ESTIMATING RECREATION DEMAND WITH ON–SITE DATA: AN APPLICATION OF TRUNCATED AND ENDOGENOUSLY STRATIFIED COUNT DATA MODELS

VILLE OVASKAINEN, JARMO MIKKOLA AND EIJA POUTA*

ABSTRACT

In travel cost models of recreation demand the dependent variable is typically the count of trips taken over the year, and data based on on-site surveys are often used. The appropriate estimator must take into account the fact that the dependent variable is a nonnegative integer from a truncated, endogenously stratified sample and that real data frequently exhibit overdispersion. In this paper truncated count data models are employed to estimate recreation demand and benefits per trip using on-site data from three adjacent forest recreation sites near Helsinki, Finland. As the Poisson model was rejected due to overdispersion in the data, the paper focuses on truncated negative binomial models with special emphasis on endogenous stratification. Resulting in somewhat better fit and smaller standard errors, the truncated and endogenously stratified Negbin model slightly outperformed the respective non-stratified model. However, adjusting for endogenous stratification had little effect on the estimated coefficients and related benefit estimates. In addition to the basic model, a specification with site specific price slopes is presented with average as well as site specific estimates for consumer surplus per predicted trip.

Keywords: choice-based sampling, consumer surplus, count data, maximum likelihood estimation, overdispersion, recreation demand, travel cost method, truncation.

~

INTRODUCTION

In travel cost models of recreation demand the dependent variable is typically the count of trips taken by the respondent over the year, and for cost-efficiency the data are often collected from an on-site sample of participants. The data therefore exhibit several problems that must be taken into account in the estimation. First, because the dependent variable is the count of trips, the only values it can take on are nonnegative integers. Second, all observed users must

* Ville Ovaskainen and Jarmo Mikkola, Finnish Forest Research Institute, Unioninkatu 40 A, FIN–00170 Helsinki. E-mail: ville.ovaskainen@metla.fi, jarmo.mikkola@metla.fi. Eija Pouta, Department of Forest Economics, P.O. Box 27, FIN–00014 University of Helsinki. E-mail: eija.pouta@helsinki.fi.

JOURNAL OF FOREST ECONOMICS, ISSN 1104–6899. © 2001 Department of Forest Economics, The Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden. Manuscript received 1999-06-29. Revised manuscript accepted 2000-11-16.

have taken at least one trip, since non-participants are not observed. That is, the sample is truncated at the zero level. Third, the on-site sampling plan is an example of what is known as choice-based sampling. Because frequent visitors are more likely to be sampled than occasional visitors, on-site data will be endogenously stratified. Fourth, the data frequently exhibit overdispersion, which is defined as variance greater than the mean.

Truncated count data models based on the discrete Poisson and negative binomial distributions have been found to be attractive tools for recreation demand modeling. Shaw (1988) introduced truncated, endogenously stratified normal and Poisson models and MLE methods with Monte Carlo experiments. Grogger & Carson (1991) provided non-stratified standard and truncated Poisson and negative binomial models with an application to real data. An empirical application of truncated Poisson and negative binomial models, with confidence intervals for the welfare measures, was provided by Creel & Loomis (1990, 1991). In Hellerstein & Mendelsohn (1993) count data models were discussed from the perspective of economic theory. Englin & Shonkwiler (1995a) completed the set of models by developing a truncated, endogenously stratified negative binomial model with applications. Recently, Englin *et al*. (1998) developed a Poisson system of demand equations.

Basically, truncated count data models allow the unbiased estimation of the unconditional demand curve and expected benefits (e.g., consumer surplus) per trip, hence the computation of aggregate social benefits of a recreation site, using non-normal count data from truncated and possibly stratified samples (Creel & Loomis, 1990; Dobbs, 1993). Furthermore, it has been suggested (Grogger & Carson, 1991; Englin & Shonkwiler, 1995a) that by correcting for both truncation and endogenous stratification one can even use data from a choice-based, on-site sample of users to infer the latent demand by the general population and estimate the use value of the site, not only for current users but for the general population.

In this paper truncated count data models are employed to estimate the demand curve for trips and consumer surplus per predicted trip using data from an on-site survey of visitors to three adjacent forest recreation sites managed

by the City of Helsinki, Finland. The paper has twin objectives. First, empirical benefit estimates are provided for an important recreational resource, the Nuuksio Lake Plain, located near Finland's most densely populated area. Besides the basic 'pooled' model, we test for site specific differences in the price slope of the demand curve to allow the value of a site to vary with differences in per trip benefits. Second, the paper considers the relative performance of alternative truncated count data models. As our data are strongly overdispersed, we focus on truncated negative binomial (Negbin) models with special emphasis on the role of endogenous stratification.

Several papers have shown that overdispersion in the data invalidates the Poisson model and have suggested Negbin models instead. However, endogenous stratification, which is always present in an on-site sample, has received relatively little attention. Englin & Shonkwiler (1995a) developed the truncated, endogenously stratified Negbin model and applied it along with the respective Poisson, but so far the model has had few other applications. We therefore examine the empirical importance of the adjustment by comparing the results of stratified and non-stratified models (for a similar consideration in the continuous context, see Dobbs, 1993).

We begin with reviewing the count data models and their estimation, and then we introduce the estimable model and the data. In the following section the estimation results are considered. We compare the truncated negative binomial model with the respective truncated, stratified model in terms of statistical performance and implications to benefit estimates. Results from the OLS and Poisson models are provided to confirm earlier findings. We end this paper with our conclusions.

