alternative cost function for each alternative is given by

$$Cost_{ts}^{A} = Distance_{ts} \times \alpha_t + N_{ts} \times \beta_s$$

where $Distance_{ts}$ is equal to the kilometers between origin and destination using chain ts as calculated in Samgods. α_t is a cost-factor for chain t, N_t is the number of transshipment points for the chain and β_s is the assumed unit cost per transshipment for shipment size s. We set the cost-factor α_t to be the ratio of average cost to average distance for each of the four modes, based on the cost that were generated for the main analysis. (These are 3.8 for road, 2.5 for rail, 3 for sea and 5.8 for air transportation). If more than one mode is used in a chain, the cost-factor is a weighted average of the mode-specific factors. This means that a chain that uses all modes would have $\alpha_t = (3.8 + 2.5 + 3 + 5.8)/4$. We set the unit cost per transshipment β_s proportional to the shipment size s by letting it equal $\beta_s = 40s$. This value produces alternative costs that are on average similar to the original values.

We restrict our analysis to commodity group 10. Table 14 shows summary statistics for the alternative cost and original cost. Columns 1 and 2 show the statistics for all alternatives. These are indeed relatively similar on average but there is more variation in the original cost set. Columns 3 and 4 show the statistics only for alternatives that were chosen in the CFS.

		(1) Original interval, all alternatives	(2) Reduced interval, all alternatives	(3) Original interval, chosen alt.	(4) Reduced interval, chosen alt.
Cost (SEK/tonne)	Mean	8,123	9,685	7,256	15,743
	St. dev	14,467	41,984	13, 954.3	69,967.9
	Median	9.5	12.6	9.5	13.4
	Min	3,58	760	2,087	753.6
	Max	150,657	642,483	113,576	642,483
Observations		685,330	685,330	68,533	68,533

Table 13. Cost by weight interval

In the next step we apply the coefficients from the main analysis to the alternative cost variable to determine whether the prediction differs. Let C^o be the original cost variable and C^A the alternative cost variable. Using the coefficient estimated in the main analysis, β and γ, we now make two sets of predictions:

$$\mathrm{P}^O(t,s) = \frac{\exp(\beta C^O + \gamma' x_{ts})}{\sum_j \exp(\beta C^O + \gamma' x_j)}$$

$$\mathrm{P}^A(t,s) = \frac{\exp(\beta C^A + \gamma' x_{ts})}{\sum_j \exp(\beta C^A + \gamma' x_j)}$$

where the only component that differ between the predictions is the cost variables. As before, we predict tonnes by summing over observations and shipment sizes:

$$\hat{Q}^O(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P^O(t,s) \times \bar{s}$$

$$\hat{Q}^A(t) = \sum_{i=1}^{N} \sum_{s=1}^{16} P^A(t,s) \times \bar{s}$$

Figure 12 shows the results from the predictions. The blue bar shows the observed tonnes for each of the chain types, $Q(t)$, the red bars are the predictions based on the original costs. $\hat{Q}^O(t)$, and the green bars the predictions using the alterantive costs $\hat{Q}^A(t)$. For out-of-sample observations, the predictions based on alterantive costs are closer to observed levels for chains 1, 7, 9 and 14 while the predictions using the original costs performs better for the other chains. Predictions using the alternative costs are thus not systematically worse (or better) than the original, giving some support to the robustness of our estimated coefficients to changes in data inputs.

