Ehrenberg-Bass Institute Working Paper

Forthcoming in the *Journal of Retailing and Consumer Services*

"Predicting future consumer purchases in grocery retailing with the condensed Poisson lognormal model"

Authors:

Assoc. Prof. Giang Trinh - Ehrenberg-Bass Institute

Prof Malcolm Wright - Massey University

University of South Australia | Ehrenberg-Bass Institute for Marketing Science

Predicting future consumer purchases in grocery retailing with the condensed Poisson lognormal model

Author's name:	Giang Trinh, PhD
Position:	Associate Professor/Senior Marketing Scientist
Affiliation:	Ehrenberg-Bass Institute, UniSA Business, University of South Australia
Mailing address:	City West Campus Level 4, Yungondi Building, North Terrace Adelaide, SA 5000
Telephone:	+61 8 830 20600
Email:	Giang.Trinh@unisa.edu.au

Author's name:	Malcolm J. Wright, PhD
Position:	MSA Charitable Trust Chair in Marketing
Affiliation	Massey University, Auckland, New Zealand
Email:	m.j.wright@massey.ac.nz

Abstract

To identify the effect of marketing actions on consumer purchasing, analysts must disentangle the dynamic component of purchasing from expected period-to-period stochastic fluctuations. This is done by comparing marketplace observations to the conditional expectation of future purchasing. Current methods of deriving the conditional expectation contain systematic bias and rely on certain unrealistic modelling assumptions. We therefore propose a new model to predict future consumer purchases in grocery retailing. The new model is a mixture of Erlang-2, Poisson and lognormal distributions or a condensed Poisson lognormal model (CPLN). Using two grocery retailing datasets from the UK, we demonstrate that the CPLN model predicts future consumer purchases well with error of 7% and 9%, respectively. Compared with the previous benchmark model, the condensed Negative Binominal Distribution (CNBD), the CPLN model reduces error by 50% (7% versus 14%) and 67% (9% versus 27%), respectively. Theoretical and practical implications for retailers are discussed.

Keywords

Consumer purchases; Retailing; Erlang process; Inter-purchase time; Negative binomial distribution; Lognormal distribution; Conditional expectation

1. Introduction

Marketing researchers have been interested in modelling and predicting consumer purchasing behaviour for more than half a century (e.g., Martin et al., 2020; Anesbury et al., 2020; Trinh et al., 2014; Chatfield and Goodhardt, 1973; Goodhardt and Ehrenberg 1967; Ehrenberg, 1959). Models of purchase behaviour are useful to summarise the empirical patterns found in sales data, provide benchmarks to assess individual brand performance, and guide managers on the feasible changes in consumer purchasing to which they may aspire.

The earliest model of purchasing behaviour and still one of the most widely used is the Negative Binomial Distribution (NBD), introduced to describe purchases in consumer packaged goods retailing by Ehrenberg (1959). The model assumes that individual purchasing is a Poisson process, with a gamma distribution of the Poisson means across the population, hence yielding the NBD. The NBD has been shown to accurately describe repeat buying over a very wide range of contexts (Ehrenberg, 1988).

Goodhardt and Ehrenberg (1967) extended the NBD to derive the conditional expectation of future purchasing. In plain language, the conditional expectation is the number of purchases in the second period that arises from each of the first period purchase classes (e.g., second period purchasing by each of first period zero buyers, once buyers, twice buyers and so on). In a dynamic situation, the conditional expectation is crucial for retailers who wish to identify the source of any change in sales. For example, if a retailer runs a price promotion for a product and finds that sales increase, it would be important to know if the sales increase comes from previous non-buyers or from existing buyers of the product. The first scenario is good news as price promotion attracts new buyers, potentially making a marginal contribution to profit and supporting longer-term growth in the customer base. But the second scenario might be bad as price promotion encourages existing buyers to stockpile for future use, doing nothing for longer-term growth in the customer base and possibly reducing the retailer's overall profit due to the cost of price promotion.

Although retailers can observe raw figures for second period purchasing, these numbers cannot be interpreted directly. Conditional expectation tells us that some zero buyers will purchase next period even without marketing intervention. Thus, the raw figures for different buyer classes observed by the retailer include much stochastic variation and regression to the mean, rather than necessarily representing

dynamic change. The true dynamic component can only be identified by subtracting the conditional expectation from the observed data.

Due to the ability to distinguish between dynamic change and stochastic variation, conditional expectation is regarded as a very important construct in the modelling of period-to-period purchases (Schmittlein et al., 1985). Morrison and Schmittlein (1988, p. 149) argue, "Fitting histograms of purchases well yields very little insight…. It is the conditional expectation that has potential to give decision making guidance." Yet, although the NBD's conditional expectation is useful, it somewhat over-predicts purchases of existing buyers and under-predicts the purchases of new buyers (Lenk et al., 1993; Trinh et al., 2014; Trinh et al., 2019). There are two established criticisms of the NBD as an approximation of purchasing behaviour that may go some way to explaining these discrepancies. The first is that the Poisson assumption implies an exponential distribution of inter-purchase times not representative of a typical buyer, due to the likelihood of a non-purchase period immediately following purchase (Chatfield and Goodhardt, 1973; Kahn and Schmittlein, 1989). The second is that the gamma assumption is merely a convenience without a strong theoretical basis (Ehrenberg, 1988), and tends to underestimate the long right tail of a small number of heavy buyers commonly present in purchasing data (Trinh et al., 2014).

These criticisms have led to some innovation in NBD model specification. One approach has been to replace the Poisson with an Erlang distribution to better capture the theorized 'dead' period after purchase, giving an Erlang gamma mixture model known as the Condensed Negative Binomial Distribution or CNBD (Chatfield and Goodhardt, 1973). Another has been to replace the gamma distribution with the lognormal distribution, not only better modelling the right tail of the purchasing distribution but offering an appealing theoretical explanation for why the distribution arises in the context of consumer purchasing behaviour (Martin et al., 2020; Page et al., 2019; Sorensen et al., 2017; Trinh, 2014).

However, these two innovations have not been fully applied to conditional expectation. The Erlang-lognormal distribution has not been developed and assessed, and the condensed Poisson-lognormal has not been extended to second period purchasing. As conditional expectation is of such great importance to managers and researchers, improvements in accuracy have considerable potential benefits. Establishing a more accurate and theoretically appealing basis for conditional

expectation has potential to encourage further practical application and research in this important area.

The present work therefore introduces and evaluates the condensed Poisson lognormal model (CPLN) to improve conditional expectation, and so provide a more accurate method of determining the dynamic component of period-to-period purchasing. The remainder of this paper (i) reviews prior work on the NBD model, the conditional expectation, and the lognormal distribution; (ii) derives the Erlang lognormal mixture, or condensed Poisson lognormal model (CPLN); (iii) derives the conditional expectation of CPLN; (iv) uses both simulation and empirical studies to compare CPLN with the benchmarking model CNBD; (v) discusses the implication of using the CPLN model; (vi) and outlines some new research directions using CPLN.

