METSÄNTUTKIMUSLAITOKSEN TIEDONANTOJA 683, 1998
FINNISH FOREST RESEARCH INSTITUTE, RESEARCH PAPERS 683, 1998

Estimating recreation demand with on–site data

An application of truncated vs truncated, endogenously stratified count data models

Ville Ovaskainen
Jarmo Mikkola
Eija Pouta

HELSINGIN TUTKIMUSKESKUS – HELSINKI RESEARCH CENTRE

METLA

METSÄNTUTKIMUSLAITOKSEN TIEDONANTOJA 683, 1998
FINNISH FOREST RESEARCH INSTITUTE, RESEARCH PAPERS 683, 1998

Estimating recreation demand with on–site data

An application of truncated vs truncated, endogenously stratified count data models

Ville Ovaskainen
Jarmo Mikkola
Eija Pouta

HELSINGIN TUTKIMUSKESKUS – HELSINKI RESEARCH CENTRE

Ovaskainen, V., Mikkola, J. & Pouta, E. 1998. Estimating recreation demand with on-site data: An application of truncated vs truncated, endogenously stratified count data models. Metsäntutkimuslaitoksen tiedonantoja 683. Finnish Forest Research Institute, Research Papers 683. ISBN 951–40–1623–8, ISSN 0358–4283. 27 p.

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 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 recreation sites near Helsinki, Finland. As the data are overdispersed, the paper focuses on truncated negative binomial models with special emphasis on endogenous stratification. Among the truncated and truncated, stratified models compared the (non-stratified) truncated Negbin model was the best suited for the data. Surprisingly, adjusting for endogenous stratification had little effect on estimated parameters and resulted in a slightly poorer fit. The results supported a specification with different price slopes for each site. Estimates of consumer surplus per predicted trip are provided for the sites on an average as well as for individual sites.

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

Tiivistelmä:

Virkistysalueiden hyötyjen taloudellinen arvottaminen: matkakustannusmallien sovellus kävijäaineistoon

Luonnon markkinattomien hyötyjen taloudellista arvottamista varten on kehitetty useita tekniikkoja. Virkistyskäytön arvottamiseen on eri maissa jo kauan käytetty matkakustannusmenetelmää, jota nykyisin sovelletaan yleensä yksilökohtaisiin havaintoihin. Mm. kustannussyistä tutkimuksissa käytetään usein alueella koottua kävijäaineistoa. Kysyntämallin selitettävä muuttuja on käyntikertojen lukumäärä vuotta kohti ja selittäjinä matkakustannukset sekä kävijän ja mahdollisesti alueen ominaisuudet. Käytettävän estimointimenetelmän on otettava huomioon kävijäaineiston erikoispiirteet: selitettävä muuttuja on kokonaisluku, sen arvo on vähintään yksi (katkaistu jakauma) ja tiheästi käyvät ovat yliedustettuina, koska aluekyselyssä poimintatodennäköisyys on verrannollinen käyntitiheyteen. Lisäksi muuttujan varianssi on usein keskiarvoa suurempi (overdispersion).

Viime aikoina on havaittu, että virkistyskäytön kysynnän estimointiin sopivat hyvin tiettyihin kokonaislukuarvoisiin jakaumiin perustuvat, nimellä 'truncated count data models' tunnetut mallit. Virkistysalueen arvoa arvioitaessa kiinnostuksen kohteena eivät ole otokseen kuuluvien saamat ex post -hyödyt, vaan koko kävijäpopulaatiota edustavat mitat odotetulle hyödylle käyntikertaa kohti. Mallien perusetuna on, että tällainen mitta (kuluttajan ylijäämä ennustettua käyntiä kohti) saadaan harhattomasti estimoiduksi myös kävijäaineistosta.

Tässä tutkimuksessa käytetty aineisto on kerätty Helsingin kaupungin Luukkaan, Pirttimäen ja Salmen ulkoilualueilla kävijäkyselyllä. Vastaajat olivat vierailleet alueella viimeisen vuoden aikana keskimäärin 6,9 kertaa ja käynnin keskipituus oli 8 tuntia (mediaani 3 tuntia). Tutkimuksen päätavoitteena oli testata em. malleja, joita ei ole Suomessa käytetty aikaisemmin. Lisäksi tuotettiin suuruusluokka-arvioita käyntikerran markkamääräiselle arvolle, jollaisia tuloksia ei myöskään ole meillä esitetty juuri lainkaan.

