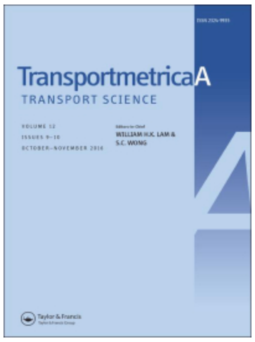


Transportmetrica A: Transport Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttra21

# Latent split of aggregate counts: revealing home deliveries per commodity types and potential freight trip implications

Elise Caspersen, Mario Arrieta-Prieto & Xiaokun (Cara) Wang

**To cite this article:** Elise Caspersen, Mario Arrieta-Prieto & Xiaokun (Cara) Wang (2021): Latent split of aggregate counts: revealing home deliveries per commodity types and potential freight trip implications, Transportmetrica A: Transport Science, DOI: 10.1080/23249935.2021.1990438

**To link to this article:** https://doi.org/10.1080/23249935.2021.1990438

Published online: 04 Nov 2021.

TRANSPORTMETRICA A: TRANSPORT SCIENCE
https://doi.org/10.1080/23249935.2021.1990438

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

# Latent split of aggregate counts: revealing home deliveries per commodity types and potential freight trip implications

Elise Caspersen[a,b], Mario Arrieta-Prieto[c] and Xiaokun (Cara) Wang[d]

[a]The Norwegian Institute of Transport Economics, Oslo, Norway; [b]The School of Economics and Business, Norwegian University of Life Science, Ås, Norway; [c]Universidad Nacional de Colombia, Bogotá, Colombia; [d]Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

**ABSTRACT**

This paper suggests a joint econometric model that allows estimating latent marginal counts when only total counts and types of commodities purchased are available. The basis for this model is the Negative binomial hurdle model, which is expanded by incorporating different features for the latent classes, allowing eventual null latent counts for one or more classes. A validation procedure for the proposed splitting is discussed. The methodology was used to estimate and validate a model for the propensity to shop online and the corresponding number of shipments per commodity group. The results confirm existing research on online shopping behaviour: elderly is less likely to buy online, while high income, education and having kids motivate online shopping. The average online shopper receives 2.4 shipments/month (0.077 shipments/day), with variations in shipments and commodities depending on the consumer profile. Correlation between commodity groups reveals that consolidation can reduce shipments of up to 30%.

**ARTICLE HISTORY**
Received 18 October 2020
Accepted 1 October 2021

**KEYWORDS**
E-commerce; freight trip generation; hurdle model; latent count estimation; hypothesis testing; consolidation opportunities

## 1. Introduction

Online shopping is growing extensively, with European e-commerce forecasted to hit €717 billion in 2020 (Ecommerce News 2020). An early assumption was that e-commerce would crowd out shopping trips and as such, reduce total traffic and its burden on society; but empirical evidence contradicts this (Zhou and Wang 2014; Cao, Xu, and Douma 2012; Pettersson, Hiselius, and Koglin 2018). Instead, online shoppers are increasingly sophisticated consumers, pushing for faster and cheaper deliveries, including new delivery services like 'instant' deliveries (Dablanc et al. 2017). As such, duplication of delivery services with low vehicle utilisation in terms of payload has been observed (Allen et al. 2018; Ni, Wang, and Zhang 2016) and consolidation is challenged as competitors fight for market shares (Allen et al. 2018; Pettersson, Hiselius, and Koglin 2018). A combination of high-level delivery services and receivers located in residential units makes online shopping one of the main contributors to growing freight traffic in residential areas (Visser, Nemoto, and Browne

**CONTACT** Xiaokun (Cara) Wang ✉ wangx18@rpi.edu Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180, USA

2014; Allen et al. 2018). Although a vast research on drivers of online shopping exists (Farag et al. 2007; Cao, Xu, and Douma 2012; Zhou and Wang 2014), there is relatively little knowledge on the relationship between consumers' online shopping activity, number of shipments and freight traffic, despite the magnitude of freight trips generated by residential units being no longer trivial (Wang and Zhou 2015; Holguín-Veras et al. 2018; Saphores and Xu 2020). In later years, online shopping has expanded from commodities with high searchability and ease of making a quality judgement prior to experience, identified as attractive commodities for online shopping (de Figueiredo 2000; Girard, Silverblatt, and Korgaonkar 2002; Lowengart and Tractinsky 2001); to clothes or even groceries, with particularly strict delivery requirements and difficulties for quality assessments online (Lagorio and Pinto 2020; Saphores and Xu 2020). This results in heterogeneous online purchases with different delivery requirements, lower degree of consolidation and a higher degree of returns, resulting in increased freight traffic.

Hence, more knowledge on shipments of different types of commodities from online shopping is necessary, motivating the following research questions addressed in this paper: (i) *which factors explain the probability to shop online and the corresponding number of shipments received, separated on commodity groups*; and (ii) *what are the freight trip generation implications of this new knowledge*? To accommodate for the potential tedious and time-consuming task it is to collect individual data about number of shipments per commodity group, this paper builds on the methodology by Afghari et al. (2016); Afghari et al. (2018) to estimate latent marginal counts when only total counts and types of commodities purchased are available. The estimation results reveal shopping patterns that suggest potential for increased consolidation between commodities to the same consumer or neighbourhood. The latter is further explored, revealing potential of reducing number of shipments up to 30%.

There is a vast literature about the use of latent-information models to address unobserved effects and manage spatio-temporal correlation patterns, especially in the context of road safety; at the level of an entire population of interest (Yasmin and Eluru 2016; Park and Lord 2009; Zou, Zhang, and Lord 2013; Heydari et al. 2017; Castro, Paleti, and Bhat 2012). However, to the best of our knowledge, this is the first paper that develops a complete statistical framework to infer latent counts and correct them based on the evidence provided by the data at the individual level, i.e. by estimating latent subcounts that differ for each individual in the sample using individual-specific covariates and shopping records (which categories were purchased). It is also the first paper that uses information about total number of shipments, indicators for commodity type purchase and joint econometric modelling to estimate number of shipments per aggregated commodity type from online shopping. This work is also pioneer in identifying the impact of consolidation strategies based on correlation analysis of the different consumption profiles of interest.

The rest of the paper is organised as follows: Section 2 introduces the proposed methodological development, which is tested and validated in Section 3. The model is applied to the total reported number of shipments from online shopping and indicator variables for commodity groups purchased to estimate the number of shipments per commodity group from online shopping. The case study is described in Section 4, followed by estimation results and implications in Sections 5 and 6. A brief conclusion is provided in Section 7.

## 2. Hurdle model and latent marginal count estimation

The proposed model attempts to describe the driving factors related with the decision of buying online or not and, in the case of online shoppers, which characteristics influence the amount of shipments received. This objective can be achieved by using a hurdle model, which firstly characterises the probability of buying online by means of a binary-outcome model (usually logit), and then, incorporates a count data model to explain the nonzero counts observed for buyers (Washington, Karlaftis, and Mannering 2003). Besides the philosophical discussions of applicability of hurdle models in comparison with zero-inflated models depending on the context (which are thoroughly discussed in the case study section), hurdle models offer an advantage in terms of the flexibility in the data requirements, since they allow to include different covariates to characterise the probability of buying versus not buying, and then to explain the observed counts for buyers, as illustrated in the details of the modelling framework.

In the context of online shopping, it is of interest to identify the count of shipments received per type of commodity and which socio-economic variables influence that marginal count. However, to avoid respondent fatigue from many detailed questions, which may lead to carelessness in answering, superficial responses or guesses (Stopher 2012), the number of shipments are rarely collected for a wide variety of commodity groups. It is also questionable whether respondents can remember their purchases in such a detailed level. Thus, questionnaires are often limited, as is the case with the data considered in this paper asking for the total number of shipments and commodity type purchased instead of the number of shipments per commodity type. To elicit the split for different commodity types, given the information about total counts and records of commodity type purchased, latent marginal-count estimation procedures can be used.

Let $Y_{\mathrm{i}}$ represent the total number of shipments received by individual *i*provided that he/she is an online shopper ($Y_{\mathrm{i}} > 0$). Let $Y_{\mathrm{ij}}$ represent the (latent) number of shipments received from commodity type *j*. It is clear, then, that if *m* commodity types are considered

$$Y_{\mathrm{i}} = \sum_{j=1}^{m} Y_{\mathrm{ij}}, \ \forall i. \tag{1}$$

If $\lambda^{(\mathrm{i})}$ represents the mean of total shipments for individual *i* and $\lambda_{\mathrm{j}}^{(\mathrm{i})}$ the mean of shipments from commodity type *j* for that same individual, regardless of the distribution assumptions and the association structure between the sub-counts, it holds that

$$\lambda^{(\mathrm{i})} = \sum_{j=1}^{m} \lambda_{\mathrm{j}}^{(\mathrm{i})}. \tag{2}$$

$Y_{\mathrm{i}}$ is usually modelled as a Negative Binomial distribution of parameters $\lambda^{(\mathrm{i})}$and $\alpha$, with $\alpha$ accounting for heterogeneity and overdispersion. With a similar framework, Afghari et al. propose a methodology for estimation of latent marginal counts in the context of road safety (Afghari et al. 2016, 2018). In their papers, each latent count mean, $\lambda_{\mathrm{j}}^{(\mathrm{i})}$, is modelled in terms of a vector of group-specific covariates, $X_{\mathrm{ij}}$, and a vector of group-specific parameters,

$\beta_{\mathrm{j}}$, such that

$$\lambda_{\mathrm{j}}^{(\mathrm{i})} = \exp(\beta_{\mathrm{j}}^{T} X_{\mathrm{ij}}). \tag{3}$$

After substituting (3) in (2), the resulting expression for $\lambda^{(\mathrm{i})}$ is replaced for each individual in the likelihood function for parameter estimation.

Afghari et al's methodology assumes that all the *j* subgroups considered have a non-negative count and relies entirely on the estimation routine to infer the estimate of the latent mean counts (Afghari et al. 2016, 2018). The dataset under analysis in this work contains information of the types of commodity purchased by the individuals, so, although a maximum number of commodity types *m* is defined, this model allows null latent counts for a given buyer, as a result of the reasonable statement that not all types of commodities are purchased by the same individual. The proposed model's goal is to identify the major drivers that affect the total mean count, $\lambda^{(\mathrm{i})}$, while allowing the impact of the covariates to vary with the commodity types the user reported to have purchased.

Following the principles of a count data model, the total mean count is characterised as

$$\ln(\lambda^{(\mathrm{i})}) = I_1^{(\mathrm{i})} \cdot \left( \sum_{k=1}^{K_1} \beta_{\mathrm{k1}} x_{\mathrm{k1}}^{(\mathrm{i})} \right) + I_2^{(\mathrm{i})} \cdot \left( \sum_{k=1}^{K_2} \beta_{\mathrm{k2}} x_{\mathrm{k2}}^{(\mathrm{i})} \right) + \ldots + I_{\mathrm{J}}^{(\mathrm{i})} \cdot \left( \sum_{k=1}^{K_J} \beta_{\mathrm{kJ}} x_{\mathrm{kJ}}^{(\mathrm{i})} \right). \tag{4}$$

The binary variable $I_{\mathrm{j}}^{(\mathrm{i})} \in \{0, 1\}$ indicates if the *i*-th individual buys from the *j*-th commodity type or not and it is derived from the information given by the buyer, i.e. it is a covariate in the model. The terms in parenthesis are the linear contributions of covariates associated with the latent count of commodity type $j \in [J] := \{1, 2, \ldots, J\}$, with $x_{\mathrm{kj}}^{(\mathrm{i})}$ being the *k*-th covariate considered to explain the count of the *j*-th commodity type for the *i*-th individual and$\beta_{\mathrm{kj}}$ its corresponding parameter. As such, the model consists of interactions between the explanatory and the indicator variables for each commodity type.