2. Literature Review

Negative Binomial Distribution

The NBD model was initially used by Greenwood and Yule (1920) to model accidents and first applied to marketing in the retailing context by Ehrenberg (1959). Ehrenberg made two assumptions:

- The purchases of a retailing product (e.g. cereal, coffee or toothpaste) by a given consumer in successive time periods follow a Poisson distribution. This implies that the variance of purchases within individual consumers is "as if" random over time (i.e. Poisson process).
- The mean rates of purchasing of different consumers in the long run differ and their distribution is a gamma distribution. In other words, the variance of mean purchases across different consumers is measured by a gamma distribution.

Following these assumptions, the frequency of consumers making 0, 1, 2, 3,…x purchases in a given time period can be modelled by the Negative Binomial Distribution.

Since the original article by Ehrenberg (1959), the NBD model has been shown to work well in numerous situations, particularly in consumer packaged goods including different brands in different categories, different time periods, and different countries (Sharp, 2010). Recently, the NBD has been applied to different types of behaviours such as gambling behaviour (Mizerski et al., 2004; Lam and Mizerski,

2009), consumption behaviour of mobile phone services (Lee et al., 2011) and industrial purchases (Wilkinson et al., 2016).

Conditional Expectation

As noted, one of the most useful properties of the NBD is its conditional expectation. In the case of modelling purchase frequency, this is the expected mean of purchases in period two made by the buyers who bought *x* purchases in period one. Based on the NBD conditional expectation, Goodhardt and Ehrenberg (1967) introduced a method to benchmark future sales change from past performance, which they called conditional trend analysis (CTA). It is regarded as a very important method with significant managerial implications (Morrison and Schmittlein, 1988). For example, CTA is crucial for brand managers who wish to identify whether a change in overall brand sales level is accounted for by previous non-buyers, light buyers or heavy buyers (Schmittlein et al. 1985). By comparing the actual purchases to the NBD conditional expectation in period two, we are able to determine which group of buyers is causing any sales gains or losses. CTA is therefore regarded as one of the most managerially useful constructs in the stochastic modelling of brand choice (Schmittlein et al. 1985).

Although it has been shown that the NBD model fits the observed data quite well in a wide range of conditions, prior literature has noted that the NBD does not give a good fit for the tail of the distribution, especially for purchase frequencies of heavily brought products (Ehrenberg, 1959; Chatfield et al., 1966). It also gives a poor fit if there are outliers such as excessively heavy buyers (Chatfield et al., 1966; Ehrenberg, 1988). These factors are exacerbated for the conditional expectation, motivating inquiry whether the NBD model and CTA may be improved.

Condensed Negative Binomial Distribution

Chatfield and Goodhardt (1973) questioned the Poisson assumption of the NBD model. They argued that if the Poisson assumption is true for any time period, then inter-purchase times should follow the exponential distribution. Thus the mode of inter-purchase times should be zero. Yet, in practice, it is not likely that a buyer purchases again immediately, making the assumption unrealistic and unrepresentative of the typical buyer. Rather, there is a 'dead period' (e.g. a week or more) between one purchase and another. The authors then proposed an alternative, the Erlang

distribution. This distribution was suggested by Herniter (1971) for modelling inter-purchase times; however, he only considered a special case in which the mean purchase rate is exponentially distributed and this rarely happens in practice (Chatfield and Goodhardt, 1973). Based on the Erlang 2 assumption, Chatfield and Goodhardt (1973) derived the distribution of an individual consumer's purchases for a given period, which they called the condensed Poisson distribution. The integration of the Erlang 2 with the gamma distribution across the whole population gives the condensed NBD model (CNBD).

In response to Chatfield and Goodhardt (1973), Schmittlein and Morrison (1983) derived the conditional expectation of the CNBD model to predict future purchases. They applied the model to empirical data and found that CNBD gives better prediction than the NBD when the number of non-buyers is large. In a later review Morrison and Schmittlein (1988 pp. 148-149) conclude, "The Erlang captures the spirit of any behaviour that we are likely to observe…. Thus changing the exponentially distributed inter-purchase times of the NBD to Erlang seems like a good first step toward 'improving' the NBD." Consequently, a considerable number of studies use Erlang distribution instead of exponential distribution (Chen and Steckel, 2012; Gupta, 1991; Gupta, 1988; Jeuland et al., 1980, Wu and Chen, 2000a; Wu and Chen, 2000b; Zufryden, 1978; Zufryden, 1977).

Lognormal Distribution

Although the conditional expectation of CNBD offers improvement over that of the NBD, it still under-predicts the purchases of non and light buyers and over-predicts the purchases of heavy buyers (Morrison and Schmittlein, 1981). This bias can be a serious problem as one might mistakenly choose to target heavy buyers to increase purchases instead of new or light buyers (Lenk et al., 1993). One possible reason for this bias is the lack of fit of the gamma distribution to heavy buyers as showed in Chatfield et al.'s (1966) study. The gamma assumption portion is questionable since it lacks theoretical support (Brockett et al., 1996; Laurence, 1980; Trinh et al., 2014). Ehrenberg (1988) notes that it is impossible to explain why a gamma distribution holds. Empirically, the gamma assumption is the primary reason for failure to model purchase frequency due its light tail on the right, especially when long-tail distributions characterise data (Chatfield and Goodhardt, 1973; Brockett et al., 1996; Ehrenberg et al., 2004). Practically, if a retailer allocates its marketing

resource for a product (e.g. cereal) on the basis of a model prediction, it is crucial to develop a model that gives better prediction to the data to ensure that the results from the analysis are valid and not due to the model inaccuracy.

A solution to these problems may lie in the lognormal distribution. Many disciplines such as geology, economics, telecommunications, biochemistry, demography, health, psychology, risk analysis, marketing and retailing have used the naturally occurring lognormal distribution as it has an attractive theoretical interpretation compared to the gamma distribution (Aitchison and Brown, 1969; Bulmer, 1974; Cassie, 1962; Crow and Shimizu, 1988; El-Basyouny and Sayed, 2009; Fahidy, 2005; Johnson et al., 1994; Martin et al., 2020; Page et al., 2019; Sorensen et al., 2017; Trinh, 2014; Winkelmann, 2008).