COUNT DATA MODELS AND THEIR ESTIMATION

This section outlines the count data models to be applied. The reader is referred to Maddala (1983) and Cameron & Trivedi (1986) for detailed presentations of the basic count data models and their estimation, and to Shaw (1988), Grogger & Carson (1991), and Creel & Loomis (1990) for truncated models with applications to recreation demand.

The simplest model for a random variable Y with only nonnegative integer values is the Poisson model. The probability density function for the basic Poisson is

$$prob(Y=y)=F_P(y)=\exp(-\lambda)\lambda^y/y!, \quad y=0,1,... \qquad (1)$$

where λ is the Poisson parameter. The model is extended to a regression setting most easily by allowing for different λ_i which vary according to $\lambda_i = \exp(X_i\beta)$, where X_i and β are the vectors of covariates and parameters to be estimated (the exponential specification serves to restrict λ_i to be positive). The conditional mean of Y is $E(Y \mid X) = \lambda = \exp(X\beta)$ and the variance $\text{var}(Y|X) = \lambda = E(Y|X)$. Note that λ is both the mean and variance of Y, which is often a problem in application to real data. A natural extension is the negative binomial model, which allows the variance to differ from the mean. The model is

$$prob(Y=y)= F_{NB}(y)=\left[\Gamma(y+1/\alpha)/\Gamma(y+1)\Gamma(1/\alpha)\right](\alpha\lambda)^y(1+\alpha\lambda)^{-(y+1/\alpha)}, \quad y=0,1,... \qquad (2)$$

where Γ indicates the gamma function and α denotes the overdispersion parameter. The conditional mean and variance are $E(Y|X) = \lambda = \exp(X\beta)$ and $\text{var}(Y|X) = \lambda(1 + \alpha\lambda)$.

For data from an on-site sample, the model must account for sample truncation. Since non-participants are not observed, all observed users must have taken at least one trip. The probability function for the zero level truncated Poisson model is

$$prob(Y=y|Y>0)=\left[\exp(-\lambda)\lambda^y/y!\right]\left[1-F_P(0)\right]^{-1}, \quad y=1,2,... \qquad (3)$$

with conditional mean $E(Y|X,\ Y>0) = \lambda\ [1 - F_p(0)]^{-1}$. The parameters of the untruncated Poisson can be consistently estimated even in the presence of overdispersion, although the standard errors are downwardly biased (Gourieroux *et al.*, 1984b; Cameron & Trivedi, 1986). For the truncated Poisson, however, overdispersion makes the estimates biased and inconsistent (Grogger & Carson, 1991). Overdispersion can be allowed for by using the truncated Negbin

$$prob(Y=y|Y>0)= \left[\Gamma(y+1/\alpha)/\Gamma(y+1)\Gamma(1/\alpha)\right](\alpha\lambda)^{y}(1+\alpha\lambda)^{-(y+1/\alpha)}\left[1-F_{NB}(0)\right]^{-1}, \quad y=1,2,... \tag{4}$$

with conditional mean $E(Y \mid X,\ Y>0) = \lambda\ [1 - F_{NB}(0)]^{-1}$.

Finally, there exist truncated count data models that also correct for endogenous stratification. This problem is present in on-site data, since the probability of being sampled on-site depends on the frequency of visits. The truncated, endogenously stratified Poisson (Shaw, 1988) is

$$prob(Y=y|Y>0)=F_{TSP}(y)=\exp(-\lambda)\lambda^{y-1}/(y-1)!, \quad y=1,2,... \tag{5}$$

with the conditional mean $E(Y|X,\ Y>0) = \lambda + 1 = \exp(X\beta) + 1$ and variance $\text{var}(Y|X) = \lambda$. Note that if we define $w_i = y_i - 1$, the Poisson case (5) coincides with the standard Poisson (1). Consequently, standard Poisson routines can be used to estimate model (5) by maximizing $\exp(-\lambda)\ \lambda^{w}\ /w!$, $w = 0,1,...$ (Englin & Shonkwiler, 1995a). The respective truncated, stratified negative binomial model (Englin & Shonkwiler, 1995a) is

$$prob(Y=y|Y>0)=F_{TSNB}(y)= y\left[\Gamma(y+1/\alpha)/\Gamma(y+1)\Gamma(1/\alpha)\right]\alpha^{y}\lambda^{y-1}(1+\alpha\lambda)^{-(y+1/\alpha)}, \quad y=1,2,... \tag{6}$$

with $E(Y|X,\ Y>0) = \lambda + 1 + \alpha\lambda$ and $\text{var}(Y|X) = \lambda(1 + \alpha + \alpha\lambda + \alpha^2\lambda)$.

Except for (6), the models can be readily estimated using the LIMDEP econometric software package (Greene, 1998). For standard count data estimators, the statistical models fitted are $Y \sim Pois(\lambda = \exp(X\beta))$ and $Y \sim NB(\lambda = \exp(X\beta), \alpha)$, where α is the overdispersion parameter, while for truncated models, Y is observed only if $Y > 0$. For OLS the semilog form was used with the model $Y \sim N(\exp(X\beta), \sigma^2 I)$. The Poisson was corrected for endogenous stratification and truncation by using $w_i = y_i - 1$ as the dependent

variable in a standard Poisson regression. Because of overdispersion, which is a form of heteroskedasticity, the standard errors for the Poisson were corrected by using White's (1980) covariance matrix estimator.