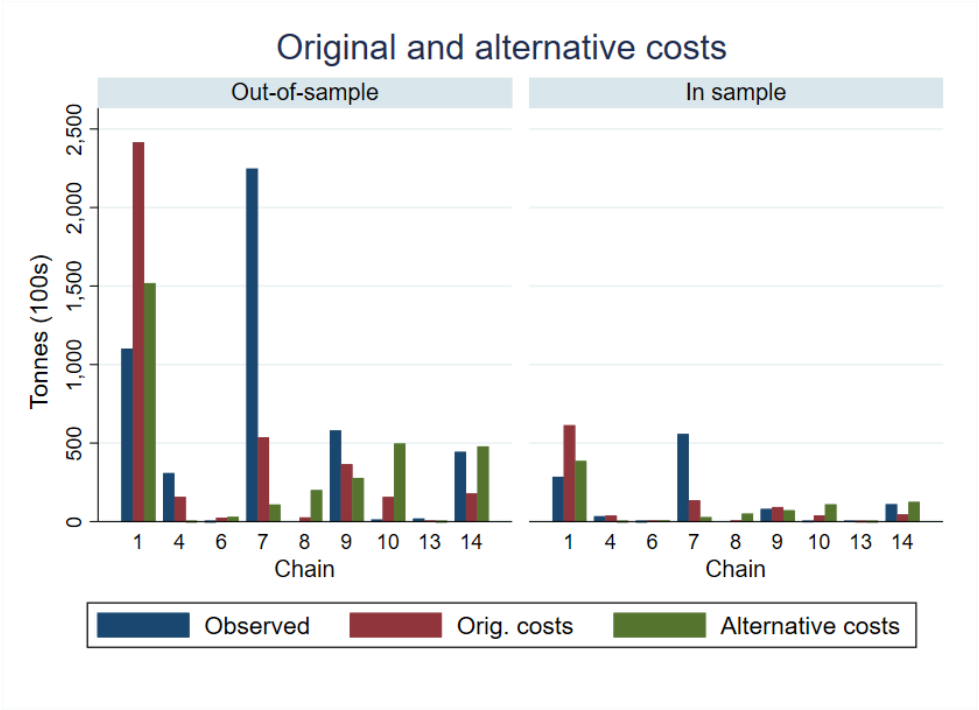


Figure 12. Predictions based on original and alternative cost

Conclusions

As part of the Swedish Transport Administration's plan to implement a stochastic logistics module in the national freight transport model Samgods, the aim of this study has been to estimate and evaluate discrete choice models, produce coefficients estimate that can be used in the implementation of a stochastic logistics module and analyze how such implementation should be conducted.

Predictive performance

The mock-deterministic model that for each shipment selects the low-cost option with probability one is still very useful when predicting the chain type. It follows the chain frequency observed in the CFS very closely, showing that the minimum-cost-principle can capture certain mode choice patterns very well.

On the other hand, the mock-deterministic model predicts choices of shipment sizes that are larger than observed. This holds true for virtually all commodity groups. The model is only accurate for shipments of coal, petroleum and natural gas where these large sizes actually are observed. This feature makes the predicted tonnes per chain by the mock-deterministic model very inaccurate.

When it comes to predicted frequencies of the logit models, we find that in-sample predictions are very close to observed levels. This is likely due to the inclusion of constants for chains and shipment size, which leads to close to perfect predictions. One concern is that including a large set of chain and shipment size constant leads to overfitting. Reassuringly though, the out-of-sample predictions are also accurate.

When it comes to predicted tonnes by the logit model, there is a high degree of consistency between observed and predicted tonnes for most groups. The predictive performance of our baseline model is superior to the mock-deterministic model. This illustrates how a logit-based model can outperform all-or-nothing assignment.

However, predictions are very poor for a few groups (2,3,5, 11 and 13). The difficulties do not appear to be caused by a particular chain being used for these groups. The chains observed vary across these groups and the same chains are accurately predicted in other groups. Rather, it appears that it is the group-chain-combination that is difficult to predict.

Additional analyses

We also make a set of additional analyses that test if the suitability of our coefficients depends on a range of changes in the procedure used in the main analysis. We show that the inaccuracy of predictions for some of the commodity groups can be handled by recoding the implied weight of each predicted weight class interval. The adjusted predictions show a high degree of consistency with the observed levels. This shows that the inaccuracy of the initial predictions is not due to incorrect modal shares but that the shipment sizes are poorly predicted for these commodity groups.

This confirms that it seems to be more difficult to predict how much firms will ship than which chain they will use. But it is not obvious that the weight class interval used in the analysis is

inappropriate and should be adjusted since the existing one works well for the other commodity groups.

Reducing the weight class interval from 16 to five categories produces lower transport cost overall and for chosen alternatives. Coefficients estimated on the alternative interval are similar to those from the original interval, although some have the incorrect sign. Predicted tonnes based on the alternative weight class interval overpredicts the levels for road transportation severely. This illustrates the benefit of keeping the more fine-grained weight class interval in the logit application.