The theory of the lognormal distribution of mean purchase rate can be described as following. Suppose x_t is the mean purchase rate of an individual buyer at time t, and e_t is a serials of random variables, which independently identical distributed, e_t is also independent of x_t

Then

$$x_t - x_{t-1} = e_t x_{t-1}$$

Or

$$x_t = (e_t + 1)x_{t-1}$$

Starting with any mean purchase rate at time 0, x_0, we have

$$x_t = (e_1 + 1)(e_2 + 1)\ldots(e_t + 1)x_0$$

Suppose the effect at each step to be small, then

$$\log(1+e) = e$$

Taking logs we obtain

$$\log x_t = \log x_0 + e_1 + e_2 + \ldots + e_t$$

Relying on the central limit theorem, $\log x_t$ is normally distributed and hence x_t is lognormally distributed (Aitchison and Brown, 1969).

In the grocery retailing context, it is reasonable to assume that the individual consumer's mean purchase rate is determined through the interaction of multiple unobserved factors, which can be both positive or negative such as advertising, promotion, word of mouth and other consumer specific factors. If there are many independent unobserved factors that affect the mean purchase rate of a given

consumer, the multiplicative process may converge them to a lognormal distribution relying on the central limit theorem as shown above (Aitchison and Brown 1969; Johnson et al. 1994; Winkelmann 2008).

Empirically, previous research has shown the lognormal gives a better fit than the gamma in long tail data (Connolly et al., 2009; Sohn, 1994; Miranda-Moreno et al., 2005). Thus the lognormal may be practically as well as theoretically suitable for purchase frequency of heavily brought products where the gamma shows a lack of fit (Ehrenberg, 1959; Chatfield et al., 1966; Ehrenberg, 1988).

The present research therefore replaces the gamma distribution in the CNBD model with the theoretically and empirically appealing lognormal distribution. This creates a new mixture model with the Erlang-2 and Poisson distribution or CPLN. The new model is then not only used to fit consumer purchase data but also to predict future consumer purchases – the conditional expectation.

3. Model development

Following Chatfield and Goodhardt (1973), Erlang inter-purchase times are assumed for purchases of a product by any given buyer, where every 2x purchase from a Poisson process is counted.

$$f(t|\lambda) = \lambda^2 t e^{-\lambda t}, t > 0 \qquad \textbf{(1)}$$

Hence, the distribution of purchase frequency in a given period follows the censored or condensed Poisson distribution:

$$f_{CP}(0) = f_P(0) + \frac{1}{2} f_P(1)$$

$$f_{CP}(x) = \tfrac{1}{2} f_P(2x-1) + f_P(2x) + \tfrac{1}{2} f_P(2x+1) \qquad (2)$$

where

$$f_p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad (3)$$

with mean

$$E_{cp}[x] = \frac{\lambda}{2} \qquad (4)$$

and variance

$$var_{CP}[x] = \frac{\lambda}{4} + \frac{1}{4}e^{-\lambda}sinh\lambda \quad (5)$$

The mean rate of purchase follows a lognormal distribution in the population:

$$dF = \frac{1}{z\sigma\sqrt{2\pi}}e^{-\frac{(log\,z-\mu)^2}{2\sigma^2}}dz \quad (6)$$

with mean

$$E[z] = e^{\pi+\frac{\sigma^2}{2}} \quad (7)$$

and variance

$$var[z] = e^{2\mu+\sigma^2}(e^{\sigma^2}-1) \quad (8)$$

Let $\lambda = 2z$ where λ is the corresponding Poisson rate before censoring. Then the distribution of mean purchase rates can be expressed as:

$$dF = \frac{1}{\lambda\sigma\sqrt{2\pi}}e^{-\frac{(log\,z-\mu)^2}{2\sigma^2}}d\lambda$$

$$= \frac{1}{\lambda\sigma\sqrt{2\pi}}e^{-\frac{\left(log\frac{\lambda}{2}-\mu\right)^2}{2\sigma^2}}d\lambda$$

$$= \frac{1}{\lambda\sigma\sqrt{2\pi}}e^{-\frac{\left(log\lambda-(log2+\mu)\right)^2}{2\sigma^2}}d\lambda$$

$$= \frac{1}{\lambda\sigma\sqrt{2\pi}}e^{-\frac{(log\lambda-\mu')^2}{2\sigma^2}}d\lambda \quad (9)$$

with $\mu' = \log2 + \mu$

The Erlang lognormal model or condensed Poisson lognormal (CPLN) model is then obtained by mixing the condensed Poisson distribution with the lognormal distribution:

$$f_{CPLN}(0) = f_{PLN}(0) + \frac{1}{2}f_{PLN}(1)$$

$$f_{CPLN}(x) = \frac{1}{2}f_{PLN}(2x-1) + f_{PLN}(2x) + \frac{1}{2}f_{PLN}(2x+1) \quad (10)$$

where $f_{PLN}(x)$ is a Poisson lognormal distribution (PLN) with parameters μ' and σ

$$f_{PLN}(x) = \int_0^\infty f_P(x)f(\lambda;\mu',\sigma)d\lambda$$

$$= \frac{1}{x!\sigma\sqrt{2\pi}}\int_0^\infty \lambda^{x-1}e^{-\lambda}e^{-\frac{(log\lambda-\mu')^2}{2\sigma^2}}d\lambda \quad (11)$$

Similar to CNBD model, the mean of CPLN is half of that of PLN:

$$E_{CPLN}[x] = \frac{E_{PLN}[x]}{2} = \frac{e^{\mu' + \frac{\sigma^2}{2}}}{2} = \frac{e^{log2+\mu+\frac{\sigma^2}{2}}}{2} \tag{12}$$

The variance of CPLN is the sum of mean variance of the condensed Poisson distribution and the variance of the lognormal distribution:

$$var_{CPLN}[x] = E(var_{CP}[x]) + var_{LN}\left[\frac{\lambda}{2}\right]$$

$$= E\left(\frac{\lambda}{4} + \frac{1}{4}e^{-\lambda}sinh\lambda\right) + e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{13}$$

The conditional expectation of CPLN can be derived using Bayes' theorem:

$$E[X_2|X_1 = x] = E[z|x] = \frac{\int_0^\infty P_{CPLN}(x|z)zdF(z)}{\int_0^\infty P_{CPLN}(x|z)dF(z)} \tag{14}$$