The truncated, stratified Negbin model (6) was estimated using the User defined optimization in Limdep and the two-step QGPML estimation procedure (Gourieroux *et al*., 1984a, 1984b; Cameron & Trivedi, 1986). The reported results are based on the parameterization $\alpha_i = \alpha$ for which the conditional mean and variance are given below (6). This formulation (a counterpart of Negbin II in Cameron & Trivedi) is consistent with the Negbin estimators in Limdep and gave the best results. Taking logarithms of (6) yields the log likelihood function used,

$$\ln L = \ln\left[\Gamma\left(y_i + 1/\alpha\right)\right] - \ln\left[\Gamma\left(y_i + 1\right)\right] - \ln\left[\Gamma\left(1/\alpha\right)\right] + y_i \ln\left(\alpha\lambda_i\right) - \left(y_i + 1/\alpha\right)\ln\left(1+\alpha\lambda_i\right) + \ln\left(y_i\right) - \ln\left(\lambda_i\right). \quad (7)$$

While maximizing (7) the nuisance parameter α was held as a constant, the value of which was first estimated from a separate regression using nonlinear least squares.

ESTIMABLE MODEL AND THE DATA

The data used comprised 656 observations from an on-site survey of visitors (Pouta, 1990) conducted on the adjacent recreation sites of Luukkaa (n=327), Salmi (n=205) and Pirttimäki (n=124) in the Nuuksio Lake Plain, Finland. The sites are located at a distance of 25–35 kilometers from the center of Helsinki, managed by the City of Helsinki, and mainly used by day visitors from Helsinki (50–60%) and the neighboring towns. The sample mean of time spent on-site was 8.1 hours with a median of 3 hours.

The dependent variable is the count of trips taken to the site during the last 12 months. As the wording of the question, '*How many times did you visit this site during the last year?*', did not explicitly specify whether the current trip should be included in or excluded from the reported number, the responses contained a non-negligible amount of zeros. This suggests that people excluded the current trip, so one trip was added to all reported numbers of less than

20 trips.[1] The sample mean of the dependent variable is 6.88 trips per year. While the median of 3 trips indicates that the distribution is skewed, the mean is not particularly low. The variance is 73.6, as much as 10.7 times the mean, which strongly suggests that overdispersion is present.

Even though the data include visitors to three recreation sites, the choice of modeling approach was restricted by limitations in the data. Separate demand equations were initially estimated for each individual site, but the results were rather poor with large standard errors. Therefore, we chose to pool the observations across the sites and estimate a single demand equation. This amounts to treating the sites as one destination (cf. Creel & Loomis, 1990, p. 437) with the individual sites interpreted as multiple entry points to the area. Potential differences between visitor groups using particular entry points were tested for (see below).

Treating the three sites as one destination is justified, first, because the sites are located adjacently so that a majority of visitors face roughly the same travel cost to any individual site. All the sites also provide similar basic facilities and recreational activities (walking, hiking, camping, swimming, and fishing). Second, a system of count demand equations (Englin *et al.*, 1998) could not be estimated in the lack of key information. For this, the respondents' trips to each site in the system should be known, but the data only contained the number of trips to one site. The inclusion of substitute prices in the single demand equation was also hampered by data limitations.[2] While the omission of substitutes can bias the consumer surplus upward if substitute prices are correlated with the own-site price (e.g., Bockstael, 1995), the empirical importance of this aspect depends on other considerations related to the travel

[1] The sensitivity of the results with respect to this assumption was tested by using an opposite solution, whereby only the zeros were corrected to one. The results were not significantly affected. While the absolute value of the travel cost coefficient slightly increased (consistently for all models), the difference of the point estimates was only slightly greater than one standard error.

[2] If site specific demand equations were to be estimated, the substitute prices could be represented by the travel costs to the other two sites in the data. However, this is hampered by the apparent multicolinearity between the own and substitute prices due to the proximity of the sites. For the present pooled, single-equation approach, the substitutes would be destinations other than Nuuksio. However, the truly relevant substitute sites are difficult to identify in the presence of a common right of access to private lands since at least "imperfect substitutes" abound, and there was no respondent based information such as self-reported closest substitutes.

cost variable.

The travel cost variable, denoted *TC*, is the round-trip vehicle cost at FIM 1.00 per kilometer. The sample mean was FIM 49.25 per trip. The unit cost was chosen so as to approximate the variable cost of using the car, which was the mode of transportation for all observations used. In effect, the travel cost variable equals the round-trip distance and allows the resulting benefit estimates to be simply adjusted to any desired level of vehicle cost.[3] Even though omission of the cost of travel time tends to bias the consumer surplus downward (e.g., Cesario & Knetsch, 1970), there is no generally valid and practicable way to measure the time cost (cf. Fletcher *et al.*, 1990; Bockstael, 1995; Feather & Shaw, 1999). The complexities related to substitutes and time cost are beyond the scope of this paper. Empirically, the results can be considered fairly realistic, because the two effects work in the opposite direction.

Potential differences between the groups visiting particular parts of the Lake Plain were tested for by using site specific dummy variables, denoted *DPIRT* and *DSALMI* (Luukkaa is the reference case). Further, besides differences in the frequency of visits the visitors to different sites could react differently to increases in travel cost. To test for site-specific price slopes, we tried the variables *TC*×*DPIRT* and *TC*×*DSALMI* which interact the travel cost and the site dummy.[4] Other independent variables are: *AGE*, the respondent's age; INC, after-tax income per year; *GEND*, respondent's gender (0 = male, 1 = female); *EQUIP*, the number of recreational equipment possessed by the family out of a list of 12 alternatives; and *MONEY*, annual expenditure on outdoor recreation.