There are several reasons to favour this model structure over a more conventional one with only first-order general terms when the main purpose is to characterise the latent counts:

(1) Based on the objective of identifying the driving factors of the number of purchases of *J* commodity types, this model structure allows to isolate and directly obtain the marginal effects of the covariates for a specific type *j*. For instance, if *age* is included in the covariates explaining the counts for types $u \in [J]$ and $t \in [J] \backslash \{u\}$, the corresponding parameter associated with each type will explicitly show the marginal effect of *age* in the purchase of that specific group. A variable can have different effects depending on the commodity types in which it is considered a determinant factor.

(2) The candidate variables to explain the count for a commodity type can be determined independently from those considered for a different type. Continuing with the example above, this implies that the marginal effect of *age* in type *u* is independent of age being present in any of the other types, i.e. *t*.

(3) Another approach to capture the purchase behaviour of the individuals in the sample could be based on developing an explanatory model with different covariates for each one of the different purchase profiles included in the sample. If *J*commodity types are included in the analysis, there are $2^J$ purchase profiles depending on the types of commodities bought by the user (a purchase profile is intended as a record of *J* 'Yes/No'

answers to the question if *j*-th commodity type was purchased or not). However, even for a moderate-size number of commodities (e.g. $J = 5$), the number of purchase profiles grows exponentially (e.g. $2^5 = 32$), adding an unnecessary computational burden to the estimation routines.

The parameters of the model are estimated via maximum likelihood of a hurdle-type discrete count model, under the assumption of negative-binomial-distributed outcomes for each one of the buyers. Following Greene (2000), the Negbin II parametrization of the random variable $Y_\mathrm{i}$, allows to specify its distribution of the number of shipments for a buyer as

$$Y_\mathrm{i} \sim NB\left(\frac{\lambda^{(\mathrm{i})}(X_\mathrm{i})}{\lambda^{(\mathrm{i})}(X_\mathrm{i}) + \alpha}, \alpha\right), \tag{5}$$

with $\lambda^{(\mathrm{i})}(X_\mathrm{i})$ being the mean of the distribution, and $\alpha$ the overdispersion parameter. $\lambda^{(\mathrm{i})}(X_\mathrm{i})$ depends on (a subset of) covariates, $X_\mathrm{i}$, for the *i*-th individual as expressed in (4). Its probability mass function can be then expressed as

$$\begin{aligned}\Pr_{NB}(Y_\mathrm{i} = y_\mathrm{i}) &:= p_{NB}^{(\mathrm{i})}(y_\mathrm{i}) \\ &= \frac{\Gamma(\alpha + y_\mathrm{i})}{\Gamma(y_\mathrm{i} + 1)\Gamma(\alpha)}\left(\frac{\lambda^{(\mathrm{i})}(X_\mathrm{i})}{\lambda^{(\mathrm{i})}(X_\mathrm{i}) + \alpha}\right)^{y_\mathrm{i}}\left(\frac{\alpha}{\lambda^{(\mathrm{i})}(X_\mathrm{i}) + \alpha}\right)^{\alpha}.\end{aligned} \tag{6}$$

The hurdle component introduces the possibility that individuals have the option of either to purchase or not, decision driven by a binary choice outcome with probability $\pi^{(\mathrm{i})}(\tilde{X}_\mathrm{i})$, which is potentially dependent on some individual features or covariates, $\tilde{X}_\mathrm{i}$. In that regard, the complete probability mass function for a given individual (either a buyer or not) results to be

$$\Pr(Y_\mathrm{i} = y_\mathrm{i}) = \begin{cases}\pi^{(\mathrm{i})}(\tilde{X}_\mathrm{i}),\ y_\mathrm{i} = 0 \\ \frac{(1-\pi^{(\mathrm{i})}(\tilde{X}_\mathrm{i}))p_{NB}^{(\mathrm{i})}(y_\mathrm{i})}{1-p_{NB}^{(\mathrm{i})}(0)},\ y_\mathrm{i} > 0\end{cases} \tag{7}$$

The influence of the negative binomial structure appears only in the second piece of the probability mass (when $y_\mathrm{i} > 0$), as it is intended to model only the behaviour of the buyers. Notice also that in the estimation of the parameter associated with the purchase decision, $\pi^{(\mathrm{i})}(\tilde{X}_\mathrm{i})$, a set of covariates potentially different from the ones included in the negative binomial mean, $\lambda^{(\mathrm{i})}(X_\mathrm{i})$ are considered; which is particularly advantageous when only a restricted set of covariates is available for non-buyers.

The likelihood function is simply

$$L(y_1, y_2, \ldots, y_\mathrm{N}) = \prod_{i=1}^{N} \Pr(Y_\mathrm{i} = y_\mathrm{i}), \tag{8}$$

the product of the marginal probability mass functions for all the individuals in the sample, since it is assumed that the total counts are independent between individuals, i.e. the information related to one individual does not impact the distribution of the total count for another individual. This is the only independence assumption introduced in this model. Although this model's purpose is to aid the splitting procedure into latent sub-counts for different commodity groups given a total observed count, the only variable that is modelled for each individual is the total count,$Y_i$; therefore, no mention or consideration of the distribution or the co-dependence structure of the marginal subcounts, $Y_{ij}$, is required. Their mean parameters are, although, considered implicitly in the expression (4) for $\lambda^{(i)}(X_i)$, as (2) also suggested. The parameters accompanying the covariates in $\lambda^{(i)}(X_i)$ and $\pi^{(i)}(\tilde{X}_i)$ are estimated by maximising the logarithm of the likelihood.

The results of the likelihood estimation determine the magnitude of the impact of a given covariate in both the total number of shipments and the sub-counts for each commodity type. However, how can this information be used to provide actual estimates of the unobserved sub-counts associated the commodity groups of interest?

Once the estimation routine is finalised, the following estimate for the latent marginal counts is proposed

$$\hat{\lambda}_j^{(i)} = \begin{cases} exp\left(\sum_{k=1}^{K_j} \hat{\beta}_{kj} x_{kj}^{(i)}\right), \text{ if } I_j^{(i)} = 1. \\ 0, \text{ if } I_j^{(i)} = 0. \end{cases} \tag{9}$$

A methodology to test the validity of these latent count estimates is presented in Section 3. Via a statistical test of hypothesis, it is possible to verify if the data provide evidence towards favouring the proposed count splitting procedure. The next section also presents a way to adjust the latent-count estimators for each commodity type if the statistical evidence suggests that some assumptions are not met; but at the same time, the deviations are not significantly large.

## 3. Hypothesis testing for validation of the latent-count estimation

### 3.1. Procedure for hypothesis testing

It is expected that the mean latent counts satisfy the assumption that their summation equals the mean total count, i.e. $\lambda^{(i)} = \lambda_1^{(i)} + \lambda_2^{(i)} + \cdots + \lambda_J^{(i)}$, as introduced in (2). However, the proposed construction of the latent-count estimators provides no guarantee of that. To test this assumption, an indicator of the magnitude of discrepancy from the assumption can be defined as

$$\Delta^{(i)} := \ln\left(\frac{\sum_{j\in J} \lambda_j^{(i)}}{\lambda^{(i)}}\right). \tag{10}$$

The natural logarithm is conveniently introduced so that $\Delta^{(i)}$can take any value on the real line. A negative value would imply that the summation of the mean latent counts underestimates the mean total count, while a positive value would be an indicator of the opposite result.

If the latent-count estimators proposed are reasonable, one would expect that $\Delta^{(i)} = 0$. Therefore, the following hypothesis system is of interest

$$\begin{cases} H_0 : \Delta^{(i)} = 0 \\ H_1 : \Delta^{(i)} \neq 0 \end{cases}. \tag{11}$$

Using the estimates for $\hat{\lambda}_j^{(i)}$, it is possible to compute an estimate of$\Delta^{(i)}(\hat{\Delta}^{(i)})$ for each e-online shopper in the sample. Due to sample variability, measurement errors and unobserved effects; these estimates will not be exactly equal to zero even if all the assumptions are met, motivating the use of statistical tools to assess the hypothesis system presented. It must also be noted that the information of online shoppers who bought only from one commodity type should not be considered in the analysis since no splitting was necessary. These individuals have a $\hat{\Delta}$ value equal to zero by definition, but those zeroes do not provide evidence in favour or against the splitting procedure proposed and must be discarded.

Assuming that $\tilde{N}$ individuals purchased shipments from more than one commodity type, the vector of $\Delta$-estimates, $\vec{\Lambda} := (\hat{\Delta}^{(1)}, \hat{\Delta}^{(2)}, \ldots, \hat{\Delta}^{(i)}, \ldots, \hat{\Delta}^{(\tilde{N})})$, can be used to assess the hypothesis system in (11). However, it cannot be assumed that these realizations are independent from each other; although they come from different individuals, because they are all obtained using the same set of estimated parameters, which induces a co-dependence structure between them. After evaluating expression (10) with the proposed estimators, it is obtained that

$$\hat{\Delta}^{(i)} := k(\hat{\beta}, X_i) := \ln\left(\frac{\hat{\lambda}_1^{(i)}(\hat{\beta}, X_i) + \hat{\lambda}_2^{(i)}(\hat{\beta}, X_i) + \ldots + \hat{\lambda}_J^{(i)}(\hat{\beta}, X_i)}{\hat{\lambda}^{(i)}(\hat{\beta}, X_i)}\right), \tag{12}$$

where $k(\cdot,\cdot)$ is the function that relates the parameter estimates, $\hat{\beta}$, and the vector of covariates, $X_i$, for individual $i$; with the delta estimate $\hat{\Delta}^{(i)}$. The $\hat{\beta}$ estimators are present in the expression for each delta value, making them dependent, up to some extent, on each other. This prevents direct consideration of the delta estimates for any standard procedure of statistical hypothesis testing of their mean. To overcome this difficulty, a three-step procedure can be put in place.

(1) Identify a matrix B such that the vector $\vec{\delta} := B\vec{\Lambda}$ has all components with unit variance and uncorrelated from each other. This first step allows to break the co-dependence structure that the delta estimates might exhibit, by decorrelating its components. Appendix B presents a discussion on how to find the matrix B.

(2) Run a normality test over the elements in the vector $\vec{\delta} := B\vec{\Lambda}$ using any of the classical goodness-of-fit tests for normality. This step is crucial in the successful implementation of this methodology, because evidence of normality in addition to no-correlation (achieved in step 1) implies independence of the elements of $\vec{\delta}$. If normality is rejected, independence cannot be claimed, preventing the use of this methodology, since absence of correlation is not enough to build a statistical test supported in random sample theory. Under this scenario, a new methodology should be developed based, for example, on bootstrapping techniques.

(3) Since the vector $\vec{\delta} := B\vec{\Lambda}$ consists of a sequence of independent, unit-variance random variables, a simple t-test to check if the mean of these components is equal to zero suffices to provide a recommendation regarding the hypothesis presented in (11), since a zero mean in the original $\Delta^{(i)}$ variables is equivalent to a zero mean in the elements of the vector $\vec{\delta}$.