Since $z = \frac{\lambda}{2}$, the conditional expectation of CPLN can be expressed as:

$$E[X_2|X_1 = x] = \frac{1}{2}\frac{\int_0^\infty P_{CPLN}(x|\lambda)\lambda dF(\lambda)}{\int_0^\infty P_{CPLN}(x|\lambda)dF(\lambda)}$$

$$= \frac{1}{2}\frac{\int_0^\infty \left[\frac{1e^{-\lambda}\lambda^{2x}}{2(2x-1)!} + \frac{e^{-\lambda}\lambda^{2x+1}}{(2x)!} + \frac{1}{2}\frac{e^{-\lambda}\lambda^{2x+2}}{(2x+1)!}\right]dF(\lambda)}{f_{CPLN}(x)} =$$

$$\frac{1}{2}\frac{\int_0^\infty \left[x\frac{e^{-\lambda}\lambda^{2x}}{(2x)!} + (2x+1)\frac{e^{-\lambda}\lambda^{2x+1}}{(2x+1)!} + (x+1)\frac{e^{-\lambda}\lambda^{2x+2}}{(2x+2)!}\right]dF(\lambda)}{f_{CPLN}(x)} =$$

$$\frac{1}{2}\frac{xf_{PLN}(2x) + (2x+1)f_{PLN}(2x+1) + (x+1)f_{PLN}(2x+2)}{f_{CPLN}(x)} \tag{15}$$

4. A simulation study to compare CPLN and CNBD

As an initial comparison we consider the theoretical distribution and conditional expectation for purchases arising from the CPLN and CNBD differ in any substantive way. The proportion of buyers, average purchase frequency and sample size are specified for two sets of simulated data. The first dataset characterises frequent purchases with a high proportion of buyers (80% of the population) and high average purchase frequency (4.5 purchases in the period). The second dataset characterises infrequent purchases with a low proportion of buyers (20% of the population) and low average purchase frequency (1.5 purchases in the period). With CNBD, the mean and zero method is used to estimate the frequency distribution and conditional expectation of purchases as it has a closed form (Chatfield and Goodhardt,

1973; Schmittlein and Morisson, 1983). With CPLN, a numerical estimation method based on 1000 random draws is used to calculate the frequency distribution as it does not have a closed form. Next, equation (15) is used to calculate the conditional predicting of purchases.

Figure 1 and Figure 2 show the theoretical purchase frequency distribution and conditional expectation for the CPLN and CNBD models. Note that in a stationary market the theoretical purchase frequency distribution shown in Figure 1 (without conditional expectation) will be identical from period to period and it is this distribution that is used to fit the model.

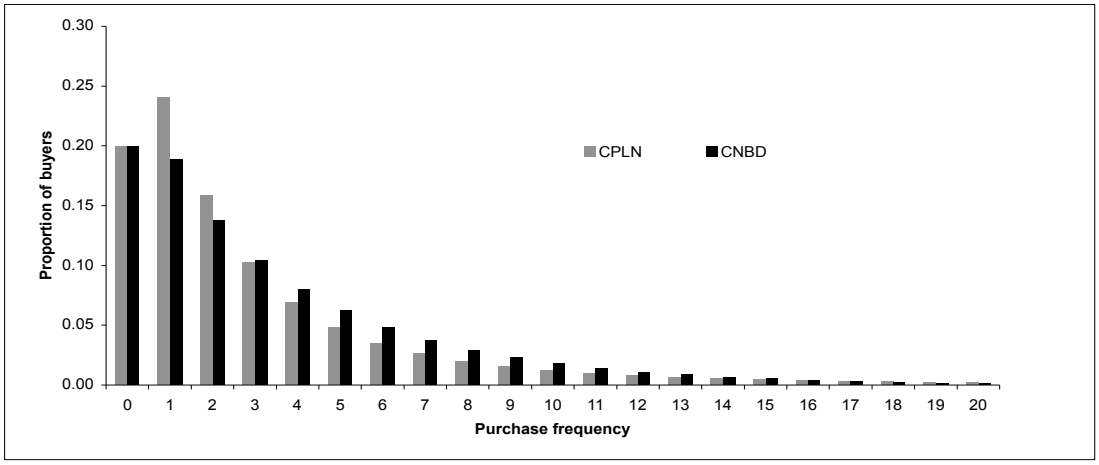

Figure 1 CPLN and CNBD for frequent purchase data.

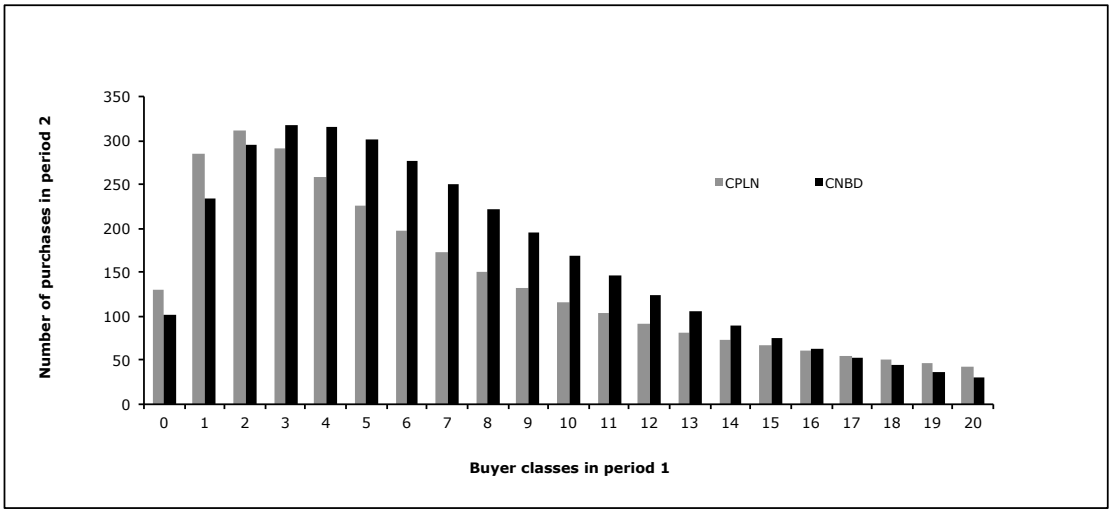

Figure 2 Conditional predictions by CPLN and CNBD for frequent purchasing data.

Figure 1 shows that the purchase frequency distributions of CPLN and CNBD look quite different. CPLN shows a higher number of light buyers (those who purchase one or two times), whereas CNBD shows a higher number of heavy buyers (those who purchase four times or more). Consequently, the second period conditional expectation in Figure 2 shows greater purchasing by light buyers for the CPLN compared to the CNBD. These results indicate that CPLN has potential to reduce the bias of CNBD in terms of predicting heavy buyers' purchases. Figure 3 and Figure 4 show the theoretical purchase frequency distribution and conditional expectation from the CPLN and CNBD models for the infrequent purchase data.