Tutkimuksessa keskityttiin negatiiviseen binomijakaumaan perustuviin malleihin ja vertailtiin toisaalta vain jakauman katkaisun, toisaalta myös käyntitiheydestä riippuvan valikoitumisen (stratifikaation) huomioon ottavia malleja. Koska selitettävän muuttujan varianssi oli selvästi keskiarvoa suurempi, näiden yhtäsuuruutta edellyttävä Poisson-malli ei soveltunut.

Tulosten mukaan aineistoon sopi parhaiten (stratifioimaton) katkaistu Negbin-malli. Käyntitiheydestä riippuvan valikoitumisen huomioon ottaminen ei juuri vaikuttanut saatuihin kertoimiin. Eri muuttujayhdistelmistä parhaaksi osoittautui malli, jossa matkakustannusmuuttujan kerroin ja sen avulla saatava käyntikerran arvo oli kullekin alueelle erisuuruinen. Käyntikerran arvot, jotka ovat lähinnä suuntaa-antavia, olivat kaikille alueille keskimäärin suuruusluokkaa 45–55 mk sekä alueittain Luukkaalle vajaat 30 mk ja Pirttimäelle 50–55 mk. Salmelle aineisto ei antanut hyväksyttävää estimaattia. Kaikkiaan käytetty menetelmä näyttää sovellettavuudeltaan lupaavalta.

Asiasanat: virkistyskäytön kysyntä, kuluttajan ylijäämä, kokonaislukumuuttujat, katkaistu jakauma, valikoituminen, ylihajonta, matkakustannusmenetelmä, suurimman uskottavuuden estimointimenetelmä

Authors: Ville Ovaskainen and Jarmo Mikkola, Finnish Forest Research Institute, Helsinki Research Centre, Unioninkatu 40 A, FIN–00170 Helsinki, email: ville.ovaskainen@metla.fi, jarmo.mikkola@metla.fi; Eija Pouta, University of Helsinki, Dept. of Forest Economics, P.O. Box 24, FIN–00014 University of Helsinki, email: eija.pouta@helsinki.fi.

Publisher: Finnish Forest Research Institute

Approved by Matti Kärkkäinen, Research Director, 28.4.1998

Distribution: Finnish Forest Research Institute, Unioninkatu 40 A, 00170 Helsinki, Finland, tel. +358-9-857 051, fax +358-9-625 308.

Acknowledgments: We wish to thank Jeff Englin and J.S. Shonkwiler for advice on the estimation problems and two anonymous referees for their criticism and valuable suggestions to an earlier version. We also acknowledge comments by the participants of the Symposium on Forest Resource Economics, Umeå, 1997, and a joint seminar of FFRI and Dept. of Forest Economics, Helsinki, 1997, as well as Pertti Ruuska for the estimations of the earlier version.

1 Introduction

One of the oldest nonmarket valuation techniques is the travel cost method (TCM), which first became known as the Clawson–Knetsch (1966) zonal approach. Several versions of the method have now been used for decades in the economics of outdoor recreation for estimating the demand curve for recreational trips and the related consumer surplus measure of economic value (see Smith & Kaoru 1990, Walsh et al. 1992, and Smith 1993 for reviews). Following a shift from the original aggregate model to the use of micro data, the recent years have seen a major development in the estimation techniques of recreation demand models.

In travel cost models of recreation demand the dependent variable is typically the count of trips taken by the respondent over the year. 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 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.

For this kind of data estimators based on the continuous normal distribution and uncorrected for sample truncation, such as ordinary least squares (OLS), are most likely to yield biased parameter estimates. Although its statistical assumptions are clearly violated, OLS was in fact used until lately. Consequently, most of the early TCM results based on individual data are potentially biased and suspect for overestimating the consumer surplus.