### 3.2. Approximate solutions when the null hypothesis gets rejected

By means of the approach previously described, if the hypothesis $\Delta^{(\mathrm{i})} = 0$, $\forall i$ cannot be rejected, there is evidence that $\hat{\lambda}_{\mathrm{j}}^{(\mathrm{i})}$, $\forall j$ are reasonable estimates for the counts of each commodity type the $i$-th online shopper purchases. On the other hand, if there is evidence to reject the hypothesis, a new proposal for the estimation of the latent counts should be proposed.

The correction method presented in this section attempts to mediate between the methodological developments done in this work and an eventual rejection of the null hypothesis. This correction is based on the idea that even if $\Delta^{(\mathrm{i})}$ cannot be claimed to be statistically equal to zero, its value is reasonably close to zero, hence there is not a substantial deviation from the assumptions already made. Nevertheless, the nonzero value for $\Delta^{(\mathrm{i})}$ is incorporated in a redefinition of the latent mean count estimates. Recall from (10) that

$$e^{\Delta^{(\mathrm{i})}} = \frac{\sum_{j \in J} \lambda_{\mathrm{j}}^{(\mathrm{i})}}{\lambda^{(\mathrm{i})}} \Leftrightarrow \lambda^{(\mathrm{i})} e^{\Delta^{(\mathrm{i})}} = \sum_{j \in J^{(\mathrm{i})}} \lambda_{\mathrm{j}}^{(\mathrm{i})}. \tag{13}$$

Since $\Delta^{(\mathrm{i})} \neq 0 \therefore \exp(\Delta^{(\mathrm{i})}) \neq 1$, the sum of the actual latent count parameters is not equal to the mean count, violating the first assumption in (2). In order to respect that expression, it is necessary to find an adjustment for the latent count estimates, so that (2) is observed. One way to adjust the parameters $\hat{\lambda}_{\mathrm{j}}^{(\mathrm{i})}, j \in J$ is to define another latent marginal count

$$\bar{\lambda}_{\mathrm{j}}^{(\mathrm{i})} := h(\Delta^{(\mathrm{i})}) \cdot \lambda_{\mathrm{j}}^{(\mathrm{i})}, \tag{14}$$

for some function $h(\cdot)$ that depends on $\Delta^{(\mathrm{i})}$. A multiplicative adjustment is reasonable as it increases the counts proportionally based on their previous value. The multiplicative constant can be derived as follows

$$\begin{aligned} \sum_{j=1}^{J} \bar{\lambda}_{\mathrm{j}}^{(\mathrm{i})} &= \lambda^{(\mathrm{i})}, \\ \sum_{j=1}^{J} [\lambda_{\mathrm{j}}^{(\mathrm{i})} \cdot h(\Delta^{(\mathrm{i})})] &= \lambda^{(\mathrm{i})}, \\ h(\Delta^{(\mathrm{i})}) \cdot \sum_{i=1}^{J} \lambda_{\mathrm{j}}^{(\mathrm{i})} &= \lambda^{(\mathrm{i})}. \end{aligned} \tag{15}$$

After replacing (13) in the last equality, it gets that

$$\begin{aligned} h(\Delta^{(\mathrm{i})}) \cdot e^{\Delta^{(\mathrm{i})}} \cdot \lambda^{(\mathrm{i})} &= \lambda^{(\mathrm{i})}, \\ h(\Delta^{(\mathrm{i})}) \cdot e^{\Delta^{(\mathrm{i})}} &= 1, \\ h(\Delta^{(\mathrm{i})}) &= e^{-\Delta^{(\mathrm{i})}}. \end{aligned} \tag{16}$$

In conclusion, $\bar{\lambda}_{\mathrm{j}}^{(\mathrm{i})} = e^{-\hat{\Delta}^{(\mathrm{i})}} \hat{\lambda}_{\mathrm{j}}^{(\mathrm{i})}$ is an adjustment that respects that the summation of the latent count estimates equals the total mean count, while maintaining the reasoning behind the entire modelling effort. The new estimates, $\bar{\lambda}_{\mathrm{j}}^{(\mathrm{i})}$, can then be used as proxies of the mean counts for each of the commodity types an individual purchase.

## 4. Data description and variable selection

The modelling methodology proposed in Chapters 2 and 3 was applied to estimate the number of shipments per commodity group from online shopping. In this context, online shopping/e-commerce is defined as any goods purchased online and delivered to consumers at home, to a pick-up point, or collected by consumers themselves in a store, warehouse, etc. All purchases in store are excluded, even though the purchase is reserved online in advance. Purchase of services not providing any shipments and online purchases between businesses (B2B) or consumers (C2C) are also excluded.

To deal with excess of zeros from non-shoppers, zero-inflated or hurdle models are commonly used alternatives, with the main difference being how they treat the process of obtaining the zeros in the data generating process (Mullahy 1986). The dataset in question considers a priori only one source of zeros: no online shopping. However, this might result from both individuals who never shop online and individuals who usually shop online, just not that particular time period. As the month in question is related to Christmas shopping (see sample description in the next subsection), it is likely that regular online shoppers did shop, justifying that zeros occurred from only one source. Additionally, the hurdle model, unlike the zero-inflated model, estimates the splitting between zero and non-zero counts, and the positive count data model in two stages. This allows for inclusion of variables that are only available for shoppers, of which are in abundance in the dataset. Hence, the authors chose to proceed with the hurdle model, despite both Wang and Zhou (2015) and Saphores and Xu (2020) used the Zero-inflated model for similar purposes in different applications.

### 4.1. Sample description

The dataset was shared with the researchers by a Norwegian logistic company. It was collected by a third party in January 2017 and contains information about online purchases in December 2016. The sample was collected through an internet panel survey among Norwegian respondents between 18 and 79 years of age with internet access (which in 2015 was 97% of the age segment (PostNord 2017)). Similar surveys are conducted on a regular basis by the logistics company itself as well as other companies to reveal consumer online shopping habits. The total sample consists of 1,515 respondents. Of these 1,515 respondents 1,019 answered that they had shopped online at least once in December 2016 and got follow-up questions about their purchase(s). At most, these respondents were asked 27 questions, including the total number of shipments from online shopping and types of commodity purchased, separating between 41 commodity types. For the remaining 496 respondents, only information about individual characteristics (demographics, socioeconomics factors and household characteristics) are available.

Out of the variables included in the analysis, 75 respondents (5%) have missing data in at least one variable due to non-response. Depending on the mechanism behind missing data, observations can be deleted, or values can be imputed. Deleting missing observations is accepted in the frequentist approach if both the missing and the observed data are at least random (Graham 2009; Rubin 1976). Although Afghari et al. (2019) show that when the data are missing completely at random or at least at random, the multiple imputation approach yields smaller standard errors and overall better goodness of fit than deletion, there is no guarantee that imputation will produce unbiased estimates of the missing variables. When the missing data points are in the dependent variable and these observations are missing at least at random, several imputation methods produce the same results as deletion (Allison 2000). Additionally, when the total number of missing data points is low, the imputation mechanism is of negligible relevance (Schafer 1999). For this research, the missing data are found in the variable for own income, the number of shipments received, main benefit of online shopping and preferred payment options when shopping online. A combination of mean tests (as presented by Enders (2010)) indicate that incomplete records in income are missing at random (somewhat more often for less educated people) while for the other variables, they are missing completely at random. As some of these correspond to missing values in the dependent variable and the frequency of individual records with at least one missing value is significantly low (5%), deletion was chosen, leaving 1440 respondents with full information for analysis.

### 4.2. Variable selection based on multiple correspondence analysis and literature review

To select variables for model estimation, the limited research on freight trip generation by household derived from online shopping was consulted. Wang and Zhou (2015) and Saphores and Xu (2020) have both estimated number of shipments from online shopping per household from the US 2009 and 2014 National Household Travel Survey. They showed that web use, education, income, age, race and lifestyle, including household composition impact both the probability to purchase online as well as the number of deliveries received at home. Gardrat et al. (2016) found tendencies towards an increase in both practice and number of deferred purchase and reception (DPR), covering deliveries to consumer from both online shopping and shopping trips, with the social profile of the head of the household. Interestingly, Gardrat et al. (2016) also distinguished between commodity groups and show the importance of distinguishing on commodity group. They found that groceries (including catering) were the most frequent commodity purchased and delivered to consumers, followed by clothing; high-tech items and culture; household appliances, furniture and other products, and last; healthcare and cosmetics. Other research on online shopping shows that income, internet use and experience, e-shopping attitude, potential for lower price, higher education, working full-time jobs or long hours, kids in the household and residency in urban areas have a positive impact on online shopping; while risk of security breach, household size, ethnicity and age have a negative impact (Farag et al. 2007; Cao, Xu, and Douma 2012; Zhou and Wang 2014; Limayem, Khalifa, and Frini 2000; Lohse, Bellman, and Johnson 2000).

The 41 commodity types were aggregated into five groups based on the degree of quality judgement online, risk of erroneous purchase and consumer characteristics, inspired by a multiple correspondence analysis (see Appendix B) and literature review, as follows:

*Children and leisure goods* include products about which the same quality judgement can be made online as in-store. No pre-knowledge of the product is necessary, and the combination of high quality judgement and low price makes it a low risk commodity for online purchases (Lowengart and Tractinsky 2001). The MCA indicates that buyers of these items suffer from 'time starvation'. The concept of 'time starvation' is inspired by Lohse, Bellman, and Johnson (2000) who associated it with people who worked many hours per week and had little time for physical shopping. In this paper, having kids under 15 years in the household and some distance to their closest urban areas are assumed to contribute to time starvation. The most frequent main benefit of online shopping for this group is better selection and lower price.

*Electronics* covers commodities about which consumers can make quality judgements based on information online, given that the customer has or gains some pre-knowledge about the product, and to a high degree meet customer's expectations at arrival. Hence online shopping induces relatively low risk of erroneous purchase or mismatch of expectations. The MCA showed that this consumer group has a predominance of male consumers, searching for the lowest price or a better selection. The same tendencies are found for electronics and computer products in Levin, Levin, and Weller (2005).

*Clothes, beauty and interior products* include commodities that need experience prior to a quality judgement, and as such, induce some risk of erroneous purchase. The MCA shows that the consumers are typically women and assess either time saving or flexibility to shop at a preferred time (for instance outside of regular opening hours) as main benefit of online shopping. The latter can be related to shopping interests, which is an attribute that is relatively more important for clothes and products of its like than other commodities (Levin, Levin, and Weller 2005). Consumers who want to enjoy shopping, tend to prefer offline to online shopping or add traditional shopping to online shopping (Levin, Levin, and Weller 2005; Zhou and Wang 2014).

*Consumables* are goods of which the quality cannot be known prior to consumption. Hence, the risk of buying a good that does not meet expectations is relatively high but can be reduced by multiple purchases of the same brand or from the same seller. The MCA indicates that these customers are usually women, have kids, assess either comfort or time savings as the main benefit of online shopping, and tend to live in urban areas. The combination of customer characteristics indicated that consumers purchasing consumables online are likely to suffer from 'time starvation'.

The last group, *other (miscellaneous)* commodities, includes commodities that do not fit in the previous four groups. These are (i) antiques and nutritional supplements, of which the quality cannot be assessed without careful inspection or long time experience (de Figueiredo 2000; Girard, Silverblatt, and Korgaonkar 2002); (ii) commodities that fall out of the 41 given commodity types and are included in the open alternative 'Other … ', typically tickets and trips; (iii) purchases where the consumer did not know or did not report the commodity type. This group has high variability and diversity. No joint assumptions can be made about quality judgement, risk of erroneous purchase or consumer characteristics.