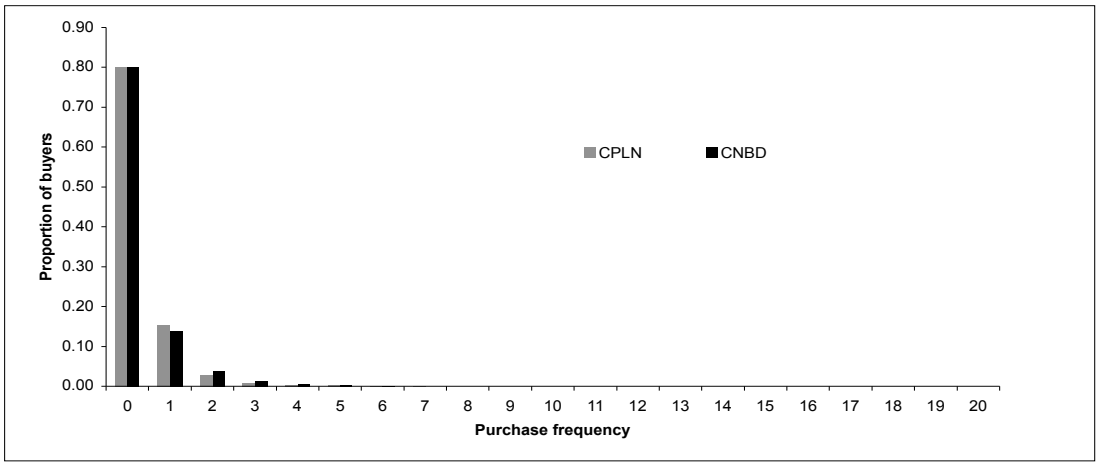

Figure 3 CPLN and CNBD for infrequent purchase data.

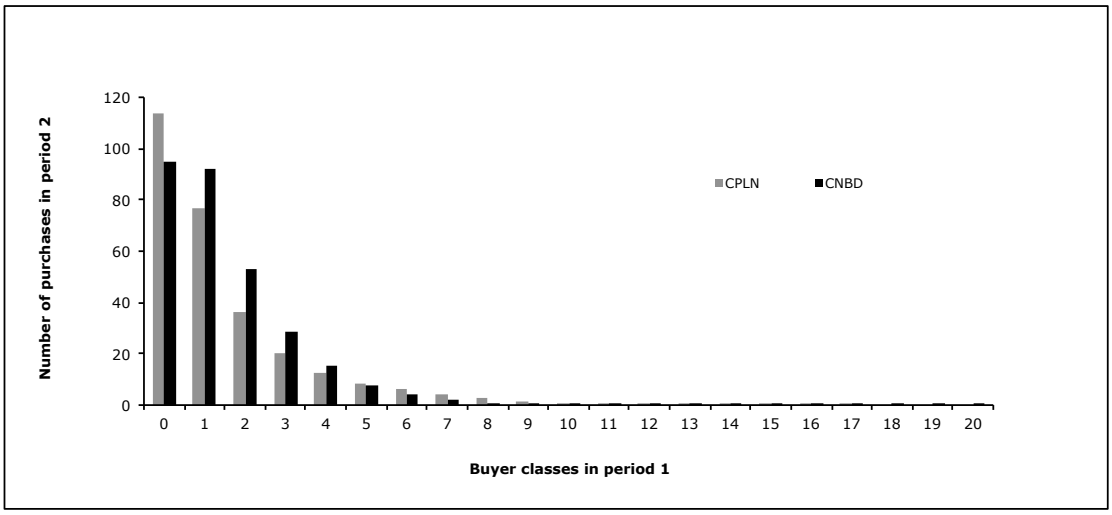

Figure 4 Conditional predictions by CPLN and CNBD for infrequent purchase data.

Although there are only slight differences in purchase frequency distribution of CPLN and CNBD in Figure 3, the conditional expectation of purchase by the two models in Figure 4 differ substantially. The purchases in period two by non-buyers in period one is greater for CPLN than CNBD. Conversely the purchases in period two of existing buyers in period one is lesse for CPLN than CNBD. These results show that CPLN has potential to reduce the bias of CNBD in terms of predicting of non-buyers' purchases.

5. Empirical analysis

For empirical analysis, two datasets are used to compare the fit and predictive accuracy of CPLN versus CNBD. The two datasets characterise two types of distribution: a long tail for a heavily bought personal care product, and a short tail for a lightly bought household care product. The data used are household purchasing over a 104-week period from Kantar Superpanel in the UK. Only aggregate data are available. The first year is used for parameter estimation and the second year as a test period for predicting future purchases.

The parameters of the CNBD are estimated using the maximum likelihood method while the parameters of the CPLN are estimated with the numerical approximation method proposed by Train (2009), using 1000 Halton draws (Halton 1960). Extant research suggests that Halton draws provide better results than random draws (Bhat 2001; Hensher 2001; Spanier and Maize 1991; Train 2000; Train 2009).

For each dataset, model fit to the year one purchases is examined, and then conditional expectation for year two is evaluated. Model fit is assessed through the Log-likelihood ratio and Theil's U coefficient of inequality, while predictive accuracy for the conditional expectation is assessed with Theil's U coefficient of inequality and weighted mean absolute percentage error (Gupta, 1988; Wu and Chen, 2000a; 2000b; Trinh et al., 2014; Ludwichowska et al., 2017; Sorensen et al., 2017; Martin et al., 2020).

5.1. Model fit and prediction for personal care product

For the personal care product the dataset contains continuously reported, purchases for 16,998 households. The year one proportion of buyers and average

purchase frequency are 88% and 3.98, respectively. Fitting the CPLN and CNBD models to the first year of these data gives log-likelihoods of -39,863 and -39,904, respectively. The log-likelihood is higher for CPLN, which suggests it outperforms CNBD. The U coefficients of equality for CPLN and CNBD were 0.038 and 0.056, respectively, which confirms that CPLN fit these data better than CNBD. Figure 5 presents the fit of both models. It is clear CPLN offers improvement over CNBD when describing lighter and heavier purchase classes.

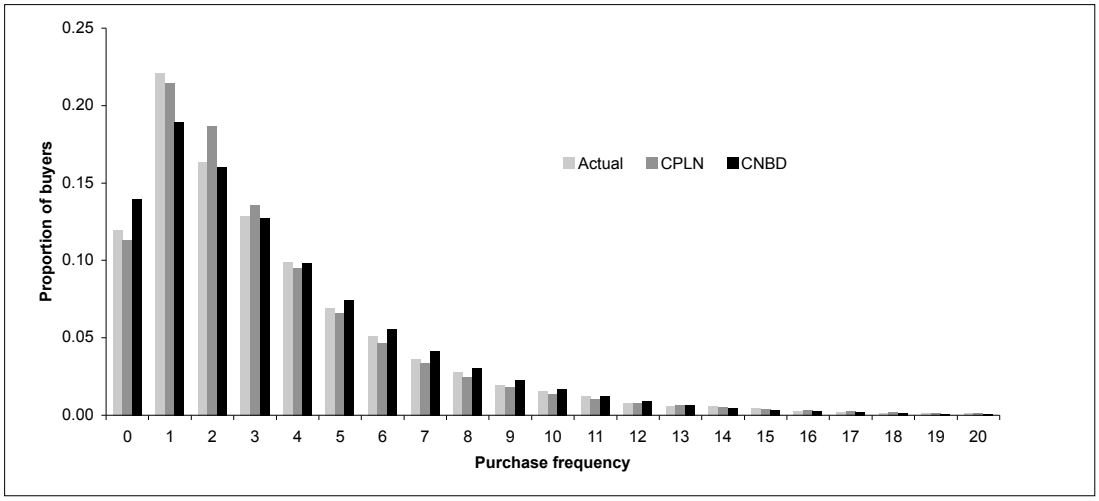

Figure 5 The fit of the CNBD and CPLN models (Personal care product).