We show that an expanded choice set that differentiates between container and non-container options leads to inclusion of additional alternatives that are similar in terms of transport time and distance but have lower transport cost on average. The difference in average costs is substantially reduced when we restrict attention to alternatives recorded as chosen in the CFS. This suggest that the alternatives included only in the expanded choice set have a relatively low probability of being selected anyway. Predictions are improved when using the expanded choice set, which is possibly a mechanic effect of using more fine-grained alternatives in combination with size- and chain-specific constants.

We compare our baseline multinomial logit to a nested logit model and show that the coefficients are almost always of identical sign but vary in size. The nested logit does slightly better in predicting tonnes.

Finally, we show that our estimated coefficients do not systematically predict worse when applied to a set of alternative-specific transport costs that are slightly different than those used in the main estimation. The predictions based on alternative cost function are more accurate for some chains and less accurate for others, compared to the original predictions. This gives some support to the robustness of our estimated coefficients.

Implications for implementing a stochastic module in Samgods

We have found that the best model specification includes constants that are specific for chain types and for weight classes, transport cost that enters separately for each chain type and a general measure of transport time. All coefficients vary across commodity groups. Our main output is the set of coefficients that can be used as a basis for the logit-formula in the stochastic logistics module. The estimates are provided in Appendix A. Applying these to the Commodity Flow Survey gives predictions that overall are in line with the observed frequencies and tonnes.

Our results show that several adjustments may have to be made when applying these coefficients in the logistics module of Samgods. In our application, a rescaling of the predicted tonnes was needed for some of the commodity groups to achieve better predictions. We varied the weight associated with each weight interval and used to observed tonnes as target. An alternative approach is to include the logarithm of observed over predicted tonnes of each chain in the model specification of the logit model, estimate and predict iteratively until the target has been reached. This was the procedure used when a stochastic logistics module was developed for two commodity groups in a previous project (Abate et al. 2016).

The choice set in our analysis and the proposed stochastic logistics module consists of 14 chain types and 16 shipment sizes. Expanding the choice set to differentiate between container and non-container option does not seem to include options that are relevant for shippers in the CFS anyway. The transport alternative considered in the original choice set thus tends to be more relevant for the choice. Prediction accuracy did improve with a more fine-grained choice set, which makes it worth to consider alternative dimensions of the choice set. However, the benefits of an expanded choice set should be weighed against the increase in work needed for implementing and running such logistics module.

There are several thinkable changes to be made to the dimension of the choice set in the stochastic module, not least the way Samgods vehicle types maps to each transport chain and which chains and shipment weights to be included. If such changes are made it is possible to combine (e.g., by averaging or replacing) the estimated coefficients for different chain types. This is particularly relevant for the three chains in the choice set containing ferry transportation and for which no coefficients were estimated in this study. These could potentially be proxied by the coefficient estimates for vessel transportation — e.g. coefficients for chain 8 (truck-ferry-truck) is approximated by those for chain 9 (truck-vessel-truck).

Our estimated coefficients appear relatively robust to alternative transport cost. Small changes to input factors determining these data should therefore not be a concern. This result is partly due to the many constants in our model specification which makes the impact of new transport cost on choice probabilities smaller. Should larger changes to input factors (and transport costs) occur, it is possible to scale the coefficients at a later point to account for the fact that variables scale has been changed. The coefficients can be adjusted by scaling each coefficient by a factor equal to the average difference between the old and the new variable. If this is not sufficient to produce reasonable coefficients, new coefficients could be estimated using the new data.

We have estimated and evaluated discrete choice models for all of the 16 commodity groups available in the next version of Samgods. However, it is not obvious that a stochastic module is appropriate or needed for all groups. One of the main reasons for using a stochastic module is to reduce the occurrences of large changes in the choice of transport solution when an underlying choice parameter (e.g. transport cost) is changed. But such consideration is less important for commodities for which there is virtually no competition between modes.