A summary of the five aggregated commodity groups is presented in Table 1, followed by descriptive statistics of variables sought relevant for analysis in Table 2. The latter presents statistics in total and per commodity group. The variable *ship,* ranging from 0 (non-shoppers) to 21, is a combination of the binary variable for online shopping and the count variable for number of shipments from online shopping. The number of shipments after the value of 10 was originally reported through intervals, with values being greater than 21 in the last interval. These were recoded using the interval midpoint values. Hence, the variable exhibits censoring from 10 on (reported as intervals) and right censoring at 21. This affects few observations (for example, only two observations lie above 21), and no correction mechanism was applied. Income was also reported using intervals and recoded using midpoint values. The variable *Km* was generated based on consumers' zip code and a distance matrix between the zip code and its nearest urban area (centre zone) using QGIS

**Table 1.** Summary and examples of the five aggregated commodity groups based on MCA and literature review.

| Group | Quality judgement | Risk of erroneous purchase | Consumer characteristics | Example |
|---|---|---|---|---|
| Children and leisure goods | High | Low | Have kids, motivated by better selection and price online | CDs, board games, toys, magazines |
| Electronics | Medium/high | Low | Male, motivated by lower price and better selection online | Phones, digital games, kitchen and beauty appliances, TVs, PCs |
| Clothing, beauty and interior | Low/medium | Medium | Female, motivated by time savings and flexibility, have interest in shopping | Sweaters, Makeup, Shoes |
| Consumables | Low | Medium/high | Female, have kids, live in urban areas | Food and other groceries, lenses and glasses |
| Other goods | - | - | - | Antiques, nutritional supplements, tickets and trips |

(QGIS Development TEAM 2018). The centre zones are retrieved from Statistics Norway (https://kart.ssb.no/). They define a centre zone as one or more centre kernels (an area with at least 3 different main types of economic activities within 50 metres distance, where government administration, health and social services, or social and personal services must be present in addition to retail trade) and a 100-metre zone surrounding them (Statistics Norway 2019).

Descriptive statistics are included for the total sample, for buyers and non-buyers as well as for each of the five commodity groups. For the variables *age*, *male*, *kids*, *education* and *income,* average numbers for the Norwegian population were collected from Statistics Norway (2017b, 2017a, 2016a, 2016b). Note that the population statistics have some discrepancies from the sample. The main difference is that population numbers include both individuals with and without internet access. Other discrepancies are highlighted in Table 2.

## 5. Estimation results and implications

### 5.1. Estimation results

Table 3 presents estimation results of the model parameters given by the methodology proposed in Chapter 2 and the variables in Table 2. The results for the hurdle splitting (buy/no buy decision) and the parameters for the actual buyers are presented separately.

Starting with the zero-hurdle splitting model, the results show that higher income, education, number of kids and distance (km) to urban area are positively correlated with, and thus motivate, online shopping. As all the factors above imply less time for traditional shopping, the broader definition of time starvation as used in this paper might provide an explanation. Age and males are negatively related to the decision of buying online, despite male individuals being expected to have a slightly higher interest in technology than females. Gender and number of kilometres seem to be less relevant due to a high *p*-value. The results are in line with existing research on the subject, as presented in Chapter 4.

The main contribution of this modelling effort lies in the results from the count model split on commodity groups. The first thing to notice is that the negative binomial parameter

**Table 2.** Descriptive statistics (min/mean/max) for the variables included in the estimation as a total and for each model.

| | | Total[1] | Binary-outcome model | | Count data model | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Var. | Descriptions | Sample (Popul. '16) | Do not shop | Shop | Children and leisure | Electronics | Interior, clothing and beauty | Consumables | Other |
| Ship | Number of shipments (dependent variable) | Na | Na | 1/2.95/21 | 1/3.76/21 | 1/3.69/21 | 1/3.62/21 | 1/4.56/21 | 1/2.90/21 |
| Age | Age | 18/50/79 (18/46/79) | 18/57/79 | 18/47/78 | 18/47/78 | 19/46/75 | 18/44/78 | 20/46/77 | 18/48/78 |
| Inc | Personal annual gross income (in 1000 NOK) | 100/442/1,000 (100/432[4]/ 1,000) | 100/426/ 1,000 | 100/450/ 1,000 | 100/453/ 1,000 | 100/467/ 1,000 | 100/427/ 1,000 | 100/505/ 1,000 | 100/448/ 1,000 |
| Km | Number of kilometres from nearest urban area | 0.0/5.1/ 63.5 | 0.04/5.43/ 46.82 | 0.01/4.94/ 63.5 | 0.09/5.23/ 56.77 | 0.09/5.37/ 56.77 | 0.04/5.17/ 63.5 | 0.13/4.26/ 42.03 | 0.01/5.02/ 40.4 |
| Male | The respondent is a man (B) | 0.50 (0.50) | 0.51 | 0.50 | 0.49 | 0.71 | 0.35 | 0.45 | 0.52 |
| Kids | Have kids under 15 years in the household (B) | 0.20 (0.50[2]) | 0.101 | 0.26 | 0.31 | 0.27 | 0.30 | 0.33 | 0.24 |
| Heduc | Have university or college education (B) | 0.60 (0.33[3]) | 0.50 | 0.59 | 0.64 | 0.53 | 0.60 | 0.64 | 0.58 |
| Empl | Working fulltime or as self-employed (B) | 0.50 | 0.46 | 0.60 | 0.61 | 0.61 | 0.60 | 0.70 | 0.58 |

(*continued*).

**Table 2.** Continued.

| | | Total[1] | | Binary-outcome model | | Count data model | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Var. | Descriptions | Sample | (Popul. '16) | Do not shop | Shop | Children and leisure | Electronics | Interior, clothing and beauty | Consumables | Other |
| Price | Lower price is most important benefit of online shopping (B) | Na | | Na | 0.21 | 0.22 | 0.29 | 0.19 | 0.22 | 0.19 |
| Select | Better selection is most important … (B) | Na | | Na | 0.19 | 0.18 | 0.18 | 0.17 | 0.15 | 0.22 |
| Time | Time savings is most important … (B) | Na | | Na | 0.09 | 0.09 | 0.08 | 0.11 | 0.09 | 0.08 |
| Flex | Shop at preferred time is most important … (B) | Na | | Na | 0.35 | 0.36 | 0.31 | 0.38 | 0.34 | 0.36 |
| Comf | Comfort is most important … (B) | Na | | Na | 0.10 | 0.11 | 0.11 | 0.11 | 0.18 | 0.08 |
| N | Number of observations | 1440 | | 479 | 961 | 455 | 327 | 437 | 132 | 245 |

Note: For the count data portion, descriptive statistics are presented per commodity group covering only individuals who purchased from that group (for the variable "Ship", statistics about the total shipments from all commodity groups received by a buyer of the commodity group listed are shown). Variable names are partly given by authors. For binary variables (B) only the proportion of "ones"/"yes" outcomes is presented.

[1] Population equivalents are presented under parenthesis, when available; [2] Includes children at 15–17 years, and people younger than 18 years and older than 79 years with kids in the household; [3] Includes children at 16 and 17 years; [4] Total income includes social welfare transfers; Na = not applicable as the sample includes observations that did not receive this question.

**Table 3.** Results from estimation of a hurdle model with five commodity groups.

| | | Estimate | Std, Error | z value | Pr( > \|z\|) |
|---|---|---|---|---|---|
| Zero hurdle splitting model coefficients (binomial with logit link): | | | | | |
| | (Intercept) | 0.135 | 1.179 | 0.114 | 0.909 |
| | age | −0.041 | 0.004 | −9.432 | $< 2e^{-16}$*** |
| | log(Inc) | 0.191 | 0.096 | 1.980 | 0.048* |
| | Male | −0.074 | 0.122 | −0.606 | 0.545 |
| | Kids | 0.512 | 0.184 | 2.788 | 0.005** |
| | Higheduc | 0.324 | 0.122 | 2.648 | 0.008** |
| | Km | $1.21e^{-04}$ | 0.008 | 0.015 | 0.988 |
| Count model coefficients (truncated negbin with log link): | | | | | |
| Other | Constant | 0.334 | 0.068 | 4.920 | 8.63e-07*** |
| Electronics | Male | 0.444 | 0.090 | 4.942 | 7.73e-07*** |
| | Price | 0.155 | 0.130 | 1.186 | 0.236 |
| | Selection | 0.352 | 0.147 | 2.395 | 0.017* |
| Clothes, beauty and interior products | Male | −0.131 | 0.103 | −1.268 | 0.205 |
| | Log(Inc) | 0.047 | 0.008 | 5.964 | 2.47e-09*** |
| | Selection | 0.068 | 0.152 | 0.446 | 0.655 |
| | Time | 0.108 | 0.165 | 0.655 | 0.513 |
| | Flexibility | 0.138 | 0.116 | 1.190 | 0.234 |
| Consumable | Km | −0.024 | 0.012 | −2.053 | 0.040* |
| | Male | −0.315 | 0.168 | −1.879 | 0.060. |
| | Comfort | 0.374 | 0.208 | 1.799 | 0.072. |
| | Time | 0.131 | 0.284 | 0.461 | 0.645 |
| | Kids | 0.194 | 0.170 | 1.142 | 0.254 |
| | Log(Inc) | 0.050 | 0.011 | 4.455 | 8.39e-06*** |
| Children and leisure goods | Km | 0.016 | 0.005 | 3.086 | 0.002** |
| | Kids | 0.404 | 0.093 | 4.358 | 1.32e-05*** |
| | Price | 0.330 | 0.120 | 2.749 | 0.006** |
| | Selection | 0.305 | 0.132 | 2.308 | 0.021* |
| Negative binomial | Log(alpha) | 0.965 | 0.139 | 6.935 | 4.06e-12*** |

Aplha: count = 2.6239
Number of iterations in BFGS optimisation: 31
Log-likelihood: −2525 on 27 Df

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

for the variance (referred to as *Log(alpha)*) is significant, exhibiting the importance of allowing heterogeneity in the count data model. Among the covariates, *gender* indicates that males are more willing to buy electronic items than females, while females purchase more clothes, beauty and interior products, and consumables. This might be related to interests, and that women (still) tend to be more occupied with the household than men. Income has a positive influence for clothes, beauty and interior products, and consumable purchases; as these are commodities typically bought by females with a lower average income than males, income might represent time starvation as well as purchasing power. Having kids under the age of 15 in the household increases purchases of consumables, and children and leisure goods; probably reflecting the corresponding demand and time-starvation that comes with having kids. Distance to urban areas favours the acquisition of children and leisure goods via internet, but for consumables, the effect is the opposite. Both might be explained by lower service in rural areas than in cities. An alternative explanation is that people in rural areas have a higher propensity to travel by car and stop by the local supermarket to buy groceries and pick up parcels on their way. Variables for the perceived main benefit of online shopping are included and provide new associations for what is important for online shoppers buying different commodity types; perceiving lower prices or a greater selection

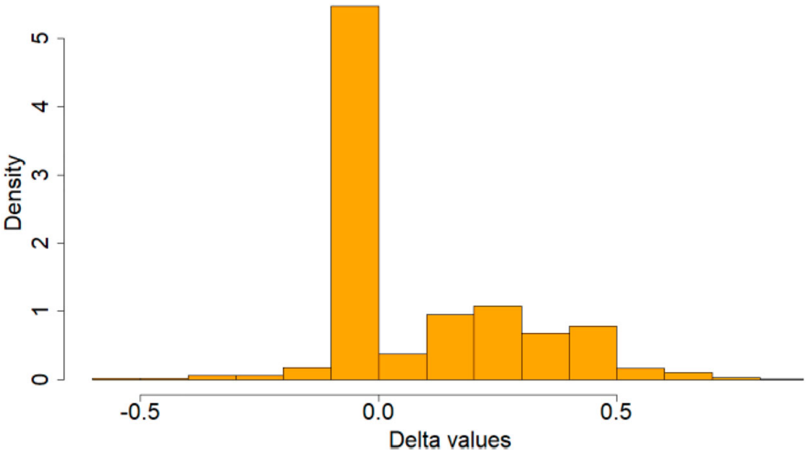


**Figure 1.** Histogram of computed delta values.

as the main benefits of online shopping motivates consumers to buy electronics and children and leisure items online. Flexibility to shop at any given time influences positively the online purchase of clothes, beauty and interior products. Time savings are conceived as an advantage for both consumables and clothes, beauty, and interior products. Comfort is associated with the online purchase of consumables.