---

[3] We also tried a travel cost variable defined as the sum of vehicle-related, out-of-pocket cost divided by the party's number of persons plus the opportunity cost of travel time evaluated at one third of hourly earnings. However, the simple round-trip vehicle cost was chosen due to its superior statistical performance, i.e., better fit and smaller standard errors. This choice follows several similar individual travel cost models (e.g., Englin & Shonkwiler, 1995a,b). An implication of using the undivided travel cost per vehicle is that the estimated per-trip consumer surplus will be for the average party.

[4] Our main concern is reliably estimating the average consumer surplus for the recreational resource as a whole. While the group specific slopes also provide site-specific benefit estimates, the present trip frequency model does not really allow us to examine whether the differences relate to differences in site characteristics, because the site choice stage is not considered and the quality effects are constant.

ESTIMATION RESULTS

An Outline

The following section presents the estimation results and considers the relative performance of several estimators. Because OLS has been much used earlier despite the violation of its assumptions, OLS results are reported to illustrate the magnitude of the bias. The truncated and truncated, stratified versions of the Poisson model (TPOIS, TSPOIS) are considered next. As overdispersion proves to be present, we then focus on the non-stratified and stratified versions of the truncated negative binomial model (TNB, TSNB) to consider the importance of correcting for endogenous stratification. A discussion of the empirical per-trip benefit estimates then follows. While the TSNB fits slightly better, the TNB differs little in terms of consumer surplus estimates. We also test for site-specific price slopes and per-trip benefits.

The Relative Performance of Different Models

The estimation results are presented in Table 1. For the comparison of estimators, the same regressors are used in each model. The coefficients have the expected signs and, based mainly on the TNB and TSNB models, most of them are statistically significant at the 5% level. The price variable in particular has a significant negative coefficient in all models. The number of trips also depends significantly on the age, equipment possessed, and amount of money spent on outdoor recreation annually. The negative coefficient of income is a common empirical finding in recreation studies (e.g., Creel & Loomis, 1990), but the effect is insignificant.

The dummy variable *DPIRT* as well as the price–site interaction *TC*×*DPIRT* had significant coefficients, suggesting that both the average frequency of visits and the slope of the demand curve for Pirttimäki differ from the reference case (Luukkaa and Salmi). A likelihood ratio test against a model with no site-specific variables supported the inclusion of each variable in turn but due to strong multicolinearity, they could not be included simultaneously. Two alternative specifications are therefore reported for the TNB and TSNB models. These have identical goodness-of-fit measures and t-statistics but differ in terms of interpretation (see below). For Salmi, both site-specific vari-

TABLE 1. ESTIMATED RECREATION DEMAND CURVES AND CONSUMER SURPLUS PER TRIP BASED ON ALTERNATIVE MODELS (*t*-STATISTICS IN PARENTHESES).

	OLS (semilog) [a]	TPOIS [a]	TSPOIS [a]	TNB with *DPIRT* [a]	TNB with *TC×DPIRT* [c]	TSNB with *DPIRT* [a, b]	TSNB with *TC×DPIRT* [b, c]
Constant	0.9134 (4.922)	1.5499 (6.018)	1.3321 (4.442)	1.0372 (3.758)	0.9966 (3.645)	0.2419 (1.354)	0.2007 (1.129)
TC	–0.01103 (–5.990)	–0.01253 (–4.488)	–0.01448 (–4.504)	–0.01484 (–5.415)	–0.01401 (–5.263)	–0.01398 (–7.882)	–0.01316 (–7.572)
TC×DPIRT	-	-	-	-	–0.00885 (–3.219)	-	–0.00844 (–3.937)
DPIRT	–0.3035 (–3.114)	–0.3379 (–2.400)	–0.3904 (–2.378)	–0.3907 (–3.192)	-	–0.3700 (–3.943)	-
MONEY	0.0536 (2.203)	0.0517 (1.432)	0.0600 (1.441)	0.0727 (2.331)	0.0699 (2.221)	0.0671 (2.865)	0.0642 (2.745)
INC	0.0065 (0.250)	–0.0093 (–0.248)	–0.0106 (–0.2490)	–0.0366 (–1.126)	–0.0382 (–1.141)	–0.0305 (–1.221)	–0.0313 (–1.252)
AGE	0.0208 (6.086)	0.0176 (4.248)	0.0202 (4.256)	0.0241 (4.402)	0.0245 (4.450)	0.0221 (6.723)	0.0225 (6.818)
EQUIP	0.0178 (0.898)	0.0414 (1.388)	0.0477 (1.393)	0.0557 (2.246)	0.0546 (2.199)	0.0510 (2.680)	0.0501 (2.633)
α	n / a	n / a	n / a	1.8626 (6.548)	1.8653 (6.533)	1.8008 (3.807)	1.8008 (3.807)
Log L	–888.74	–3240.87	–3501.03	–1821.70	–1821.79	–1890.91	–1891.15
Restricted log L	–923.06	–3454.17	–3747.27	–3240.87	–3243.17	–3501.03	–3503.38
Pseudo-R^2	0.037	0.062	0.066	0.473	0.473	0.495	0.495
CS / Y, FIM	90.66	79.82	69.05	67.38	66.16	71.55	70.37

[a] *For models with one common slope CS/Y = –1/β_P is for the representative case with $\beta_P = \beta_{TC}$.*
[b] *Estimated using the QGPML procedure.*
[c] *For models with site specific slopes the reported CS/Y is the weighted average of CS/Y measures for the reference case (Luukkaa & Salmi) with $\beta_P = \beta_{TC}$ and for Pirttimäki with $\beta_P = \beta_{TC} + \beta_{TC}$×DPIRT.*

ables were excluded as *DSALMI* was insignificant and *TC×DSALMI* was too strongly correlated with the price variable.