For most commodity groups there is variation in the tonnes shares based on the 14 chains and the CFS. Only for commodity group 1 (agriculture and forestry products (excluding round wood) and fishing) and 2 (coal, crude petroleum and natural gas) has a single chain more than 90 percent of the market. Another reason for opting for a deterministic approach for some commodities is that the mode choice is difficult to predict using a stochastic approach. The results from the main analysis shows that the predicted tonnes for commodities 2, 3, 5, 7 and 14 are very far from the observed levels. However, we showed that the model predicted shares accurately for these groups and the predicted value of shipment weight can be adjusted to align the predicted tonnes with the observed levels. Based on this, we conclude that although some commodities were harder to predict than others and have limited variation in modal

choice, there are issues concerning implementation of a mixed logistics module (i.e. with both stochastic and deterministic modelling) that should be given more weight in the decision to continue model some commodities deterministically.

Finally, we find that the nested logit does slightly better in predicting tonnes. Given that the nested logit formulation we used was fairly simple and that model fit and performance therefore probably can be improved upon, this type of model could also be used in the implementation of a stochastic logistics module. It does come with greater computational requirements and the simplicity of the MNL is one argument for using it in implementation and future work.

References

Abate, M., Vierth, I., de Jong, G., Baak, J. (2014) Joint econometric models of freight transport chain and shipment size choice. CTS working paper 2014:9

Abate, M., Vierth, I., Karlsson, R., de Jong, G., Baak, J. (2016): Estimation and implementation of joint econometric models of freight transport chain and shipment size choice. CTS Working Paper 2016:1.

Lindgren, S., Johansson, M., Vierth, I., Karlsson, R., Baak, J., de Jong, G. (2019): A stochastic logistics module in Samgods — Work tasks for full implementation.

Train, K. (2003) Discrete Choice Methods with Simulation. Cambridge University Press, MA.

Windisch, E., de Jong, G.C and van Nes, R., 2010: A disaggregate freight transport model of transport chain choice and shipment size choice. Paper presented at ETC 2010, Glasgow.

Appendix A: MNL Coefficients

Commodity	(1)		(2)		(3)		(4)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Chain 1	8.90***	(0.41)	3.56***	(1.04)	5.81***	(0.20)	13.9***	(0.41)
Chain 4					-1.04*	(0.51)	10.0***	(1.04)
Chain 6							1.01	(1.46)
Chain 7			0.83	(1.24)	0.60***	(0.17)	10.7***	(0.43)
Chain 8								
Chain 9					2.69***	(0.28)	8.42***	(0.42)
Chain 10					-4.54***	(1.13)	2.20**	(0.71)
Chain 13			0.61	(0.60)	2.21***	(0.47)		
Chain 14								
Size 2	-1.07***	(0.072)			0.035	(0.32)	-1.70***	(0.021)
Size 3	-0.37***	(0.059)			0.029	(0.31)	-2.35***	(0.026)
Size 4	0.64***	(0.050)			0.50	(0.29)	-2.85***	(0.028)
Size 5	0.52***	(0.052)			0.53	(0.28)	-3.69***	(0.031)
Size 6	-0.100	(0.057)			1.25***	(0.27)	-4.75***	(0.039)
Size 7	-0.10	(0.057)			2.16***	(0.26)	-4.91***	(0.041)
Size 8	0.0045	(0.056)			1.39***	(0.26)	-4.78***	(0.041)
Size 9	-0.23***	(0.058)			1.58***	(0.26)	-5.97***	(0.056)
Size 10	0.39***	(0.052)			1.47***	(0.26)	-4.82***	(0.041)
Size 11	-0.25***	(0.057)	1.95*	(0.81)	1.23***	(0.26)	-6.09***	(0.058)
Size 12	-3.09***	(0.16)			1.93***	(0.26)	-7.00***	(0.084)
Size 13	-4.91***	(0.38)			-1.52***	(0.42)	-9.54***	(0.22)
Size 14	-5.77***	(0.55)			-2.71***	(0.64)	-10.8***	(0.40)
Size 15	-6.87***	(1.00)			-2.53***	(0.57)	-11.9***	(0.60)
Size 16	-6.18***	(0.71)	5.19***	(0.77)	1.97***	(0.26)	-10.7***	(0.40)
Chain 1 x cost					-0.0041***	(0.00025)	-0.0019***	(0.000040)
Chain 4 x cost							-0.0022***	(0.00027)
Chain 6 x cost					-0.0023***	(0.00046)	-0.00015	(0.00019)
Chain 7x cost							-0.0040***	(0.00018)
Chain 8 x cost					-0.0017**	(0.00055)		
Chain 9 x cost							-0.0015***	(0.000077)
Chain 10 x cost							-0.0000022	(0.0000035)
Chain 13 x cost								
Cost			-0.0050***	(0.0013)	-0.0025	(0.0015)		
Time								
Com.	1		2		3		4	
Obs	146,340		840		19,100		1,618,860	
LL	-12013.1		-33.5		-2237.2		-53433.1	
Parameters	16		6		25		27	