### 5.2. Validation of the methodology for count splitting

The significant variables (with $p$-value $\leq 0.1$) in Table 3 are used to calculate the average number of shipments in total and per commodity group. To validate the latent count estimates for each commodity group, the delta values must be computed and a hypothesis system to determine if they are statistically equal to zero or not must be tested.

The corresponding delta values computed from the estimation are shown using a histogram in Figure 1. As it can be seen from Figure 1, the values for delta are relatively small, ranging from $-0.6$–0.9 with an average value of 0.11, which means that, in average, the summation of the latent marginal counts corresponds to $e^{0.11} \approx 1.12$times the estimated total count (12% larger). In terms of median performance, the median value of the deltas is equal to zero. This suggests that the discrepancies between the observed counts and the proposed split are minor.

The histogram for the standardised delta values (after using a first-order approximation to compute the variance-covariance structure and using it to decorrelate and standardise them according to Appendix A) is presented in Figure 2. A normality test using the Jarque-Bera method was conducted with a $p$-value of 0.734, allowing to conclude that the normality assumption is valid (at a 5% significance level). Figure 2 also includes the overlaid curve of a normal density function with parameters provided by the standardised delta values to see the fit graphically. However, the $p$-value for the t-test establishes that there is evidence to reject the zero-mean hypothesis ($p$-value $= 0.008$).

Given the outcome of the hypothesis testing and the low value of the original delta values (presented in Figure 1), the adjustment presented in Subsection 3.2 was considered to

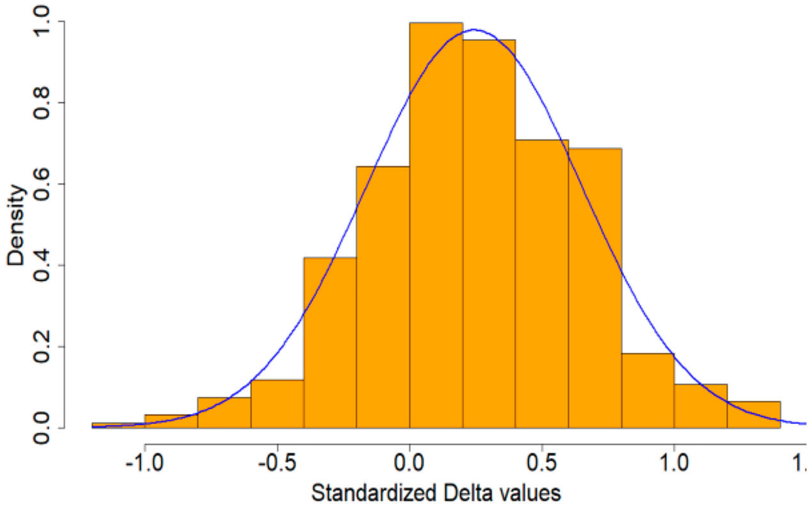


**Figure 2.** Histogram of standardised deltas.

be a reasonable approach to account for non-null, yet small, delta values in the estimate of the latent marginal counts.

A drawback with the values (number of shipments) presented in this analysis is its reliance on information of online shopping in December only. December might not be representative of online shopping as it includes Christmas shopping and holidays. However, a consumer survey reveals that as much as 67% of the population between 18 and 79 years old (with internet access) shop online at least once a month (Postnord 2020).

## 6. Application: prospective analysis of the impact of consolidation

### 6.1. Number of shipments per commodity group

Using the hurdle model (with only significant variables) and the splitting methodology, it was found that, on average, an individual attracts 2.4 shipments per month (0.077 shipments per day, considering that the data collecting process was conducted in December, which has 31 days). *Clothes, beauty and interior products* and *Children and leisure products* have the highest number of shipments with 0.77 shipments/person-month (0.025 shipments/person-day) and 0.64 shipments/person-month (0.02 shipments/person-day) respectively. They are followed by *Electronics* (0.46 shipments/person-month, or 0.015 shipments/person-day), *Other* (0.31 shipments/person-month, or 0.01 shipments/person-day), and lastly, *Consumables* (0.23 shipments/person-month, or 0.0075 shipments/person-day).

Table 4 presents the average number of shipments per person-month for each commodity group according to population segments for *age*, *income*, *gender* and *kids*. Some key findings are that consumers with kids have the highest average number of total shipments, followed by consumers in the second highest income interval and consumers in the age group 38–47 years. Females receive almost twice as many shipments of consumables, clothes, beauty and interior products than men, while men purchase electronics to a much

**Table 4.** Average number of shipments per person per month for each commodity group according to different population segmentations (age, income, gender, presence of kids).

| Age | Electronics | Interior, clothing and beauty products | Consumables | Children and leisure products | Other goods | Row total |
|---|---|---|---|---|---|---|
| 18–27 | 0.46 | 0.78 | 0.09 | 0.48 | 0.31 | 2.12 |
| 28–37 | 0.48 | 0.97 | 0.33 | 0.79 | 0.27 | 2.85 |
| 38–47 | 0.52 | 0.91 | 0.34 | 0.89 | 0.30 | 2.96 |
| 48–57 | 0.54 | 0.72 | 0.25 | 0.62 | 0.29 | 2.42 |
| 58–67 | 0.35 | 0.61 | 0.18 | 0.42 | 0.36 | 1.92 |
| 68–78 | 0.36 | 0.50 | 0.09 | 0.52 | 0.34 | 1.81 |
| Group average | 0.46 | 0.77 | 0.23 | 0.64 | 0.31 | 2.40 |
| **Income (NOK)*** | **Electronics** | **Interior, clothing and beauty products** | **Consumables** | **Children and leisure products** | **Other goods** | **Row total** |
| 100,000 | 0.42 | 0.73 | 0.11 | 0.53 | 0.32 | 2.11 |
| 250,000 | 0.39 | 0.85 | 0.26 | 0.65 | 0.35 | 2.49 |
| 349,999 | 0.43 | 0.75 | 0.19 | 0.55 | 0.30 | 2.22 |
| 450,000 | 0.45 | 0.81 | 0.26 | 0.68 | 0.26 | 2.46 |
| 549,999 | 0.40 | 0.79 | 0.28 | 0.65 | 0.37 | 2.48 |
| 650,000 | 0.60 | 0.73 | 0.19 | 0.69 | 0.28 | 2.49 |
| 749,999 | 0.41 | 0.92 | 0.19 | 0.69 | 0.24 | 2.46 |
| 899,999 | 0.79 | 0.86 | 0.30 | 0.76 | 0.33 | 3.04 |
| 1000000 | 0.64 | 0.32 | 0.43 | 0.70 | 0.36 | 2.45 |
| Group average | 0.46 | 0.77 | 0.23 | 0.64 | 0.31 | 2.40 |
| **Gender** | **Electronics** | **Interior, clothing and beauty products** | **Consumables** | **Children and leisure products** | **Other goods** | **Row total** |
| Female | 0.16 | 0.98 | 0.30 | 0.62 | 0.30 | 2.36 |
| Male | 0.75 | 0.56 | 0.16 | 0.65 | 0.32 | 2.44 |
| Group average | 0.46 | 0.77 | 0.23 | 0.64 | 0.31 | 2.40 |
| **Kids** | **Electronics** | **Interior, clothing and beauty products** | **Consumables** | **Children and leisure products** | **Other goods** | **Row total** |
| No | 0.44 | 0.68 | 0.19 | 0.46 | 0.31 | 2.09 |
| Yes | 0.51 | 1.00 | 0.34 | 1.13 | 0.30 | 3.28 |
| Group average | 0.46 | 0.77 | 0.23 | 0.64 | 0.31 | 2.40 |

Max (and min) in each column are highlighted.
*Midpoint for each income interval, as reported in the original dataset.

higher extent than females (almost five times). Measured in total number of shipments, the gender difference is small.

## 6.2. Freight trip implications of home deliveries per commodity type

For establishments, industry grouping is a popular indicator of heterogeneity in freight trip generation estimates. Similarly, splitting home deliveries in commodity groups can explain some of the heterogeneity in freight transport generated by consumers. This section provides an example of freight trip reduction opportunities by exploring correlation and thus consolidation opportunities between commodity types purchased. Correlations are explored using linear regression models for purchase behaviour between two commodity groups for respondents with and without kids. The methodology and correlations are presented in Appendix C.

There are several implications derived from the correlation analysis. First, shipments of consumables and clothes are positively related in almost all combinations. The size varies, with the largest value corresponding to people 37 years or younger, with kids, and in the

**Table 5.** Analysis of the impact of consolidation for three boroughs in Oslo, based on the linear regression coefficients previously introduced and presented in Appendix C.

| | Sector | | Number of shipments per commodity | | | | | Total shipments | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Borough | Income (in 1000) | Age | Electr. | Clothing | Consum. | Children | Others | No consolid. | Full consolid. | Relative gap |
| Nordstrand | 550–750 | 37–57 | 12,522 | 24,043 | 6,418 | 18,340 | 9,751 | 71,075 | 59,419 | 16.40% |
| Vestre Aker | > 750 | 37–57 | 12,105 | 23,672 | 6,365 | 17,972 | 9,470 | 69,584 | 50,575 | 27.32% |
| Ullern | > 750 | 37–57 | 8,365 | 16,299 | 4,379 | 12,199 | 6,544 | 47,787 | 34,705 | 27.38% |

lowest income group. For people with kids, a general positive correlation for consumables and children and leisure products also exists, except for individuals with both low age and high income. The linear regression between consumables and clothes, interior and beauty products shows that for several consumer types, as many as 50% of consumers who received shipments of consumables, also received shipments of clothes, interior and beauty products the same month. Similarly, up to 42% of consumers receiving shipments of children and leisure goods also received shipments of consumables. This result suggests consolidation opportunities between consumables (like groceries) and clothing products, beauty and interior products; and/or children and leisure goods.

For consolidation of different types of commodities to work in practice, there must be either a sufficiently large or homogenous demand. Statistics for boroughs in Oslo (the capital of Norway) shows clear patterns regarding demographics (see Appendix D for a table of descriptive statistics and a map of the boroughs). Suburban residents with medium to large households, high income, medium age and low share of people living in apartments (like the boroughs *Nordstrand, Vestre Aker* and *Ullern* in Appendix D) are areas that the model identifies with a potentially large demand of the commodity groups in question. The low share of people living in apartments indicate a relatively low density, and hence the need to consolidate shipments to achieve small routes with high vehicle load factor. Table 5 shows the potential impact that a best-case consolidation effort, that is a full consolidation scenario according to the consolidation opportunities elaborated on above, applied to the three boroughs in Oslo might have (in terms of total shipments). Total counts per commodity group are estimated using an average individual profile derived from the information presented in Appendix D and the slope coefficients for the population segment in each borough (assumed to be predominantly composed of households with kids). Considerable consolidation opportunities exist, and it can be seen that the number of separate shipments can be reduced significantly if knowledge about the demand profiles is used for classification, and cooperative efforts to develop multi-sector and multi-company partnerships are supported.