Recently, *truncated count data models* based on discrete Poisson and negative binomial distributions have been found as attractive tools for recreation demand modeling. Shaw (1988) presented truncated, endogenously stratified normal and Poisson models and MLE methods with Monte Carlo experiments. Grogger and Carson (1991) presented 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 and Loomis (1990, 1991). In Hellerstein and Mendelsohn (1993) count data models were discussed from the perspective of economic theory. Finally, Englin and Shonkwiler (1995) completed the set of models by developing a truncated, endogenously stratified negative binomial model with applications.

When estimating the benefits associated with a recreation site, we are not interested in *ex post* benefits received by the persons in the sample. Rather, we wish to estimate expected benefits (e.g., consumer surplus) per trip for the population of users as a whole that can be used to compute aggregate social benefits (Creel & Loomis 1990, Dobbs 1993). Accordingly, the basic advantage of truncated count data models is that they allow the unbiased estimation of the unconditional demand curve and consumer surplus per predicted trip with truncated, possibly stratified data. Furthermore, it has been suggested (Grogger & Carson 1991, Englin & Shonkwiler 1995) that by correcting for both truncation and stratification and simulating with population rather than sample means, one can even infer the latent demand by the general population and estimate the use value of a site not only for current users but for the general population with data from a choice-based, on-site sample of users.

In this paper truncated count data models are employed to estimate the demand curve for trips and consumer surplus per trip using data from an on-site survey of visitors to three adjacent recreation sites managed by the City of Helsinki. The paper has twin objectives. First, empirical benefit estimates are provided for this set of intensively used recreation sites. Two specifications are compared to test whether the price slope of the demand curve differs between individual sites. Second, the paper provides evidence on the relative performance of alternative truncated count data models and discusses several estimation issues.

As our data are strongly overdispersed, we focus on truncated negative binomial (Negbin) models with special emphasis on adjustments required to correct for endogenous stratification. Several papers have shown that overdispersion in the data seriously invalidates the Poisson and suggested Negbin models instead. However, endogenous stratification which is always present in an on-site sample has received relatively little attention in this context. Englin and Shonkwiler (1995), who developed the truncated, stratified Negbin model and applied it along with the respective Poisson, did not directly indicate the empirical importance of the related adjustment, as the results of stratified and non-stratified models were not compared (for the continuous context, see Dobbs 1993).

Section 2 reviews the count data models and their estimation, and section 3 introduces the estimable model and the data. In section 4 the estimation results are considered. We compare truncated and truncated, stratified negative binomial models in terms of statistical performance and implications to benefit estimates. Results from OLS and Poisson models are also presented to confirm earlier findings. Section 5 concludes the paper.

2 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 and Trivedi (1986) for detailed presentations of the basic count data models and their estimation, and to Shaw (1988), Grogger and Carson (1991), and Creel and 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

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

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 $E(Y|X) = \lambda = \exp(X\beta)$ 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 with real data. A natural extension is the negative binomial model which allows the variance to differ from the mean. The model is

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

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 is

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

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, instead, overdispersion makes the estimates biased and inconsistent (Grogger & Carson 1991). Overdispersion can be allowed for by using the truncated Negbin

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

$$y = 1,2,...$$

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

Finally, truncated count data models exist that 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 model (Shaw 1988) is

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

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 1995). The respective truncated, stratified negative binomial model (Englin & Shonkwiler 1995) is

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

$$y = 1,2,...$$

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 1995). 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, and for truncated models Y is observed only if $Y > 0$. For OLS the semilog form was used with the model $Y \sim N(\exp(X\beta), \sigma^2 I)$. In the Poisson case endogenous stratification and truncation were corrected for by using $w_i = y_i - 1$ as the dependent variable in a standard Poisson regression. Due to 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, endogenously stratified negative binomial model (6) was estimated using the User defined optimization in Limdep and the QGPML estimation procedure (Gourieroux et al. 1984a, 1984b, Cameron & Trivedi 1986). This is a two-step procedure with α computed in a separate regression. The reported results are based on the parameterization $\alpha_i = \alpha$ (i.e., α is an estimated constant) for which the conditional mean and variance are those given below (6). This formulation (a truncated, stratified counterpart of Negbin II in Cameron & Trivedi) is consistent with the negative binomial estimators used in Limdep and it gave the best results.