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Commmodity	5		6		7		8	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Chain 1	17.2***	(1.51)	7.61***	(0.14)	7.30***	(0.20)	8.42***	(0.28)
Chain 4	10.7***	(1.47)	6.36***	(0.30)			4.12***	(0.34)
Chain 6			-0.42	(0.29)			0.30	(0.43)
Chain 7			4.18***	(0.15)			5.05***	(0.32)
Chain 9	11.8***	(1.40)	4.59***	(0.15)			5.82***	(0.27)
Chain 10	8.98***	(1.37)	0.60***	(0.16)			2.15***	(0.28)
Chain 13			-0.75	(0.96)			1.68*	(0.78)
Size 1	10.3***	(0.45)						
Size 2			-1.78***	(0.052)	0.93**	(0.33)	-1.81***	(0.076)
Size 3			-1.83***	(0.053)	1.03**	(0.32)	-2.20***	(0.083)
Size 4			-1.53***	(0.049)	1.76***	(0.28)	-2.35***	(0.086)
Size 5			-2.05***	(0.055)	1.76***	(0.28)	-3.18***	(0.10)
Size 6			-2.57***	(0.064)	1.88***	(0.28)	-3.59***	(0.11)
Size 7			-2.27***	(0.058)	7.13***	(0.25)	-2.67***	(0.095)
Size 8			-1.49***	(0.052)	-0.49	(0.45)	-3.35***	(0.11)
Size 9			-2.99***	(0.071)	-2.07**	(0.75)	-4.58***	(0.15)
Size 10			-2.91***	(0.070)	-2.41**	(0.75)	-4.38***	(0.14)
Size 11			-3.99***	(0.094)	-3.23***	(0.84)	-3.78***	(0.12)
Size 12			-2.69***	(0.067)	-3.90***	(1.09)	-5.13***	(0.17)
Size 13			-5.39***	(0.17)	-3.61***	(1.08)	-7.87***	(0.52)
Size 14			-5.83***	(0.20)	-1.81**	(0.56)	-8.63***	(0.72)
Size 15			-6.06***	(0.21)	3.38***	(0.26)	-7.29***	(0.39)
Size 16			-5.87***	(0.20)			-7.69***	(0.51)
Chain 1 x cost	-0.0013***	(0.00012)	-0.0012***	(0.000036)	-0.0031***	(0.00029)	-0.00086***	(0.000043)
Chain 4 x cost	-0.00069**	(0.00024)	-0.0014***	(0.00011)			-0.00093***	(0.00014)
Chain 7x cost			-0.00059***	(0.000054)			-0.00084***	(0.00011)
Chain 9 x cost			-0.00099***	(0.000068)			-0.0010***	(0.000075)
Chain 13 x cost			-0.00081	(0.00077)			-0.00096	(0.00054)
Obs	111,254		154,600		225,480		79,150	
LL	-331.2		-12414.6		-1471.2		-4927.6	
Parameters	7		27		16		27	