In this paper, consolidation opportunities are found between commodities with different transport and delivery requirements: the transportation operation and characteristics for children and leisure goods, and clothes, interior and beauty products is flexible compared to consumables, with strict requirements regarding transport, handling and delivery (Lagorio and Pinto 2020; Saphores and Xu 2020). By consolidating shipments from these commodity groups, multiple freight trips to a given household may be avoided and the amount of failed deliveries reduced. As 1 in every 20 orders are not delivered on their first attempt (Loqate 2018), failed deliveries generates both economic losses and unnecessary freight trips, and motivates for instance e-grocers to experiment with alternative deliveries increasing the probability of successful delivery (Saphores and Xu 2020). Consolidation might also contribute to reduced operating costs, which falls with vehicle load factor and drop density per round trip (Allen et al. 2018).

With online shoppers pushing for faster and cheaper deliveries, new trends like 'instant deliveries' or 'one-day rush' reduce the ability of logistic firms to plan for deliveries with a high degree of consolidation and low road freight externalities. This research shows how unfortunate that is, as there are unused consolidation opportunities in the last-mile delivery market, both within and between commodity groups. It should be stressed that the time dimension is not considered in this example. For consolidation to work in practice, consumers must be given the opportunity to pile up their orders and have them jointly delivered. To increase consumer acceptance of delivery time, transport companies can work together with online retailers and grocers to inform their clients about the benefits of consolidation (reduced costs and externalities) or even offer discounts or other benefits for those who choose to consolidate their purchases. In any case, consolidation of shipments to consumers could be further researched and explored by both transport companies and public authorities to reveal its potential to reduce externalities.

## 7. Conclusions

This paper builds on the econometric model to estimate latent marginal counts based on an observable total count by Afghari et al. (2016); Afghari et al. (2018), and complements it with a statistical test to assess its validity. The model is applied to an e-commerce dataset to estimate online shopping behaviour given by the propensity to shop online and corresponding number of shipments in total and applies the splitting methodology to infer the marginal counts of five different commodity groups of interest. The results are assessed in terms of concordance of the purchase profiles obtained and the literature consulted, and the potential freight trip implications of the results, exploiting the co-dependence structure between the demands for different types of commodities. Accordingly, the novelty in this research are (i) the incorporation of available information of purchase profiles to guide the estimation of latent marginal counts (ergo, allowing for potential null marginal counts), (ii) proposing a statistical test to measure the validity of the assumptions made in the modelling process and a correction mechanism, based on the evidence provided by the data, in case the deviations from the assumptions are minor, (iii) estimating number of shipments per commodity and hence revealing some of the heterogeneity in the freight trip attraction to consumers, and (iv) suggesting how knowledge of commodity groups and purchase behaviour can help policy makers and transport companies achieve more sustainable freight transport in urban areas through consolidation.

The outcome of the methodology proposed, and the results obtained for the case study of online shopping, shows a promising development in inferring non-observable features, and the implications that these results might have for the different stakeholders and decision makers involved. This research effort is particularly timely, especially during this unprecedented time of pandemic situation, in which online shopping and household deliveries have become a new standard, in order to avoid life-endangering situations. However, more studies should be undertaken to strengthen the findings in this paper in other context.

## Acknowledgements

The authors would like to thank the Norwegian logistics company for sharing their data.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Elise Caspersen's work is undertaken as part of the research project 250432 NORSULP (Sustainable Urban Logistics Plans in Norway), financed by the Research Council of Norway and the Norwegian Public Roads Administration. Further information is available from www.norsulp.no (in Norwegian only).

## References

Afghari, Amir Pooyan, M. Mazharul Haque, Simon Washington, and Tanya Smyth. 2016. "Bayesian Latent Class Safety Performance Function for Identifying Motor Vehicle Crash Black Spots." *Transportation Research Record* 2601 (1): 90–98. doi:10.3141/2601-11.

Afghari, Amir Pooyan, Simon Washington, Md Mazharul Haque, and Zili Li. 2018. "A Comprehensive Joint Econometric Model of Motor Vehicle Crashes Arising from Multiple Sources of Risk." *Analytic Methods in Accident Research* 18: 1–14. doi:10.1016/j.amar.2018.03.002.

Afghari, Amir Pooyan, Simon Washington, Carlo Prato, and Md Mazharul Haque. 2019. "Contrasting Case-Wise Deletion with Multiple Imputation and Latent Variable Approaches to Dealing with Missing Observations in Count Regression Models." *Analytic Methods in Accident Research* 24. doi:10.1016/j.amar.2019.100104.

Allen, J., M. Piecyk, M. Piotrowska, F. McLeod, T. Cherrett, K. Ghali, T. Nguyen, et al. 2018. "Understanding the Impact of e-Commerce on Last-Mile Light Goods Vehicle Activity in Urban Areas: The Case of London." *Transportation Research Part D: Transport and Environment* 61: 325–338. doi:10.1016/j.trd.2017.07.020.

Allison, Paul D. 2000. "Multiple Imputation for Missing Data: A Cautionary Tale." *Sociological Methods & Research* 28 (3): 301–309. doi:10.1177/0049124100028003003.

Cao, Xinyu Jason, Zhiyi Xu, and Frank Douma. 2012. "The Interactions Between e-Shopping and Traditional in-Store Shopping: an Application of Structural Equations Model." *Transportation* 39 (5): 957–974. doi:10.1007/s11116-011-9376-3.

Castro, Marisol, Rajesh Paleti, and Chandra R. Bhat. 2012. "A Latent Variable Representation of Count Data Models to Accommodate Spatial and Temporal Dependence: Application to Predicting Crash Frequency at Intersections." *Transportation Research Part B: Methodological* 46 (1): 253–272. doi:10.1016/j.trb.2011.09.007.

Dablanc, Laetitia, Eleonora Morganti, Niklas Arvidsson, Johan Woxenius, Michael Browne, and Neïla Saidi. 2017. "The Rise of on-Demand 'Instant Deliveries' in European Cities." *Supply Chain Forum: An International Journal* 18 (4): 203–217. doi:10.1080/16258312.2017.1375375.

de Figueiredo, John M.. 2000. *Finding Sustainable Profitability in the E-commerce Continuum*. MIT Sloan Management Review.

Ecommerce News. "Ecommerce in Europe: €717 billion in 2020.".

Enders, Craig K. 2010. *Applied Missing Data Analysis*. New York, USA: The Guilford Press.

Farag, Sendy, Tim Schwanen, Martin Dijst, and Jan Faber. 2007. "Shopping Online and/or in-Store? A Structural Equation Model of the Relationships Between e-Shopping and in-Store Shopping." *Transportation Research Part A: Policy and Practice* 41 (2): 125–141. doi:10.1016/j.tra.2006.02.003.

Gardrat, Mathieu, Florence Toilier, Danièle Patier, and Jean-Louis Routhier. 2016. "The Impact of New Practices for Supplying Households in Urban Goods Movements: Method and First Results. An application for Lyon, France." In *VREF conference on Urban Freight* 2016. Göteborg, Sweden.

Girard, Tulay, Ronnie Silverblatt, and Pradeep Korgaonkar. 2002. "Influence of Product Class on Preference for Shopping on the Internet." *Journal of Computer-Mediated Communication* 8 (1). doi:10.1111/j.1083-6101.2002.tb00162.x.

Graham, John W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60 (1): 549–576. doi:10.1146/annurev.psych.58.110405.085530.

Greene, William H. 2000. Econometric Analysis 4th Edition. *International Edition*, 201–215. New Jersey: Prentice Hall.

Heydari, Shahram, Liping Fu, Luis F Miranda-Moreno, and Lawrence Jopseph. 2017. "Using a Flexible Multivariate Latent Class Approach to Model Correlated Outcomes: A Joint Analysis of Pedestrian and Cyclist Injuries." *Analytic Methods in Accident Research* 13: 16–27.

Holguín-Veras, José, Stacey Hodge, Jeffrey Wojtowicz, Caesar Singh, Cara Wang, Miguel Jaller, Felipe Aros-Vera, et al. 2018. "The New York City Off-Hour Delivery Program: A Business and Community-Friendly Sustainability Program." *Interfaces* 48 (1): 70–86. doi:10.1287/inte.2017.0929.

Kessy, Agnan, Alex Lewin, and Korbinian Strimmer. 2018. "Optimal Whitening and Decorrelation." *The American Statistician* 72 (4): 309–314. doi:10.1080/00031305.2016.1277159.

Lagorio, Alexandra, and Roberto Pinto. 2020. "Food and Grocery Retail Logistics Issues: A Systematic Literature Review." *Research in Transportation Economics*. doi:10.1016/j.retrec.2020.100841.

Levin, Aron M., Irwin Levin, and Joshua Weller. 2005. "*A Multi-Attribute Analysis of Preferences for Online and Offline Shopping: Differences Across Products, Consumers, and Shopping Stages*." *Journal of Electronic Commerce Research* 6: 281–290.

Limayem, M., M. Khalifa, and A. Frini. 2000. "What Makes Consumers buy from Internet? A Longitudinal Study of Online Shopping." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30 (4): 421–432. doi:10.1109/3468.852436.

Lohse, Gerald L., Steven Bellman, and Eric J. Johnson. 2000. "Consumer Buying Behavior on the Internet: Findings from Panel Data." *Journal of Interactive Marketing* 14 (1): 15–29. doi:10.1002/(SICI)1520-6653(200024)14:1 < 15::AID-DIR2 > 3.0.CO;2-C.

Loqate. 2018. "Fixing Failed Deliveries." In *Improving Data Quality in Retail*. https://www.loqate.com/resources/thank-you/data-quality-report/?submissionGuid = e5417777-de99-4620-b51c-f1627e4d8027.

Lowengart, Oded, and Noam Tractinsky. 2001. "Differential Effects of Product Category on Shoppers' Selection of Web-Based Stores: A Probabilistic Modeling Approach." *J. Electron. Commerce Res.* 2: 142–156.

Mullahy, John. 1986. "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics* 33 (3): 341–365. doi:10.1016/0304-4076(86)90002-3.

Ni, Linglin, Xiaokun Wang, and Dapeng Zhang. 2016. "Impacts of Information Technology and Urbanization on Less-Than-Truckload Freight Flows in China: An Analysis Considering Spatial Effects." *Transportation Research Part A: Policy and Practice* 92: 12–25. doi:10.1016/j.tra.2016.06.030.

Park, Byung-Jung, and Dominique Lord. 2009. "Application of Finite Mixture Models for Vehicle Crash Data Analysis." *Accident Analysis & Prevention* 41 (4): 683–691.

Pettersson, Fredrik, Lena Winslott Hiselius, and Till Koglin. 2018. "E-commerce and Urban Planning – Comparing Knowledge Claims in Research and Planning Practice." *Urban, Planning and Transport Research* 6 (1): 1–21. doi:10.1080/21650020.2018.1428114.

PostNord. 2017. "Netthandel i Norden 2017." *Norden - en digitalisert region: Slik ser nettkjøpsatferden ut i Norden i 2017*.