3 ESTIMABLE MODEL AND THE DATA

The data used comprised 656 observations from an on-site survey of visitors (Pouta 1990) conducted on the recreation sites of Luukkaa (n=327), Salmi (n=205) and Pirttimäki (n=124). The sites are managed by the City of Helsinki and, located next to the capital region at 25–35 kilometers from the center of Helsinki, mainly used by day trippers from relatively short distance. The sample mean of time spent on-site was 8.1 hours with a median of 3 hours.

Separate models for each site were tried, but the results were rather poor with large standard errors. Therefore, the observations were pooled across the sites into a single data set. This is justified as the areas are located adjacently so that most of the users have practically the same distance to any one of them, and because all of them provide opportunities for the same main activities (walking, hiking, camping, swimming and fishing at small lakes). On the other hand, there are differences in the landscape, scenery and constructed facilities. Incorporating the between-site variation is likely to improve the reliability of the coefficients and might also provide some information on the effects of site characteristics.

The dependent variable of the demand model is the count of trips taken to the site during the last 12 months. The respondents were asked the question 'How many times did you visit this site during the last year?'. As the phrasing did not specify whether the current trip should be included or excluded, the responses contained a significant amount of zeros. This suggests that people excluded the current trip, so all reported numbers of less than 20 trips were added by one trip. The sample mean of the dependent variable is 6.88 trips per year (for distributions and descriptive statistics, see Appendix I and II). While the median of 4 trips indicates that the distribution is rather skewed, the mean is not particularly low. The variance is 73.63, as much as 10.7 times the mean. Since this strongly suggests that overdispersion is present, the data exhibit all the estimation problems typical of recreation demand models based on on-site data.

The travel cost variable, denoted by 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 as a 'best guess' estimate for the variable cost of using the car, since the mode of transportation was by car for all observations used. A small number of cyclists (n=13) and visitors arriving by bus (n=22) were omitted for simplicity of interpretation, but this had no impact on the results. In addition to providing the most plausible consumer surplus estimates directly, the price variable actually equals the round-trip distance and allows the benefit estimates to be simply adjusted to any desired level of vehicle cost.[1]

Several variables for site characteristics were tried, such as kilometers of shoreline, kilometres of constructed trails, and the existence and number of developed campgrounds. However, the differences seemed to be captured best by site specific dummy variables, denoted DSALMI and DPIRT (Luukkaa is the reference case). Further, besides differences in the frequency of visits (i.e., position of the demand curve) the visitors could react differently to increases in travel cost. To test for site specific price slopes, we included the variables TC*DSALMI and TC*DPIRT (travel cost times site dummy) which interact price and site (cf. price*gender in Englin & Shonkwiler 1995). Other independent variables are: AGE, the respondent's age; INC, the 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, money spend on outdoor recreation annually.

4 Estimation results

4.1 An outline

Section 4.2 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 related bias. The simplest count data model is the Poisson, and its truncated and truncated, stratified versions (TPOIS,

---

[1] 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 vehicle cost was chosen due to its superior statistical performance, i.e., better fit and smaller standard errors. Regarding the average level of travel cost, both specifications had roughly the same mean, since leaving the vehicle cost undivided compensates for the failure to explicitly recognize the cost of time.

TSPOIS) are considered next. As overdispersion proves to be present, we then focus on the negative binomial models and compare the truncated and truncated, stratified (TNB, TSNB) models to consider the importance of correcting for endogenous stratification. The models' empirical implications (i.e., benefit estimates) are considered in section 4.3. For the most reliable estimate of consumer surplus per trip we refer to the best fitting model, TNB. Two spesifications are compared to test for site specific price slopes and per trip benefits.

4.2 The relative performance of different models

The estimation results are provided in Table I. For the comparison of different models, the same list of regressors was included in each. The coefficients have the expected signs and, based mainly on the full and restricted TNB models, most of them are statistically significant at the 5% or 10% levels. In particular, the price variable has a significant negative coefficient for all models. The number of trips also depends significantly on the visitor's age, equipment possessed, and amount of money spent on outdoor recreation annually.