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Com.	9		10		11		12	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Chain 1	9.87***	(0.74)	5.46***	(0.085)	9.28***	(0.71)	11.0***	(1.01)
Chain 4			0.76***	(0.15)	6.46***	(0.71)	8.51***	(1.08)
Chain 6			0.52*	(0.25)	1.62*	(0.80)		
Chain 7	4.19***	(0.89)	3.72***	(0.10)	2.19**	(0.76)	7.30***	(1.11)
Chain 9	5.54***	(0.74)	2.37***	(0.082)	6.21***	(0.71)	8.84***	(1.00)
Chain 10	0.31	(0.41)	-0.20**	(0.062)	7.62***	(0.71)	6.19***	(1.02)
Chain 13			1.35	(1.18)	1.84	(1.23)	8.72**	(2.79)
Size 2	-0.83***	(0.10)	-1.29***	(0.045)	-0.37***	(0.033)	-1.04***	(0.065)
Size 3	-0.75***	(0.10)	-1.57***	(0.048)	-0.68***	(0.036)	-0.98***	(0.068)
Size 4	-0.19*	(0.092)	-1.87***	(0.052)	-1.45***	(0.049)	-1.53***	(0.074)
Size 5	-0.30**	(0.095)	-2.33***	(0.058)	-2.41***	(0.074)	-2.04***	(0.083)
Size 6	-0.46***	(0.10)	-2.65***	(0.064)	-3.44***	(0.11)	-2.65***	(0.092)
Size 7	-0.21*	(0.095)	-2.70***	(0.065)	-3.81***	(0.14)	-2.99***	(0.10)
Size 8	-0.78***	(0.11)	-2.10***	(0.057)	-4.98***	(0.25)	-3.43***	(0.12)
Size 9	-1.39***	(0.12)	-4.09***	(0.10)	-6.45***	(0.50)	-6.43***	(0.33)
Size 10	-1.23***	(0.11)	-4.14***	(0.10)	-7.83***	(1.00)	-7.19***	(0.46)
Size 11	-1.76***	(0.13)	-5.19***	(0.15)			-3.98***	(0.13)
Size 12	-2.09***	(0.15)	-3.50***	(0.084)			-6.50***	(0.32)
Size 13	-5.57***	(0.59)	-6.74***	(0.31)			-7.33***	(0.46)
Size 14	-6.72***	(1.01)	-7.17***	(0.36)			-7.12***	(0.42)
Size 15	-6.65***	(1.01)	-7.86***	(0.50)			-9.04***	(1.01)
Size 16	-4.40***	(0.34)	-6.16***	(0.23)			-7.94***	(0.59)
Chain 1 x cost	-0.00074***	(0.00013)	-0.00073***	(0.000028)			-0.00063***	(0.000045)
Chain 4 x cost			-0.00070***	(0.000087)			-0.00063***	(0.000093)
Chain 6 x cost			-0.00042***	(0.000080)				
Chain 7x cost	-0.00070	(0.00054)	-0.0012***	(0.000067)			-0.0031***	(0.00068)
Chain 9 x cost			-0.00037***	(0.000027)			-0.00045***	(0.000027)
Chain 10 x cost							-0.0000048**	(0.0000016)
Chain 13 x cost			-0.0042**	(0.0015)			-0.0027	(0.0015)
Obs	50,140		137,790		215,130		80,560	
LL	-3589.8		-11113.9		-21773		-5649.2	
Parameters	21		28		16		27	

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Com.	13		14		15		16	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Chain 1	5.55***	(0.12)	7.17***	(1.11)	7.94***	(0.12)	9.30***	(0.29)
Chain 6								
Chain 7			2.56	(2.90)				
Chain 9	2.67***	(0.13)						
Chain 10	2.11***	(0.13)					3.43***	(0.23)
Chain 13								
Size 1							10.6***	(0.40)
Size 2	0.32***	(0.059)	0.68	(1.19)	1.36***	(0.17)	8.05***	(0.40)
Size 3	-0.14*	(0.066)	0.45	(1.19)	1.09***	(0.20)	7.65***	(0.40)
Size 4	-0.86***	(0.083)	-0.85	(1.70)	1.63***	(0.21)	6.56***	(0.40)
Size 5	-1.95***	(0.12)	3.11***	(0.71)	2.43***	(0.20)	4.75***	(0.42)
Size 6	-3.65***	(0.24)	2.99***	(0.69)	2.38***	(0.20)	4.12***	(0.44)
Size 7	-4.28***	(0.34)	1.21	(0.78)	2.49***	(0.21)	3.73***	(0.45)
Size 8	-2.18***	(0.14)	2.96***	(0.65)	2.45***	(0.21)	3.73***	(0.43)
Size 9	-5.13***	(0.51)	1.89**	(0.72)	1.71***	(0.21)		
Size 10	-4.20***	(0.34)	2.50***	(0.69)	2.65***	(0.21)		
Size 11			0.63	(0.97)	3.25***	(0.21)		
Size 12	-6.61***	(1.00)			2.20***	(0.21)		
Size 13	-5.79***	(0.71)			-1.65***	(0.25)		
Size 14					-2.73***	(0.31)		
Size 15			-0.97	(1.32)	-3.74***	(0.41)		
Size 16			-0.92	(1.26)	-1.12***	(0.24)		
Chain 1 x cost			-0.0051***	(0.0011)	-0.0035***	(0.00031)	-0.0017***	(0.000067)
Chain 7 x cost			-0.0041	(0.0048)				
Time							-0.085***	(0.019)
Obs	57,320		1,320		406,390		113,010	
LL	-4250		-113.3		-24463.8		-3004.4	
Parameters	14		16		17		12	