PostNord. 2020. "Netthandel i Norden - Oppsummering 2019." In *Netthandel i Norden*, edited by PostNord. PostNord.

QGIS Development TEAM. 2018. "QGIS Geographic Information System. Open Source Geospatial Foundation Project." In.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–592. doi:10.1093/biomet/63.3.581.

Saphores, Jean-Daniel, and Lu Xu. 2020. "E-shopping Changes and the State of E-Grocery Shopping in the US - Evidence from National Travel and Time use Surveys." *Research in Transportation Economics*. doi:10.1016/j.retrec.2020.100864.

Schafer, Joseph L. 1999. “Multiple Imputation: a Primer.” *Statistical Methods in Medical Research* 8 (1): 3–15. doi:10.1177/096228029900800102.

Sen, P. K., and J. M. Singer. 2017. *Large Sample Methods in Statistics (1994): An Introduction with Applications*. Boca Raton, FL: Chapman & Hall/CRC.

Statistics Norway. 2016a. “07778: Registered Incomes for Residents Persons (mill.NOK) 2006-2017.” In *Income and wealth statistics for households*, edited by Statistics Norway.

Statistics Norway. 2016b. “08921: Educational Attainment, by County, Age and Sex (C) 1980-2018.” In *Educational attainment of the population*, edited by Statistics Norway.

Statistics Norway. 2017a. “06071: Persons, by Type of Household, Contents and Year.” In *Families and households*, edited by Statistics Norway.

Statistics Norway. 2017b. “07459: Population, by Sex and One-Year Age Groups (M) 1986-2019.” In, edited by Statistics Norway.

Statistics Norway. 2019. “Concept Variable Centre Zone.” Accessed Oktober 4th. https://www.ssb.no/a/metadata/conceptvariable/vardok/2598/en.

Stopher, Peter. 2012. *Collecting, Managing, and Assessing Data Using Sample Surveys*. Cambridge: Cambridge University Press.

Visser, Johan, Toshinori Nemoto, and Michael Browne. 2014. “Home Delivery and the Impacts on Urban Freight Transport: A Review.” *Procedia - Social and Behavioral Sciences* 125: 15–27. doi:10.1016/j.sbspro.2014.01.1452.

Wang, Xiaokun, and Yiwei Zhou. 2015. “Deliveries to Residential Units: A Rising Form of Freight Transportation in the U.S.” *Transportation Research Part C: Emerging Technologies* 58: 46–55. doi:10.1016/j.trc.2015.07.004.

Washington, Simon P., Matthew G. Karlaftis, and Fred L. Mannering. 2003. *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Yasmin, Shamsunnahar, and Naveen Eluru. 2016. “Latent Segmentation Based Count Models: Analysis of Bicycle Safety in Montreal and Toronto.” *Accident Analysis & Prevention* 95: 157–171.

Zhou, Yiwei, and Xiaokun Wang. 2014. “Explore the Relationship Between Online Shopping and Shopping Trips: An Analysis with the 2009 NHTS Data.” *Transportation Research Part A: Policy and Practice* 70: 1–9. doi:10.1016/j.tra.2014.09.014.

Zou, Yajie, Yunlong Zhang, and Dominique Lord. 2013. “Application of Finite Mixture of Negative Binomial Regression Models with Varying Weight Parameters for Vehicle Crash Data Analysis.” *Accident Analysis & Prevention* 50: 1042–1051.

# Appendices

## Appendix A: First-order approximation of the covariance structure

In order to characterise the association structure of the delta estimators, an approximation of the second-order moments for the vector $\vec{\Lambda}$, through a stochastic first-order Taylor expansion of its components, was utilised. The first-order Taylor expansion for $\hat{\Delta}^{(\mathrm{i})}$around $\beta$ corresponds to

$$\hat{\Delta}^{(\mathrm{i})} = k(\hat{\beta}, X_{\mathrm{i}}) = k(\beta, X_{\mathrm{i}}) + \nabla_{\beta} k(\beta, X_{\mathrm{i}})^{T} (\hat{\beta} - \beta) + o_{\mathrm{p}}(||\hat{\beta} - \beta||), \tag{A17}$$

where $\nabla_{\beta}$ is the gradient operator, $o_{\mathrm{p}}(\cdot)$ is the probabilistic equivalent of the little o-notation for random objects and $||\cdot||$ the Euclidean norm (Sen and Singer 2017). The variance of $\hat{\Delta}^{(\mathrm{i})}$, based on this first-order approximation, can be computed as

$$\mathrm{var}(\hat{\Delta}^{(\mathrm{i})}) \simeq \nabla_{\beta} k(\beta, X_{\mathrm{i}})^{T} \cdot \Sigma_{\hat{\beta}} \cdot \nabla_{\beta} k(\beta, X_{\mathrm{i}}), \tag{A18}$$

with $\Sigma_{\hat{\beta}}$ being the variance-covariance matrix of the estimators $\hat{\beta}$. Since the true vector of parameters is not observed, this variance has to be estimated by replacing all the parameters involved in (18) with their corresponding estimates

$$\tau_{\mathrm{ii}} := \widehat{\mathrm{var}}(\hat{\Delta}^{(\mathrm{i})}) \simeq \nabla_{\beta} k(\hat{\beta}, X_{\mathrm{i}})^{T} \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla_{\beta} k(\hat{\beta}, X_{\mathrm{i}}). \tag{A19}$$

The estimated variance and covariance matrix, $\hat{\Sigma}_{\hat{\beta}}$, can be easily obtained from the maximum likelihood estimation of the count data model as the inverse of the observed Fisher's information matrix. The elements of the gradient can be computed as

$$\frac{\partial k}{\partial \beta_{\mathrm{l},\mathrm{K}_\mathrm{l}}}(\hat{\beta}, X_\mathrm{i}) = -\left[\frac{\sum_{j\in J\setminus\{l\}} \hat{\lambda}_\mathrm{j}^{(\mathrm{i})}}{\sum_{j\in J} \hat{\lambda}_\mathrm{j}^{(\mathrm{i})}}\right] x^{(\mathrm{i})}_{\mathrm{l},\mathrm{K}_\mathrm{l}} I^{(\mathrm{i})}_\mathrm{l}, \tag{A20}$$

where $x^{(\mathrm{i})}_{\mathrm{l},\mathrm{K}_\mathrm{l}}$ is the is the $K_\mathrm{l}$-th covariate associated with the $l$-th commodity type for individual $i$, $I^{(\mathrm{i})}_\mathrm{l}$ is the indicator variable for the $l$-th commodity type for individual $i$, and $\beta_{\mathrm{l},\mathrm{K}_\mathrm{l}}$ is their corresponding parameter. The covariance between any two components can also be estimated using the approximation in (17) as

$$\tau_{\mathrm{ij}} := \widehat{\mathrm{cov}}(\hat{\Delta}^{(\mathrm{i})}, \hat{\Delta}^{(\mathrm{j})}) \simeq \nabla_\beta k(\hat{\beta}, X_\mathrm{i})^T \cdot \hat{\Sigma}_{\hat{\beta}} \cdot \nabla_\beta k(\hat{\beta}, X_\mathrm{j}). \tag{A21}$$

Then, the matrix $\mathrm{T} := (\tau_{\mathrm{ij}})_{\mathrm{i,j}}$ is an estimate of the variance-covariance matrix of the vector $\vec{\Lambda}$ that can be used to break the correlation between its components by means of the transformation

$$\vec{\delta} = \mathrm{T}^{-1/2}\vec{\Lambda}, \tag{A22}$$

because

$$\mathrm{cov}(\vec{\delta}) = \mathrm{T}^{-1/2}\mathrm{cov}(\vec{\Lambda})\mathrm{T}^{-1/2} \simeq \mathrm{T}^{-1/2}\mathrm{T}\mathrm{T}^{-1/2} = I_{\tilde{\mathrm{N}}\times\tilde{\mathrm{N}}}; \tag{A23}$$

being $I_{\tilde{\mathrm{N}}\times\tilde{\mathrm{N}}}$ the identity matrix of size $\tilde{N}$. In the case that T is not a positive definite matrix (it is non-invertible), a similar result using the Moore-Penrose generalised inverse of T,$\mathrm{T}^+$, can be implemented

$$\vec{\delta} = (\mathrm{T}^+)^{1/2}\vec{\Lambda}, \tag{A24}$$

as long as $(\mathrm{T}^+)^{1/2}\mathrm{T}(\mathrm{T}^+)^{1/2} \approx I_{\tilde{\mathrm{N}}\times\tilde{\mathrm{N}}}$, which means that the variance-covariance matrix of $\vec{\delta}$ is close to the identity. There are other decorrelation matrices different from the proposed ones. For further details, please consult Kessy, Lewin, and Strimmer (2018).

## Appendix B: multiple correspondence analysis

Figure B1 summarises the results from the multiple correspondence analysis (MCA) on the 41 commodity groups. The axes are interpreted in terms of attributes of the online shopping experience inspired by Lowengart and Tractinsky (2001); Levin, Levin, and Weller (2005); Zhou and Wang (2014); de Figueiredo (2000); Girard, Silverblatt, and Korgaonkar (2002). Dimension 1 could be associated with the risk (or the propensity) of unmet expectations or erroneous purchase. A high (positive) value suggests a high risk of making purchases that do not meet the buyers' expectations. For instance, some groceries have a high risk of not meeting the expectations because of freshness, transporting and handling conditions; while electronical products do not present that risk due to standard manufacturing procedures and vast marketing campaigns to inform the consumers. Dimension 2 can be interpreted as the amount of knowledge a buyer possesses about a product prior to online purchase. A high value means that the buyer is informed about the characteristics of the product he/she is about to buy. Once again, groceries have a high value in this dimension since the majority of individuals is familiar with the standard expected quality and appearance of the food they are about to purchase. Electronics is also a well-informed purchase, as opposed to clothes, beauty and interior products, of which qualities like texture and colour are important aspects and difficult to describe with accuracy on the internet.

Even though the MCA itself did not allow to reduce the dimensionality of the dataset due to low percentages of explained variability in the first components, it was of great help in motivating the number and composition of the groups.