Further, the significant coefficients for the site specific dummy variables and interaction terms in the full models indicate that both the position and slope of the demand curve for Salmi and Pirttimäki differ from the reference case; the visitors to these two take fewer trips per year but are less easily priced out. The negative coefficient for income is a usual empirical finding in recreation demand studies (e.g., Creel & Loomis 1990), but the effect fails to be significant. The gender is clearly insignificant.

Table I. Estimated recreation demand curves and consumer surplus per trip based on alternative models (t-statistics in parentheses).

	Full OLS (semilog)	Full TPOIS	Full TSPOIS	Restricted TNB	Full TNB	Restricted TSNB[a]	Full TSNB[a]
constant	1.625	2.549	2.478	1.408	2.0493	0.531	1.389
	(5.736)	(7.225)	(6.114)	(3.462)	(5.189)	(1.913)	(4.388)
TC	-0.02386	-0.03242	-0.03724	-0.01816	-0.03548	-0.01710	-0.03423
	(-6.084)	(-6.349)	(-6.524)	(-4.525)	(-6.242)	(-6.717)	(-7.811)
TC*DSALMI	0.02007	0.03027	0.03471		0.03242		0.03141
	(3.946)	(4.743)	(4.796)		(4.297)		(5.528)
TC*DPIRT	0.01488	0.01825	0.02059		0.01753		0.01737
	(2.115)	(1.602)	(1.550)		(1.959)		(2.210)
MONEY	0.0484	0.041	0.047	0.063	0.066	0.059	0.059
	(1.999)	(1.226)	(1.248)	(1.901)	(1.974)	(2.197)	(2.192)
GEND	-0.095	-0.128	-0.148	-0.104	-0.084	-0.108	-0.092
	(-1.218)	(-1.297)	(-1.301)	(-0.953)	(-0.804)	(-1.254)	(-1.049)
INC	-0.006	-0.028	-0.033	-0.046	-0.049	-0.040	-0.045
	(-0.216)	(-0.754)	(-0.769)	(-1.299)	(-1.355)	(-1.369)	(-1.488)
AGE	0.020	0.017	0.019	0.023	0.023	0.021	0.021
	(5.927)	(4.087)	(4.093)	(4.085)	(4.113)	(5.620)	(5.587)
EQUIP	0.023	0.048	0.054	0.052	0.056	0.048	0.051
	(1.185)	(1.729)	(1.738)	(2.031)	(2.186)	(2.207)	(2.337)
DSALMI	-0.997	-1.318	-1.502	0.249	-1.500	0.228	-1.435
	(-3.469)	(-3.938)	(-3.930)	(1.363)	(-3.361)	(1.906)	(-4.465)
DPIRT	-0.901	-0.977	-1.103	-0.344	-1.012	-0.330	-0.987
	(-3.030)	(-2.120)	(-2.075)	(-2.677)	(-2.713	(-2.993)	(-2.970)
α	n/a	n/a	n/a	1.836	1.672	1.996	1.341
				(6.516)	(6.985)	(6.276)	(6.989)
Log L	-880.05	-3130.47	-3374.68	-1820.31	-1810.56	-1881.21	-1888.97
Restricted (β=0) log L	-923.06	-3454.17	-3747.27	-3226.39	-3130.47		
LR index	0.047	0.094	0.099	0.436	0.422		
CS/Y, FIM[b]	67.68	51.24	44.45	55.08	45.39	58.49	47.32

[a] Estimated using the QGPML procedure. [b] For restricted models CS/Y = $-1/\beta_P$ is for the representative case with $\beta_P = \beta_{TC}$. For full models CS/Y is an average figure computed using the weighted average $\beta_P = d_0\beta_{TC} + \Sigma d_k(\beta_{TC} + \beta_{TC*Dk})$, k=1,2, where d_k (k=0,1,2) is the proportion of site k of all observations.

Goodness-of-fit of the models

In addition to the basic log-likelihood statistic, we also report the likelihood ratio index LRI = 1 – ln L/ln L_0 (e.g., Greene 1993). This is based on testing the model against the hypothesis that all coefficients are zero (i.e., on the improvement of fit as compared to a model with only a constant term). As an analog to the R^2, the LRI simply displays the information in the maximized and restricted log-likelihood values in a single figure bounded by zero and 1.