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix B: Correspondence key

Shortlist chaintypes to Samgods chaintypes

SHORTLIST CHAINTYPE		SAMGODS CHAINTYPES	
NO.	Description	Sub-ModeNr	Description
1	Truck	A	Heavy lorry
		B	Light Lorry
		C	Light Lorry
		c	Extra heavy lorry
2	Vessel	J	Direct Sea
		KL	Feeder vessel-Long-Haul vessel
		LK	Long-Haul vessel-Feeder vessel
		M	Direct Sea
3	Rail	GH	Feeder train-Wagonload train
		Gh	Feeder train-Long wagonload train
		GHG	Feeder train-Wagonload train-Feeder train
		HG	Wagonload train-Feeder train
		hG	Long wagonload train-Feeder train
		I	System train
		i	Long system train
		T	System train
		U	System train
		f	Long Wagonload train
4	Truck-Vessel	AJ	Heavy lorry-Direct Sea
		AKL	Heavy lorry-Feeder vessel-Long-Haul vessel
		AV	Heavy lorry-IWW
		CM	Heavy lorry-Direct Sea
		JA	Direct Sea-Heavy lorry
		LKA	Long-Haul vessel-Feeder vessel-Heavy lorry
		MC	Direct Sea-Heavy lorry
		VA	IWW-Heavy lorry
		WB	IWW-Light Lorry
5	Rail-Vessel	GHM	Feeder train-Wagonload train-Direct Sea
		GHMI	Feeder train-Wagonload train-Direct Sea-System train
		GHMT	Feeder train-Wagonload train-Direct Sea-System train
		GHMU	Feeder train-Wagonload train-Direct Sea-System train
		IM	System train-Direct Sea
		iM	Long system train-Direct Sea
		IMHG	System train-Direct Sea-Wagonload train-Feeder train
		MHG	Direct Sea-Wagonload train-Feeder train
		MI	Direct Sea-Long system train
		MT	Direct Sea-Long system train
		MU	Direct Sea-Long system train
		TM	System train-Direct Sea
		TMGH	System train-Direct Sea-Feeder train-Wagonload train