Goodness-of-fit of the Models

In addition to the basic log-likelihood statistic, we report the pseudo-R^2 (or likelihood ratio index) $R^2_{LRI} = 1 - \ln L/\ln L_0$ (e.g., Greene, 1993; 1997). This is based on testing the improvement of the fit over a restricted model with only a constant term (i.e., with the restriction β=0). As an analog

to the standard R^2, the R^2_{LRI} summarizes the maximized and restricted log-likelihood values in a single figure bounded by zero and 1.[5]

As OLS and truncated count estimators are compared, the Negbin models have R^2_{LRI} values of 0.47–0.50 while OLS falls short of 0.04. On the other hand, the R^2_{LRI} values for the Poisson models are below 0.07, indicating that the Poisson performs only slightly, if at all, better than OLS (R^2 measures for OLS and Poisson are 0.091 and 0.089, respectively). Accordingly, both log-likelihood statistics and R^2_{LRI} suggest that all the Negbin models clearly outperform OLS as well as the Poisson. While there is less difference between various Negbin models, the endogenously stratified TSNB seems to fit slightly better than the simple TNB.

Poisson vs. Negative Binomial Models: the Role of Overdispersion

Based on a parametric restriction on the overdispersion parameter α, the Poisson model can be tested against the Negbin using the likelihood ratio test statistic $LR = -2(\ln L_R - \ln L_U)$, where the subscripts R and U stand for restricted and unrestricted models (e.g., Cameron & Trivedi, 1986; Greene, 1998). Further, testing the significance of α in the Negbin model provides a simple test for overdispersion.

The LR test statistics for TPOIS vs TNB and TSPOIS vs TSNB (with *DPIRT*) obtain values as high as 2,838.3 and 3,220.2, respectively. The t-statistics for all Negbin models indicate that α is significantly different from zero, so the data are obviously overdispersed. That is, both LR and overdispersion tests strongly reject the Poisson. This confirms earlier findings (Cameron & Trivedi, 1986; Grogger & Carson, 1991) that violation of the mean–variance equality is most serious to the performance of the Poisson. Consequently, we focus on the truncated Negbin models, which are strongly favored over the Poisson for this data.

---

[5] For a common point of reference, and for the overall rather than incremental fit for the Negbin, the *restricted* log L of the relevant Poisson (with restrictions β=0, α=0) was used as the restricted log L for both the Poisson and Negbin when computing the pseudo-R^2. In contrast, the restricted log L values in Table I follow the usual practice in Limdep (Greene, 1998), where the Negbin's restricted log L equals the log L of the respective full Poisson (with restriction α=0 only). Thus, the Chi-squared statistic in Limdep output for the Negbin readily gives the likelihood ratio (LR) statistic for testing the Negbin against the Poisson (e.g., Cameron & Trivedi, 1986).

The Importance of Endogenous Stratification

Our data are drawn from a choice-based sampling scheme. As frequent visitors are more likely than occasional visitors to be sampled on-site, frequent visitors tend to be over-represented in comparison to the visitor population. In theory, the estimation problems associated with the data can be addressed by using the truncated, endogenously stratified Negbin model (Englin & Shonkwiler, 1995a). However, this model has not become routinely applied so far even though on-site data are frequently used. While this kind of analysis for a model based on a continuous distribution is found in Dobbs (1993), we do not know of published results on the empirical importance of adjusting for stratification in the count data context. From an applied point of view, the stratified Negbin is more costly to estimate than the non-stratified TNB for which routines already exist. Therefore, it is interesting to compare the performance of the stratified and non-stratified versions of the truncated Negbin model (TNB, TSNB).[6]

Based on the pseudo-R^2 the endogenously stratified Negbin (TSNB) has a slightly better fit than the non-stratified TNB. The t-statistics also indicate that the TSNB estimates have smaller standard errors. However, although the endogenous selection is *a priori* an apparent problem with on-site data, the related adjustment had little effect on the estimated parameters. No major differences can be found between the respective TNB and TSNB models.

Since truncation and endogenous stratification are both instances of choice-based sampling, an interpretation of the small difference could be that "more complicated forms of endogenous stratification" (Pudney, 1989, p. 76) have little effect beyond sample truncation, which implies a zero sampling probability to non-visitors. Considering probability function (4), the TNB accounts for unobserved zeros by multiplying the standard probability (2) by $[1 - F_{NB}(0)]^{-1}$,

[6] Reported non-stratified and stratified models are based on the same parameterization $\alpha_i = \alpha$. Estimation differed in that the non-stratified TNB was estimated with Limdep's ML estimator (Greene 1998), while for the TSNB the two-step QGPML estimation procedure (Gourieroux *et al.* 1984a,b, Cameron & Trivedi 1986) was used. Earlier results do not suggest that the ML and QGPML estimators should systematically give different results, so the potential difference between TNB and TSNB can be assumed to reflect the impact of adjusting for stratification.

which is greater than 1 and inflates the probabilities by a constant proportion. For the TSNB in (6) the standard probability is adjusted for truncation and endogenous selection by the weighting factor y_i/λ. Since this is greater (less) than 1 as the observed value is greater (less) than the mean, the probability is inflated (deflated) for y_i above (below) mean. Although both adjustments shift the probability mass in the same direction, the way and extent they do so differs. Consequently, the conditional means and variances also differ in a way that apparently depends on the properties of the actual data. The finding that the difference need not have major effects on the estimated coefficients is similar to the conclusion in Dobbs (1993, p. 339). According to Dobbs, over-presentation of particular types of individuals in itself seems no reason to expect bias in slope coefficients.