Comparing OLS and truncated count data estimators, the TNB models have LR indices of 0.42–0.44 while OLS falls short of 0.05. That is, the Negbin models clearly outperform OLS. On the other hand, the LR indices for the Poisson models are below 0.10 indicating that the Poisson performs only slightly better than OLS (R^2 for OLS and Poisson models are 0.109, 0.135 and 0.123, respectively). Accordingly, both log-likelihood statistics and LR indices suggest that all negative binomial models fit substantially better than the Poisson.

Poisson vs negative binomial models: the role of overdispersion

The fit measures suggested that the Negbin model is superior. More formally, the Poisson vs Negbin models can be compared using the likelihood ratio (LR) test based on a parametric restriction on the overdispersion parameter α. The statistic LR = –2(ln L_R – ln L_U), where the subscripts R and U stand for restricted and unrestricted models, is distributed $\chi^2(1)$ (e.g., Cameron & Trivedi 1986, Greene 1993). Further, testing the significance of α in the Negbin model provides a simple test for overdispersion.

The LR test statistics for TPOIS vs full TNB and TSPOIS vs full TSNB obtain values as high as 2,639.8 and 2,986.9, respectively. The t-statistics for all Negbin models also indicate that α is significantly different from zero, so the data are obviously overdispersed. That is, both the LR and overdispersion tests strongly reject the Poisson. This confirms earlier findings (e.g., Cameron & Trivedi 1986, Grogger & Carson 1991) that the violation of mean–variance equality is most serious to the performance of the Poisson. Consequently, we focus on the truncated negative binomial models which are strongly favored over the Poisson for this data.

The importance of endogenous stratification

Further, our data are drawn from a choice-based sampling scheme. As frequent visitors are more likely to be sampled on-site than occasional ones, the frequent visitors tend to be over-represented in the sample as compared to the visitor population (e.g., Dobbs 1993). In theory, then, addressing all the estimation problems associated with our data calls for the truncated, endogenously stratified Negbin model by Englin and Shonkwiler (1995).

However, this model has few applications so far although on-site data are commonly used in recreation research. We do not know of any published findings that explicitly compare the stratified and non-stratified Negbin models to analyze the empirical importance of adjusting for the stratification. For a model based on a continuous distribution this kind of analysis was presented in Dobbs (1993). Second, from an applied point of view the truncated, stratified Negbin is more costly to estimate than the non-stratified TNB for which routines readily exist. Therefore, it is interesting to consider the performance of the truncated vs truncated, stratified negative binomial (TNB, TSNB) models.[2]

---

[2] The non-stratified and stratified models reported are both based on the same parameterization $\alpha_i = \alpha$. The estimation techniques differed in that the non-stratified TNB was estimated using the ML estimator in Limdep (Greene 1995), 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 reflected the impact of adjusting for stratification.

Interestingly, although the endogenous selection is *a priori* an apparent problem with on-site data, the results suggest that the related adjustment had no substantial effect on the estimated coefficients. No remarkable differences can be found as the restricted TNB and TSNB are compared, and the same is true for the full TNB and TSNB. (Poisson models do not suggest major differences either; the TSPOIS seems to give larger slopes in absolute value, but one should recall that the TPOIS is suspect for bias.) Second, based on log-likelihood values for the restricted as well as full models the non-stratified TNB seems to fit slightly better than the respective TSNB.

Truncation and endogenous stratification are instances of the same phenomenon, i.e., choice-based or endogenous sampling. Thus, an intuitive interpretation to the small difference above could be that "more complicated forms of endogenous stratification" (Pudney 1989, p. 76) have little effect beyond sample truncation, its basic form that implies a zero sampling probability to non-visitors. Consider the probability functions (4) and (6). The TNB accounts for the unobserved zeros by multiplying the standard probability (2) by $[1 - F_{NB}(0)]^{-1}$, 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 the adjustments basically work in the same direction shifting the probability mass to the right, the way and extent they do so differs. Consequently, the conditional means and variances also differ in a way that apparently depends on the parameters of the actual distribution.