		UM	System train-Direct Sea
		UMGH	System train-Direct Sea-Feeder train-Wagonload train
		HM	Wagonload train-Direct Sea
		MH	Direct Sea-Wagonload train
6	Truck-Truck-Truck	BS	Light Lorry-Consolidated heavy lorry
		BSB	Light Lorry-Consolidated heavy lorry-Light Lorry
		SB	Consolidated heavy lorry-Light Lorry
		cB	Extra heavy lorry-Light Lorry
		cS	Extra heavy lorry-Consolidated heavy lorry
		cC	Extra heavy lorry-Heavy lorry
		Bc	Light Lorry-Extra heavy lorry
		XA	Extra heavy lorry-Heavy lorry
		AX	Heavy lorry-Extra heavy lorry
7	Truck-Rail-Truck	ADA	Heavy lorry-Kombi train-Heavy lorry
		AdA	Heavy lorry-Long kombi train-Heavy lorry
		CGH	Heavy lorry-Feeder train-Wagonload train
		CGHC	Heavy lorry-Feeder train-Wagonload train-Heavy lorry
		CH	Heavy lorry-Wagonload train
		Ch	Heavy lorry-Long wagonload train
		ch	Extra heavy lorry-Long wagonload train
		ChC	Heavy lorry-Long wagonload train-Heavy lorry
		CHG	Heavy lorry-Wagonload train-Feeder train
		CHGC	Heavy lorry-Wagonload train-Feeder train-Heavy lorry
		GHC	Feeder train-Wagonload train-Heavy lorry
		HC	Wagonload train-Heavy lorry
		hC	Long wagonload train-Heavy lorry
		hc	Long wagonload train-Extra heavy lorry
		HGC	Wagonload train-Feeder train-Heavy lorry
		XdX	Extra heavy lorry-Kombi train-Extra heavy lorry
		cH	Extra heavy lorry-Wagonload train
		XF	Extra heavy lorry-Wagonload train
8	Truck-Ferry-Truck	APA	Heavy lorry-Road Ferry-Heavy lorry
		CPC	Heavy lorry-Road Ferry-Heavy lorry
9	Truck-Vessel-Truck	AJA	Heavy lorry-Direct Sea-Heavy lorry
		AVA	Heavy lorry-IWW-Heavy lorry
		CMC	Heavy lorry-Direct Sea-Heavy lorry
		CWC	Heavy lorry-IWW-Heavy lorry
		cWc	Extra heavy lorry-IWW-Extra heavy lorry
10	Truck-Air-Truck	BR	Light Lorry-Plane
		BRB	Light Lorry-Plane-Light Lorry
		RB	Plane-Light Lorry
11	Truck-Ferry-Rail-Truck	HQH	Wagonload train-Rail Ferry-Wagonload train
12	Truck-Rail-Ferry-truck	GHQH	Feeder train-Wagonload train-Rail Ferry-Wagonload train
13	Truck-Vessel-Rail-Truck	AJDA	Heavy lorry-Direct Sea-Kombi train-Heavy lorry
		CMI	Heavy lorry-Direct Sea-System train

		CMT	Heavy lorry-Direct Sea-System train
		CMU	Heavy lorry-Direct Sea-System train
		LKDA	Long-Haul vessel-Feeder vessel-Kombi train-Heavy lorry
		MHGC	Direct Sea-Wagonload train-Feeder train-Heavy lorry
		AJdX	Heavy lorry-Direct Sea-Kombi train-Extra heavy lorry
14	Truck-Rail-Vessel-Truck	ADJA	Heavy lorry-Kombi train-Direct Sea-Heavy lorry
		ADJDA	Heavy lorry-Kombi train-Direct Sea-Kombi train-Heavy lorry
		ADKL	Heavy lorry-Kombi train-Feeder vessel-Long-Haul vessel
		CGHM	Heavy lorry-Feeder train-Wagonload train-Direct Sea
		CUM	Heavy lorry-System train-Direct Sea
		IMC	System train-Direct Sea-Heavy lorry
		TMC	System train-Direct Sea-Heavy lorry
		UMC	System train-Direct Sea-Heavy lorry
		XdJA	Extra heavy lorry-Kombi train-Direct Sea-Heavy lorry
		CHM	Heavy lorry-Wagonload train-Direct Sea

Appendix C: Commodity groups

Code	Name	NST 2007
01	Products of agriculture, hunting, and forestry; fish and other fishing products. Not timber	01 excl. timber
02	Coal and lignite; crude petroleum and natural gas	02
03	Metal ores and other mining and quarrying products; peat	03
04	Food products, beverages and tobacco	04
05	Textiles and textile products; leather and leather products	05
06	Wood and products of wood and cork (except furniture); pulp, paper and paper products; printed matter and recorded media	06
07	Coke and refined petroleum products	07
08	Chemicals, chemical products, and man-made fibers; rubber and plastic products; nuclear fuel	08
09	Other non metallic mineral products	09
10	Basic metals; fabricated metal products, except machinery and equipment	10
11	Machinery and equipment; medical, precision and optical instruments	11
12	Transport equipment	12
13	Furniture; other manufactured goods	13
14	Secondary raw materials; municipal wastes and other wastes	14
15	Timber	01, part of
16	Air freight (fractions of some of the commodity groups)	…