In conclusion, the results support the truncated, endogenously stratified Negbin as the best-suited model for the present data. However, the TSNB fit only slightly better than the non-stratified TNB and suggested minor differences in the estimated coefficients. For the TSNB, we also tried an alternative parameterization $\alpha_i = \alpha_0/\lambda_i$ implying $E(Y|X, Y>0) = \lambda + 1 + \alpha_0$ and $\text{var}(Y|X) = \lambda + \alpha_0 + \alpha_0\lambda + \alpha_0^2$ (Englin & Shonkwiler, 1995a). However, the results favored the 'Negbin II' type of model (Cameron & Trivedi, 1986) for the present data since this fit the data best and gave more reasonable estimates for the overdispersion parameter in the first step of the estimation.

Estimated Consumer Surplus per Predicted Trip

From an applied point of view a central outcome of the travel cost model is the estimated net economic value per trip.[7] For the Marshallian measure, consumer surplus (willingness-to-pay over and above the amount actually paid), consider the exponential demand function or its semi-logarithmic equivalent

[7] For the interpretation and use of the benefit measures and other results in the context of truncated and possibly stratified models, see Creel & Loomis (1990), Dobbs (1993), and Englin & Shonkwiler (1995a). Formulas for the Hicksian welfare measures, compensating and equivalent variation, have been developed as well (Bockstael et al., undated). However, simple consumer surplus will do, since the Marshallian and Hicksian measures are very close when the income coefficient is small (e.g., Creel & Loomis, 1991).

$$Y = \exp\left(\beta_0 + \beta_P P + \beta_1 X_1 + \ldots + \beta_K X_K\right) \Leftrightarrow$$
$$\ln Y = \beta_0 + \beta_P P + \beta_1 X_1 + \ldots + \beta_K X_K \qquad (8)$$

where P is the price variable (i.e., travel cost) and X_k's (k = 1,..., K) denote other independent variables. Total consumer surplus is the integral of the demand function from the beginning price P_B to the choke price with zero trips, P_C. Because

$$CS = \int_{P_B}^{P_C} Y(p)dp = -Y/\beta_P$$

the formula for consumer surplus per predicted trip is

$$CS/Y = -1/\beta_P. \qquad (9)$$

Models with the dummy variable *DPIRT* impose a common slope for the demand curve, so *CS*/*Y* for the representative trip is directly obtained by using formula (9) and the travel cost coefficient. On the other hand, a representative case does not exist for the model including the interaction *TC*×*DPIRT*, once the slope differs between the sites. However, an average per trip consumer surplus measure can be computed to characterize the average per trip benefits associated with the recreational resource as a whole and to compare the specifications.[8]

Average Consumer Surplus Estimates

According to the best-fit model, the TSNB, the average consumer surplus per predicted trip (the lowermost line of Table 1) is on the order of FIM 70–72 per trip. The TNB suggests only slightly lower estimates of FIM 66–67 per trip (FIM 1.00 is roughly equivalent to USD 0.20 or EUR 0.17). As expected due to the small difference in estimated coefficients, there is little difference between the consumer surplus estimates from the respective stratified and non-strati-

[8] The travel cost coefficient now represents the slope for the reference case, Luukkaa and Salmi, while the interaction term indicates how the price coefficient for Pirttimäki differs from that. The case specific price slopes were the basis for computing the respective *CS*/*Y* measures, and the weighted average of the latter is used as the average *CS*/*Y*. Consider the TNB with *TC*×*DPIRT*, for example. Using the shares of observations (0.811 and 0.189) as the weights, we get *CS*/*Y* = 0.811(1/0.01401) + 0.189[1/(0.01401+0.00885)] = FIM 66.16.

fied Negbin models. The average *CS*/*Y* estimates from models with site specific slopes are consistent with the models imposing one common slope.

Besides the point estimates, it is useful to consider the standard errors of the benefit estimates. Following Englin and Shonkwiler (1995b), an approximation to the standard error of the consumer surplus ($1/\beta_p$) is obtained by using the second-order Taylor series approximation of the variance of $1/\beta_p$. This is

$$Var\left(1/\beta_P\right) = S^2/\beta_P^4 + 2\left(S^4/\beta_P^6\right), \qquad (10)$$

where S is the standard error of β_p. Of most interest is the comparison between the TNB and TSNB. For models with common slope, the standard errors of *CS*/*Y* are FIM 12.86 or 19.1% of the point estimate for the TNB and FIM 9.22 or 12.9% for the TSNB. That is, the TSNB yields benefit estimates with smaller standard errors than the other models, including the TNB (for OLS, TPOIS and TSPOIS the relative standard errors are 17.2%, 23.4% and 23.3%). However, the welfare measures arising from the TNB and TSNB do not differ significantly: the point estimates (FIM 66–67 vs. FIM 70–72) differ by far less than one standard error.