Figure 6 shows the conditional predictions by the CNBD and CPLN models. CPLN predicts future purchases better than CNBD for most buyer classes. Theil's U coefficients were 0.068 for CPLN and 0.086 for CNBD, which corroborate the superiority of CPLN when predicting of future purchases for this dataset. Weighted mean absolute percentage error are 0.07 for CPLN and 0.14 for CNBD, which show that the CPLN model reduces the error in future purchase prediction by half compared to the CNBD model.

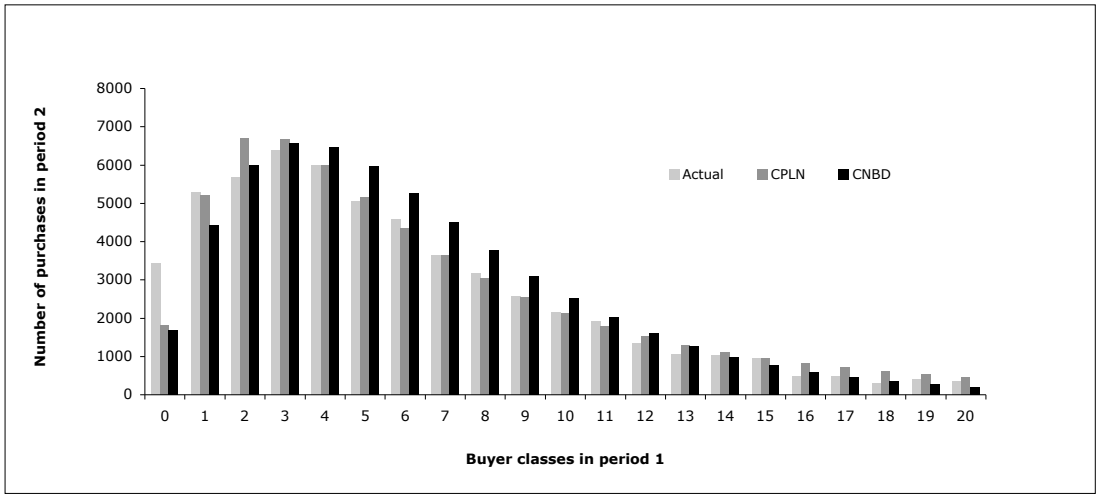

Figure 6 CPLN and CNBD conditional predictions (Personal care product).

5.2. Model fit and prediction for household care product

For the household care product, the dataset contains continuously reported purchases for 9,063 households. Contrary to the personal care dataset, this dataset has a much lower observed proportion of buyers and average purchase frequency: 13% and 1.87, respectively. Fitting CPLN and CNBD models to the first year data gives log-likelihoods of -5,068 and -5,033, respectively. Assessing log-likelihoods, CNBD outperforms CPLN. Theil's U coefficients of equality confirms this result, with values of 0.003 for CNBD and 0.009 for CPLN. Figure 7 presents the fit of both models which is clearly close in both cases albeit slightly better for the CNBD.

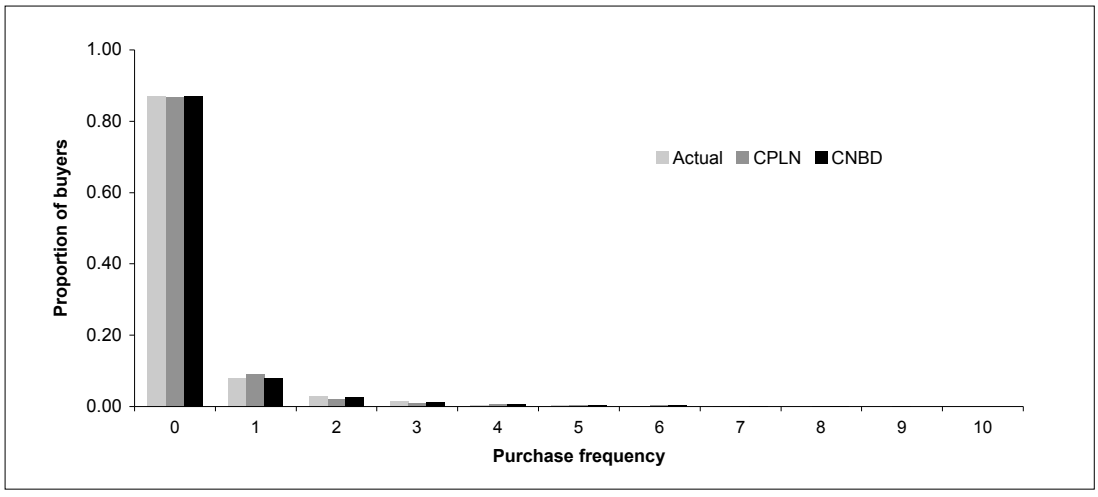

Figure 7 The fit of CNBD and CPLN models (Household care product)

Figure 8 shows the conditional predictions by the CNBD and CPLN models. Although CPLN underperformed CNBD regarding in-sample fitting, it predicted future purchase better. Theil's U coefficients of equality for the CPLN and CNBD models are 0.100 and 0.146, respectively. Weighted MAPE for the CPLN and CNBD models are 0.09 and 0.27, respectively. In comparison to the CNBD model, CPLN demonstrates significant improvements in the conditional predicting of future purchases for the dataset. The CPLN model reduces two thirds of the error in terms of future purchase prediction.