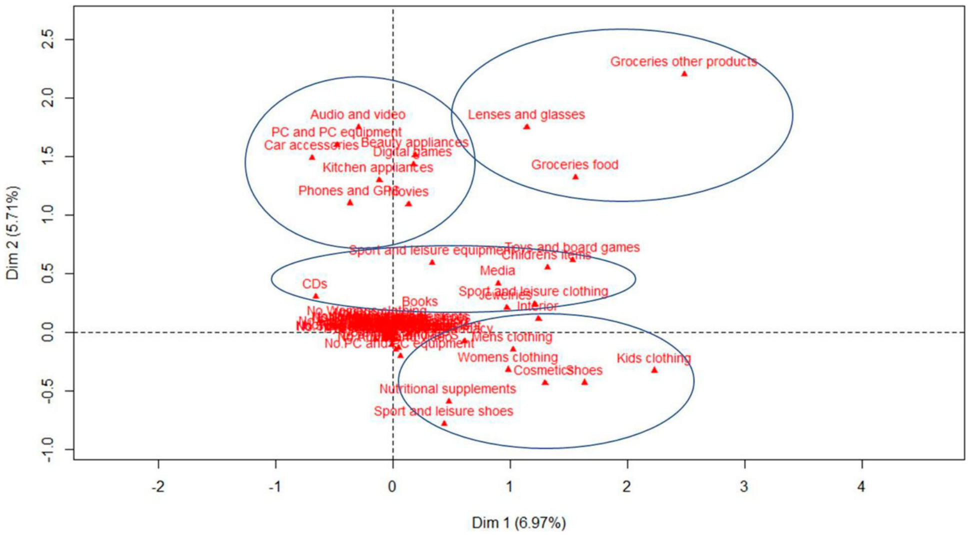


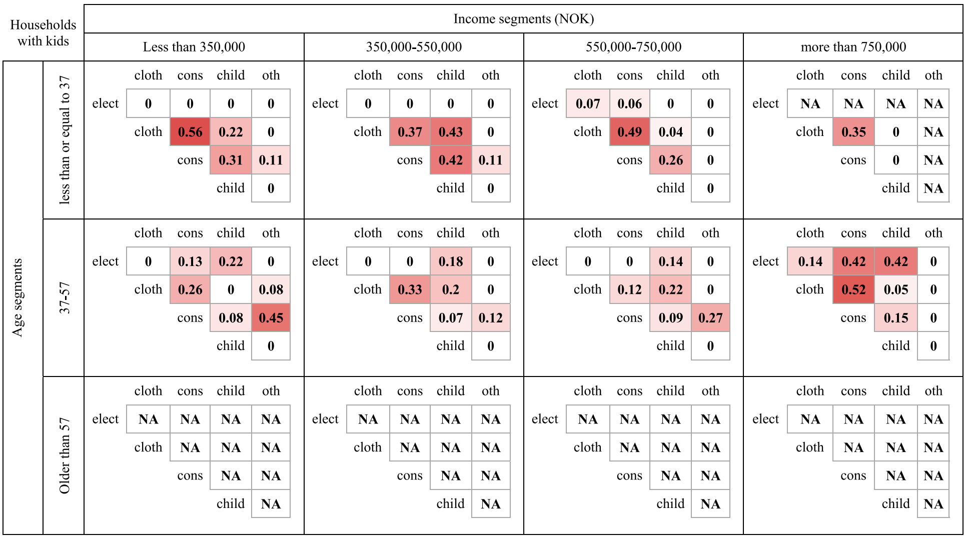
**Figure B1.** MCA results for the commodity clustering.

## Appendix C: correlation between commodity groups

Figure C1 and Figure C2 present positive correlation (red-coloured cells, with a stronger colour denoting a stronger correlation) and the slope of a linear regression model for purchase behaviour between two commodity groups (numbers) for respondents with and without kids. The estimated latent counts are used as input for this calculation based on the expression

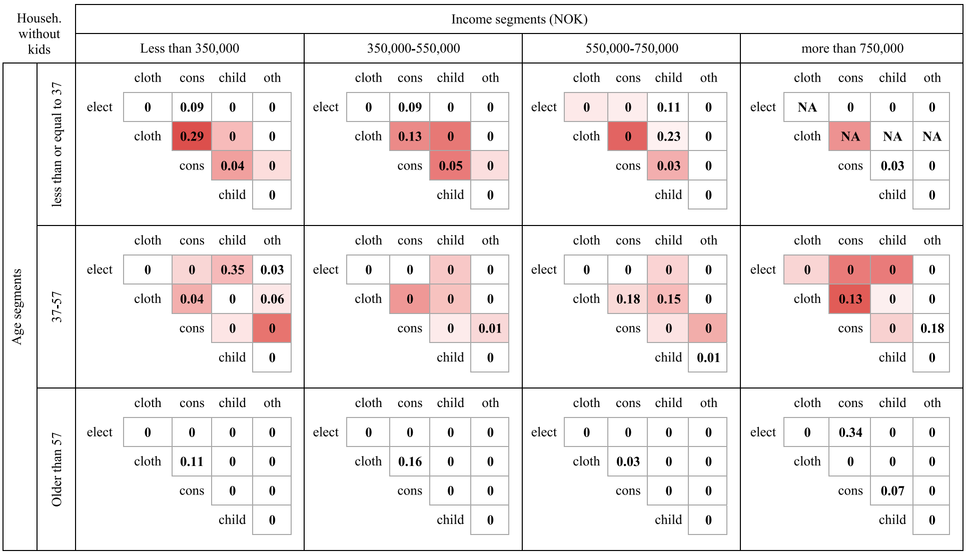
$$\hat{\beta}_{\mathrm{y|x}} = \hat{\rho}(x,y)\frac{\hat{\sigma}_{\mathrm{y}}}{\hat{\sigma}_{\mathrm{x}}}, \tag{A25}$$

where $\hat{\beta}_{\mathrm{y|x}}$ is the slope of the regression of shipments for commodity type*y* given shipments of commodity type $x$, $\hat{\rho}(x,y)$ is the estimate of Pearson's correlation coefficient for the two commodity groups and $\hat{\sigma}_{\mathrm{y}}, \hat{\sigma}_{\mathrm{x}}$ are their corresponding (estimated) standard deviations. The coefficients are not symmetric, i.e. $\hat{\beta}_{\mathrm{y|x}} \neq \hat{\beta}_{\mathrm{x|y}}$. Still, only half of the matrix of coefficients is presented to illustrate their use due to space limitations. Both zero and negative correlation between commodities are represented as zeros because negative correlations do not translate into consolidation efforts. NA values correspond to cases lacking information to provide reliable estimates for the regressions.



| Households with kids | | | Income segments (NOK) | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Less than 350,000 | | | | 350,000-550,000 | | | | 550,000-750,000 | | | | more than 750,000 | | | |
| | | | cloth | cons | child | oth | cloth | cons | child | oth | cloth | cons | child | oth | cloth | cons | child | oth |
| Age segments | less than or equal to 37 | elect | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.07 | 0.06 | 0 | 0 | NA | NA | NA | NA |
| | | cloth | | 0.56 | 0.22 | 0 | | 0.37 | 0.43 | 0 | | 0.49 | 0.04 | 0 | | 0.35 | 0 | NA |
| | | cons | | | 0.31 | 0.11 | | | 0.42 | 0.11 | | | 0.26 | 0 | | | 0 | NA |
| | | child | | | | 0 | | | | 0 | | | | 0 | | | | NA |
| | 37-57 | elect | 0 | 0.13 | 0.22 | 0 | 0 | 0 | 0.18 | 0 | 0 | 0 | 0.14 | 0 | 0.14 | 0.42 | 0.42 | 0 |
| | | cloth | | 0.26 | 0 | 0.08 | | 0.33 | 0.2 | 0 | | 0.12 | 0.22 | 0 | | 0.52 | 0.05 | 0 |
| | | cons | | | 0.08 | 0.45 | | | 0.07 | 0.12 | | | 0.09 | 0.27 | | | 0.15 | 0 |
| | | child | | | | 0 | | | | 0 | | | | 0 | | | | 0 |
| | Older than 57 | elect | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| | | cloth | | NA | NA | NA | | NA | NA | NA | | NA | NA | NA | | NA | NA | NA |
| | | cons | | | NA | NA | | | NA | NA | | | NA | NA | | | NA | NA |
| | | child | | | | NA | | | | NA | | | | NA | | | | NA |

**Figure C1.** Beta values of linear regression model between commodity groups. Consumers with kids.



| Househ. without kids | | | Income segments (NOK) | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Less than 350,000 | | | | 350,000-550,000 | | | | 550,000-750,000 | | | | more than 750,000 | | | |
| | | | cloth | cons | child | oth | cloth | cons | child | oth | cloth | cons | child | oth | cloth | cons | child | oth |
| Age segments | less than or equal to 37 | elect | 0 | 0.09 | 0 | 0 | 0 | 0.09 | 0 | 0 | 0 | 0 | 0.11 | 0 | NA | 0 | 0 | 0 |
| | | cloth | | 0.29 | 0 | 0 | | 0.13 | 0 | 0 | | 0 | 0.23 | 0 | | NA | NA | NA |
| | | cons | | | 0.04 | 0 | | | 0.05 | 0 | | | 0.03 | 0 | | | 0.03 | 0 |
| | | child | | | | 0 | | | | 0 | | | | 0 | | | | 0 |
| | 37-57 | elect | 0 | 0 | 0.35 | 0.03 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | | cloth | | 0.04 | 0 | 0.06 | | 0 | 0 | 0 | | 0.18 | 0.15 | 0 | | 0.13 | 0 | 0 |
| | | cons | | | 0 | 0 | | | 0 | 0.01 | | | 0 | 0 | | | 0 | 0.18 |
| | | child | | | | 0 | | | | 0 | | | | 0.01 | | | | 0 |
| | Older than 57 | elect | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.34 | 0 | 0 |
| | | cloth | | 0.11 | 0 | 0 | | 0.16 | 0 | 0 | | 0.03 | 0 | 0 | | 0 | 0 | 0 |
| | | cons | | | 0 | 0 | | | 0 | 0 | | | 0 | 0 | | | 0.07 | 0 |
| | | child | | | | 0 | | | | 0 | | | | 0 | | | | 0 |

**Figure C2.** Beta values of linear regression model between commodity groups. Consumers with NO kids

## Appendix D: the boroughs in Oslo

**Table D1.** Demographics for the boroughs in Oslo, Norway for year 2018 due to delayed income data for 2019.

| | Adult* population | Adults* per household | Live in apartment building (share) | Age (average) | Net person income per year** | Female (share) | Kids*** per adult** |
|---|---|---|---|---|---|---|---|
| Alna | 36,738 | 1.65 | 78.4 | 37.7 | 394,000 | 0.50 | 0.30 |
| Bjerke | 23,317 | 1.60 | 72.2 | 35.7 | 440,500 | 0.49 | 0.33 |
| Frogner | 49,535 | 1.41 | 91.6 | 39.0 | 670,300 | 0.50 | 0.14 |
| Gamle Oslo | 44,594 | 1.47 | 93.2 | 34.4 | 449,800 | 0.48 | 0.21 |
| Grorud | 20,502 | 1.61 | 75.8 | 38.4 | 384,400 | 0.50 | 0.29 |
| Grünerløkka | 49,735 | 1.43 | 93.4 | 33.5 | 456,900 | 0.49 | 0.17 |
| Nordre Aker | 37,495 | 1.58 | 43.3 | 37.3 | 623,200 | 0.50 | 0.30 |
| Nordstrand | 37,109 | 1.64 | 38.9 | 39.2 | 625,300 | 0.51 | 0.31 |
| Østensjø | 35,877 | 1.57 | 61.1 | 38.7 | 490,500 | 0.52 | 0.32 |
| Sagene | 36,341 | 1.38 | 95.0 | 34.3 | 482,300 | 0.51 | 0.17 |
| Sentrum | 1,059 | 1.30 | 86.6 | 34.1 | 352,200 | 0.40 | 0.04 |
| Søndre Nordstrand | 28,091 | 1.87 | 47.1 | 35.5 | 394,700 | 0.50 | 0.36 |
| St. Hanshaugen | 32,691 | 1.39 | 94.7 | 35.1 | 493,500 | 0.50 | 0.14 |
| Stovner | 23,801 | 1.79 | 63.8 | 38.0 | 375,300 | 0.49 | 0.33 |
| Ullern | 24,680 | 1.60 | 57.0 | 40.9 | 767,100 | 0.52 | 0.29 |
| Vestre Aker | 34,974 | 1.66 | 41.5 | 39.0 | 826,400 | 0.52 | 0.33 |

*18–79years,**0–17 years,***All adults 17years and older.

Note: All statistics are collected from Statistics Norway in August 2020.