Even though the violation of the mean–variance equality resulted in a strikingly poorer fit for the Poisson, the coefficient estimates are quite consistent with the Negbin as regards the TSPOIS. This was expected since the TSPOIS was estimated as a standard Poisson, which is consistent despite overdispersion if sample size is large enough for asymptotic unbiasedness. However, ignoring overdispersion could result in errors of inference (cf. Grogger & Carson, 1991); the uncorrected t-values were drastically inflated in comparison to the corrected ones. In contrast, the truncated Poisson differs from the Negbin suggesting that there is a bias due to overdispersion. Finally, the OLS estimates differ considerably from the count data models due to a failure to take into account the properties of the data. The average *CS*/*Y* from OLS, FIM 90.66, exceeds the TSNB or TNB by roughly one third. This confirms earlier findings (e.g., Creel & Loomis, 1990; 1991; Hellerstein, 1991; Dobbs, 1993) that uncorrected estimators such as OLS could result in substantial overestimation of the benefit measures.

Site Specific Values

Besides the average figures considered above, models with site specific slopes provide *CS*/*Y* estimates, on the one hand, for the reference case (Luukkaa and Salmi) and, on the other hand, for Pirttimäki. In the latter case the sum of the travel cost and interaction coefficients is used. This allows the value of the sites to vary with differences in benefits per trip, not only with differences in the number of visits. Comparing these estimates with *CS*/*Y* measures computed from the models estimated separately for Luukkaa and Salmi vs Pirttimäki (see Appendix) can also give insight into the reliability of the results.

For Luukkaa and Salmi, the 'pooled' TNB and TSNB models suggest per trip benefits of FIM 71 and FIM 76, respectively (Appendix, last line). The site specific models have the expected significant price coefficients and imply estimates very similar to those from the pooled models, FIM 68 and FIM 72 per trip. For Pirttimäki, the pooled models suggest a *CS*/*Y* of FIM 44–46 per trip, while the site specific models yield somewhat higher benefits per trip, FIM 51–56. Still the results from both modeling approaches seem quite consistent.

DISCUSSION

Truncated count data models were employed to estimate recreation demand using on-site survey data. Generally, estimators based on the truncated negative binomial distribution were found to be the best-suited models for the data. Using OLS led to an overestimation of per trip benefits by roughly one third. Results from the Poisson model confirmed earlier findings on its poor fit in the presence of overdispersion (note, however, that expanded Poisson models have been developed that allow for under- as well as overdispersion; e.g., Cameron & Johansson, 1997). The paper's main focus was on the properties of different estimators. When using the results in computing the aggregate recreation benefits of the sites, it should be noted that the travel costs and consumer surplus per trip were for the vehicle and average party, respectively. Thus, the total number of annual visits should be adjusted for the average party size to reflect the number of parties visiting annually. Also, the sensitivity of the empirical benefit estimates to the level of vehicle cost could be considered.

The paper focused on the truncated negative binomial model, comparing its stratified and non-stratified versions to consider the empirical importance of endogenous stratification. According to the results, the endogenously stratified TSNB resulted in a slightly better fit and smaller standard errors than the non-stratified TNB. An interesting result, however, was that although endogenous selection was *a priori* an apparent problem with on-site data, the related adjustment had little effect on the estimated coefficients and consumer surplus per trip.

The latter finding may have convenient implications to applied work. If the objective is simply to estimate the aggregate benefits of a recreation site (consumer surplus *per predicted trip* multiplied by the total number of annual visits), the non-stratified truncated Negbin, which is easily estimated using standard econometric software, can be an acceptable tool even when on-site data are used. This may not always be the case as the importance of stratification can depend on the properties of the actual data, but additional support is given by similar findings based on a different type of distribution (Dobbs, 1993). On the other hand, the adjustment for stratification cannot be omitted if the model is to be used in simulating the expected number of trips demanded to compute the benefits *per individual* or to project future demands.

ACKNOWLEDGMENTS

The authors are indebted to anonymous referees of this Journal for helpful comments, to Jeff Englin and J.S. Shonkwiler for advice on the estimation of heteroskedasticity-consistent standard errors and endogenously stratified negative binomial models, and to two anonymous referees for their criticism and helpful suggestions on an early version of the paper. However, responsibility for any remaining errors or omissions is the authors' alone.

REFERENCES

Bockstael, N. E., Hanemann, W. M. & I. E. Strand, I. E., (Eds.), (undated). *Benefit Analysis Using Indirect or Imputed Market Methods, Vol. II*. Dept of Agricultural and Resource Economics, University of Maryland.

Cameron, A. C. & Johansson, P., 1997. Count Data Regression Using Series Expansions: With Applications. *Journal of Applied Econometrics*, 12, 203–223.

Cameron, A. C. & Trivedi, P. K., 1986. Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators. *Journal of Applied Econometrics,* 1 (1), 29–53.

Cesario, F. J. & Knetsch, J. L., 1970. Time Bias in Recreation Benefits Studies. *Water Resources Research*, 6, 700–704.

Creel, M. D. & Loomis, J. B., 1990. Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California. *American Journal of Agricultural Economics*, 72, 434–441.

Creel, M. D. & Loomis, J. B., 1991. Confidence Intervals for Welfare Measures with Application to a Problem of Truncated Counts. *The Review of Economics and Statistics*, LXXIII(2), 370–373.

Dobbs, I. M., 1993. Adjusting for Sample Selection Bias in the Individual Travel Cost Method. *Journal of Agricultural Economics*, 44, 335–342.

Englin, J. & Shonkwiler, J. S., 1995a. Estimating Social Welfare Using Count Data Models: An Application to Long-Run Recreation Demand under Conditions of Endogenous Stratification and Truncation. *The Review of Economics and Statistics*, LXXVII (1), 104–112.