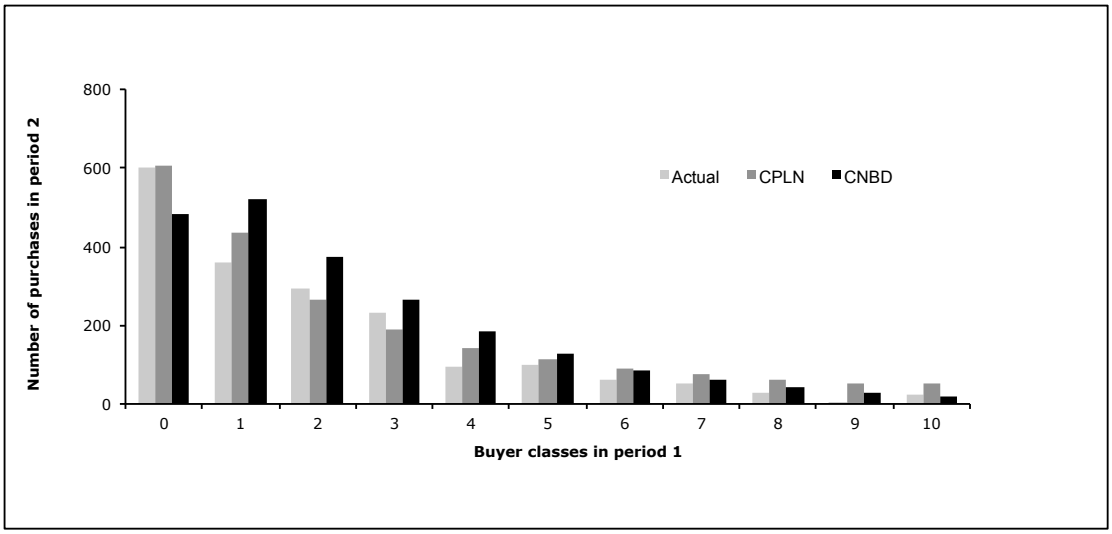

Figure 8 CPLN and CNBD conditional predictions (Household care product).

6. Conclusion, Implications and Future Research

From a theoretical perspective, this research contributes to the literature by introducing a new model for predicting consumer purchases in grocery retailing that mixes the lognormal distribution with the Erlang-2 and Poisson distributions. Not only does the lognormal distribution have an attractive theoretical interpretation that can explain differences in purchasing behaviour, previous empirical research has found that lognormal based models give a better fit for count data compared to gamma based models (e.g. Connolly et al. 2009; Kaas and Heseelager 1995; Miranda-Moreno et al. 2005; Sohn 1994; Tsionas 2010; Winkelmann 2008). Simulation shows the new CPLN model provides quite different predictions to the CNBD model for initial period purchasing and conditional expectation. Empirical analysis shows the

new CPLN model provides accurate conditional expectation with an error of less than 10%. Compared to the previous CNBD benchmark model, the CPLN model reduces the error in conditional expectation by 50% or more.

From a managerial perspective, a retailer can use the CPLN model to obtain more accurate predictions by future purchases of their customers. This is particularly useful to evaluate the source of any sales change following marketing activities such as price promotion and advertising. By comparing the actual purchases with the conditional expectation following a price promotion, the retailer can identify the source of any sales growth. If the source of growth were mainly from new buyers, then price promotion works well as the growth is not borrowed from future sales. In contrast, if the source of growth were mainly from existing buyers, the retailer would need to worry about the effect of price promotion as it could potentially borrow sales from the future. Another example would be in the management of the most valuable customers. Conditional expectation suggests some reduction in purchasing by heavy buyers in subsequent periods due to regression to the mean. But has there been too much reduction, or is normal, or is it less than expected indicating good performance in managing heavy buyers? The more accurate the conditional expectation, the more accurate the evaluation of marketing performance.

There are some limitations of this study. First, although the CPLN model is theoretically sound and empirically validated, it does not have a closed form. Hence, simulation is needed for parameter estimation. Second, for purchase prediction, despite the CPLN model achieving errors of less than 10%, it is not an error-free model, so further model development may be possible. Third, the study uses only two datasets, so extensive replication over time will provide clearer guidance on how well the model fits in different circumstances and any boundary conditions to its application.

For future research, the CPLN model could be tested against the previous applications of the CNBD model in other contexts. For example, it could be compared with Jeuland et al.'s (1980) model of multi-product purchase mixing the CNBD with the Dirichlet distribution, Gupta's (1991) model of inter-purchase time with time dependent covariates, and Zufryden's (1977; 1978) model of product choice and purchase timing. Researchers could replace the CNBD component in these studies with CPLN to determine if it leads to improvements in these areas. The CPLN could also be tested against other modelling approaches such as Bayesian non-parametric or

machine learning to see how well it predicts future purchase compared to these models.

The second direction of future research could be to extend the CPLN model to a bivariate model to examine switching among competing retailers. For example, we could subdivide the buyers according to the number of purchases of a particular retailer (say, retailer A). For each retailer A purchase class we could examine how much those buyers purchase from another retailer (say, retailer B) in the same period as well as in the next period. The duplication of purchases in the same period analysis is simply the relationship between two count variables (firms A and B purchase distributions), in which the relationship between two latent purchase rates is examined. The higher the correlation between these rates, the stronger the competition between the retailers. Analysis can then provide insights into customer migration between retailers based on conditional expectation. For example, do heavy buyers of a retailer switch to be heavy buyers of another retailer? Or do they switch gradually to be light buyers of another retailer? Or are light buyers those who switch the most?

A third direction of future research is to consider the use of covariates in the model. This was not possible under the assumption of a gamma distribution, as the lack of theoretical support for this assumption means it lacks validity as a basis for the inclusion of explanatory variables. However, the lognormal form is based on a multiplicative combination of explanatory factors that is quite familiar to marketing mix modelling, and so invites exploration of how a multiplicative marking mix regression could be combined with the lognormal distribution to start to explain the mechanisms by which marketing actions affect the distribution of consumer purchasing.

Acknowledgements

Thanks to Kantar World Panel for supporting the University of South Australia and providing the data to enable this research.

References

Aitchison, J., & Brown, J. (1969). *The lognormal distribution*. Cambridge University Press, New York.

Anesbury, Z. W., Talbot, D., Day, C. A., Bogomolov, T., & Bogomolova, S. (2020). The fallacy of the heavy buyer: Exploring purchasing frequencies of fresh fruit and vegetable categories. *Journal of Retailing and Consumer Services*, *53*, 101976.

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research*, *35B*(7), 677-695.

Brockett, P. L., Golden, L. L., & Panjer, H. H. (1996). Flexible purchase frequency modelling. *Journal of Marketing Research*, *33*, 94-107

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, *30*, 101-110.

Cassie, R.M. (1962). Frequency distribution models in the ecology of plankton and other organisms. *Journal of Animal Ecology*, *31*, 65-92.

Chatfield, C., & Goodhardt, G. J. (1973). A consumer purchasing model with Erlang inter-purchase time. *Journal of the American Statistical Association*, *68*(344), 828-835

Chatfield, C., Ehrenberg, A. S. C., & Goodhardt, G. J. (1966). Progress on a simplified model of stationary purchasing behaviour, *Journal of the Royal Statistical Society. Series A (General)*. *129*(3), 317-367.