Englin, J. & Shonkwiler, J. S., 1995b. Modeling Recreation Demand in the Presence of Unobservable Travel Costs: Toward a Travel Price Model. *Journal of Environmental Economics and Management*, 29, 368–377.

Englin, J., Boxall, P. & Watson, XX, 1998. Modeling Recreation Demand in a Poisson System of Equations: An Analysis of the Impact of International Exchange Rates. *American Journal of Agricultural Economics,* 80: 255–263.

Feather, P. & Shaw, W. D., 1999. Estimating the Cost of Leisure Time for Recreation Demand Models. *Journal of Environmental Economics and Management*, 38: 49–65.

Fletcher, J. J., Adamowicz, W. L. & Graham-Tomasi, T., 1990. The Travel Cost Model of Recreation Demand: Theoretical and Empirical Issues. *Leisure Sciences*, 12: 119–147.

Gourieroux, C., Monfort, A. & Trognon, A., 1984a. Pseudo Maximum Likelihood Methods: Theory. *Econometrica*, 52 (3), 681–700.

Gourieroux, C., Monfort, A. & Trognon, A., 1984b. Pseudo Maximum Likelihood Methods: Applications to Poisson Models. *Econometrica*, 52 (3), 701–720.

Greene, W. H., 1993. *Econometric Analysis, 2nd ed.* (New York: Macmillan Publishing Company).

Greene, W. H., 1997. *Econometric Analysis, 3rd ed.* (New Jersey: Prentice Hall).

Greene, W. H., 1998. *LIMDEP Version 7.0 User's Manual, revised edition* (New York: Econometric Software, Inc., Plainview).

Grogger, J. T. & Carson, R. T., 1991. Models for Truncated Counts. *Journal of Applied Econometrics*, 6, 225–238.

Hellerstein, D. M., 1991. Using Count Data Models in Travel Cost Analysis with Aggregate Data. *American Journal of Agricultural Economics*, 73, 860–866.

Hellerstein, D. & Mendelsohn, R., 1993. A Theoretical Foundation for Count Data Models. *American Journal of Agricultural Economics*, 75, 604–611.

Maddala, G. S., 1983. *Limited Dependent and Qualitative Variables in Econometrics* (New York: Cambridge University Press).

Pouta, E., 1990. *Ulkoilualueen virkistyshyötyjen taloudellinen arviointi [Economic Evaluation of the Benefits of a Recreation Site]*. University of Helsinki, Department of Forest Economics, M.Sc. thesis.

Pudney, S., 1989. *Modelling Individual Choice: The Econometrics of Corners* (Oxford, UK: Kinks and Holes, Basil Blackwell Ltd).

Shaw, D., 1988. On-site Samples' Regression, Problems of Non-negative Integers, Truncation, and Endogenous Stratification. *Journal of Econometrics*, 37, 211–223.

White, H., 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48 (4), 817–838.

APPENDIX. ESTIMATION RESULTS FROM SITE SPECIFIC MODELS, TRUNCATED NEGATIVE BINOMIAL MODELS (TNB, TSNB; *t*-STATISTICS IN PARENTHESES).

	Luukkaa, TNB	Salmi, TNB	Luukkaa and Salmi, TNB	Pirttimäki, TNB	Luukkaa and Salmi, TSNB	Pirttimäki, TSNB
Constant	2.1157 (6.008)	–0.7206 (–0.916)	0.8584 (3.002)	1.6765 (2.093)	0.0842 (0.434)	0.7547 (1.690)
TC	–0.0337 (–6.526)	–0.00506 (–0.850)	–0.01466 (–5.246)	–0.01969 (–1.949)	–0.01388 (–7.520)	–0.01777 (–2.880)
MONEY	0.0192 (0.481)	0.1395 (2.115)	0.0751 (2.215)	0.0311 (0.401)	0.0695 (2.664)	0.0260 (0.477)
INC	0.0251 (0.467)	–0.0217 (–0.288)	0.0084 (0.211)	–0.1866 (–2.606)	0.0099 (0.354)	–0.1677 (–2.917)
AGE	0.0144 (2.253)	0.0406 (3.462)	0.0233 (4.117)	0.0243 (1.378)	0.0217 (6.053)	0.0223 (2.636)
EQUIP	0.0546 (1.702)	0.0687 (1.227)	0.0626 (2.358)	0.0276 (0.387)	0.0583 (2.828)	0.0210 (0.420)
α	1.2561 (5.854)	2.0521 (3.358)	1.7423 (6.216)	2.1975 (2.100)	1.7009 (3.299)	2.2007 (1.723)
Log L	–929.77	–550.33	–1495.97	–322.42	–1551.29	–333.97
Restricted log L	–1540.08	–974.81	–2646.52	–569.14	–2856.54	–615.49
Pseudo-R^2	0.472	0.486	0.473	0.466	0.496	0.491
CS/Y, site specific model	29.63	(197.63) [b]	68.21	50.79	72.07	56.26
CS/Y, pooled model [a]	-	-	71.38	43.74	75.98	46.30

[a] *Computed from the results in Table I using* $CS/Y = -1/\beta_P$, *where* $\beta_P = \beta_{TC}$ *for the reference case (Luukkaa and Salmi) and* $\beta_P = \beta_{TC} + \beta_{TC \times DPIRT}$ *for Pirttimäki.*

[b] *Price coefficient not significantly different from zero.*