Chen, Y., & Steckel, J. H. (2012). Modelling credit card share of wallet: Solving the incomplete information problem. *Journal of Marketing Research*, *49*(5) 655-669.

Connolly, S. R., Dornelas, M., Bellwood, D. R., & Hughes, T. P. (2009). Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. *Ecology*, *90*, 3138-3149.

Crow, E. L., & Shimizu, V. (1988). Lognormal distributions: Theory and applications. Marcel Dekker. New York.

Ehrenberg, A. S. (1959). The pattern of consumer purchases. *Applied Statistics*, 26-41.

Ehrenberg, A.S.C. (1988). Repeat buying, second ed. Griffin, London

El-Basyouny, K. and Sayed, T. (2009). Accident prediction models with random corridor parameters. *Accident Analysis and Prevention*, *41*, 1118–1123.

Fahidy, T. Z. (2005). Electrochemical horizons for the Poisson-lognormal distribution of probability theory. *Journal of Electroanalytical Chemistry*, *581*, 11–15.

Goodhardt, G. J., & Ehrenberg, A. S. C. (1967) Conditional trend analysis: A breakdown by initial purchasing level. *Journal of Marketing Research*, *4*(2), 155-161.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research*, *25*(4), 342-355.

Gupta, S. (1991). Stochastic models of interpurchase time with time-dependent covariates. *Journal of Marketing Research*, *28*(1), 1-15.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals. *Numerische Math*, *2*, 84-90.

Hensher, D. A. (2001). The valuation of commuter travel time savings for car drivers in New Zealand: Evaluating alternative model specifications. *Transportation*, *28*(2), 101-118.

Jeuland, A.P., Bass, F.M., & Wright, G.P. (1980). A multibrand stochastic model compounding heterogeneous erlang timing and multinomial choice process. *Operations Research*, *28*, 255–277.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*, *vol. 1*, John Wiley & Sons, New York.

Kahn, B. E., & Schmittlein, D. C. (1989). Shopping trip behaviour: an empirical investigation, *Marketing Letters*, *1*(1), 55-69.

Kaas R., and Heseelager, O. (1995). Ordering claim size distributions and mixed Poisson probabilities. *Insurances: Mathematics and Economics*, *17*, 193-201.

Lenk, P. J., Rao, A. G., & Tibrewala, V. (1993) Nonstationary conditional trend analysis: An application to scanner panel data. *Journal of Marketing Research*, *30*, 288-304.

Ludwichowska, G., Romaniuk, J., & Nenycz-Thiel, M. (2017). Systematic response errors in self-reported category buying frequencies. *European Journal of Marketing, 51*(7/8), 1440-1459.

Martin, J., Nenycz-Thiel, M., Dawes, J., Tanusondjaja, A., Cohen, J., McColl, B., & Trinh, G. (2020). Fundamental basket size patterns and their relation to retailer performance. *Journal of Retailing and Consumer Services*, *54*, 102032.

Miranda-Moreno, L. F., Fu, L., Saccomano, F. F., & Labbe, A. (2005). Alternative risk model for ranking locations for safety improvement. *Transportation Research Record*, *1908*, 1-8.

Morrison, D. G., & Schmittlein, D. C. (1981). Predicting future random events based on past performance. *Management Science*, *27*, 1006-1023.

Morrison, D. G., & Schmittlein, D. C. (1988). Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort?. *Journal of Business & Economic Statistics*, *6*(2), 145-159.

Page, B., Trinh, G., & Bogomolova, S. (2019). Comparing two supermarket layouts: The effect of a middle aisle on basket size, spend, trip duration and endcap use. *Journal of Retailing and Consumer Services*, *47*, 49-56.

Schmittlein, D. C., & Morrison, D. G. (1983). Prediction of future random events with the condensed negative binomial distribution. *Journal of the American Statistical Association*, *78*, 449-456.

Schmittlein, D. C., Bemmaor, A. & Morrison, D. G. (1985). Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Science*, *4*(3), 255-266.

Sharp, B. (2010). *How Brands Grow*. South Melbourne: Oxford University Press.

Sohn, S. Y. (1994). A comparative study of four estimators for analyzing the random event rate of the Poisson process. *Journal of Statistical Computation and Simulation*, *49*, 1-10.

Sorensen, H., Bogomolova, S., Anderson, K., Trinh, G., Sharp, A., Kennedy, R., Page, B., & Wright, M. (2017). Fundamental patterns of in-store shopper behavior. *Journal of Retailing and Consumer Services*, *37*, 182-194.

Spanier, J., & Maize, E. (1991). Quasi-random methods for estimating integrals using relatively small samples. *SIAM Review*, *36*, 18–44

Train, K. (2000). Halton sequences for mixed logit. *Working paper No. E00-278*, Department of Economics, University of California, Berkeley.

Train, K. (2009). *Discrete choice methods with simulation*, 2nd ed., Cambridge University Press, Cambridge.

Trinh, G., Rungie, C., Wright, M., Driesener, C. and Dawes, J., (2014). Predicting future purchases with the Poisson log-normal model. *Marketing Letters*, *25*(2), 219-234.

Trinh, G., Khan, H., & Lockshin, L. (2020). Purchasing behaviour of ethnicities: Are they different?. *International Business Review*, *29*(4), 101519.

Trinh, G., Corsi, A., & Lockshin, L. (2019). How country of origins of food products compete and grow. *Journal of Retailing and Consumer Services*, *49*, 231-241.

Tsionas, E. G. (2010). Bayesian analysis of Poisson regression with lognormal unobserved heterogeneity: With an Application to the Patent-R&D Relationship. *Communications in Statistics—Theory and Methods*, *39*, 1689-1706.

Wilkinson, J. W., Trinh, G., Lee, R., & Brown, N. (2016). Can the negative binomial distribution predict industrial purchases?. *Journal of Business & Industrial Marketing*, *31*(4), 543-552

Winkelmann, R. (2008). *Econometric analysis of count data*. *5th ed*. Berlin: Springer-Verlag.

Wu, C., & Chen, H.-L. (2000a). Counting your customers: Compounding customer's in-store decisions, interpurchase time, and repurchasing behaviour. European *Journal of Operational Research*, *127*(1), 109-119.

Wu, C., & Chen H.-L. (2000b). A consumer purchasing model with learning and departure behaviour. *The Journal of the Operational Research Society*, *51*(5), 583-591.

Zufryden, F. (1977). A composite heterogeneous model of brand choice and purchase timing behaviour. *Management Science*, *24*(2) 121–136.

Zufryden, F. (1978). An empirical evaluation of a composite heterogeneous model of brand choice and purchase timing behaviour. *Management Science*, *24*(7) 761–773.