However, the results suggest that the difference need not have major effects on the estimated coefficients. This finding is very similar to the results in Dobbs (1993) based on a different type of distribution. As Dobbs concludes, "over-presentation of particular types of individuals in itself seems no reason to expect bias in slope coefficients" (1993, p. 339). As the adjustment also resulted in a slightly poorer fit rather than improving it, the results support the non-stratified TNB as the best suited model for the data.

Testing for specification in the Negbin models

A remaining choice is between the two specifications denoted as 'restricted' and 'full' models. Based on the log-likelihood values of the non-stratified models, the full TNB (–1810.6) seems slightly better than the restricted TNB (–1820.3). More formally, one may use the likelihood ratio (LR) test based on parametric restrictions with respect to the two price–site interaction terms. For the non-stratified TNB, the statistic $-2(\ln L_R - \ln L_F)$ is $-2[-1820.31 - (-1810.56)] = 19.5$ which exceeds the relevant critical value (d.f. 2) at any conventional significance level. Thus, the LR test supports the full TNB with site specific slope coefficients for the price variable. For the stratified TSNB models, instead, the opposite is obviously true since the full model's log-likelihood is larger in absolute value, yet the difference is small.

As a conclusion, the results support the truncated Negbin models as the best suited for the data, in particular the full specification of the truncated, non-stratified TNB. The TSNB fit slightly worse than the TNB and both had roughly similar t-values, and from among the non-stratified TNB models the LR test supported the full model. 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 $\mathrm{var}(Y|X) = \lambda + \alpha_0 + \alpha_0\lambda + \alpha_0^2$ (Englin & Shonkwiler 1995). However, the results favored the 'Negbin II' type of model (in line with Cameron & Trivedi 1986), because the alternative formulation appeared unstable in this case. While the estimates usually seemed to differ from those reported, further conclusions could not be drawn since the results were sensitive to the choice of the starting values for the overdispersion parameter (i.e., first-step estimation method).

4.3 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.[3] We use the basic Marshallian measure, consumer surplus, defined as the willingness-to-pay over and above the amount actually paid. To derive this measure, consider the exponential demand function or its semi-logarithmic equivalent

$$(7) \quad Y = \exp(\beta_0 + \beta_P P + \beta_1 X_1 + \ldots + \beta_K X_K) \Leftrightarrow \ln Y = \beta_0 + \beta_P P + \beta_1 X_1 + \ldots + \beta_K X_K,$$

where P is the price variable (i.e., travel cost) and X_k's (k = 1,..., K) denote other independent variables. Total consumer surplus for the representative visitor is the integral of the demand function from the beginning price P_B to the choke price with zero trips, P_C. Because it can be shown that $CS = \int_{P_B}^{P_C} Y(p)dp = -Y / \beta_P$, the formula for consumer surplus per predicted trip is

$$(8) \quad CS/Y = -1 / \beta_P.$$

The restricted models impose a common slope for the demand curve independent of site, so CS/Y for the representative trip is directly obtained by using formula (8) and the coefficient for the travel cost. For the full models, instead, a 'representative' case does not really exist once the slope of the demand curve differs between the sites. However, a measure of average

[3] For useful discussion on the interpretation and use of the benefit measures and other results in the context of truncated and possibly stratified models, see Creel and Loomis (1990), Dobbs (1993) and Englin and Shonkwiler (1995). Formulas for the Hicksian welfare measures compensating and equivalent variation have been developed as well (Bockstael et al., undated). However, the 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).

per trip consumer surplus can be computed to characterize the per trip benefits associated with the sites on an average and to facilitate comparison with the restricted model.[4]

Estimates for the sites on an average

The estimated consumer surplus per predicted trip is displayed on the lowermost line of Table I. The model considered the best suited and most reliable, the full TNB, suggests an average consumer surplus of FIM 45.39 per trip. The estimate from the restricted TNB, the next best alternative, is FIM 55.08 (FIM 1 is roughly equivalent to USD 0.20 or ECU 0.17). That is, the restricted model suggests somewhat higher benefit estimates than the full specification. The same is true for the stratified Negbin models. The CS/Y estimates from the full and restricted TSNB are FIM 47.32 and FIM 58.49, respectively.

As is expected given the small difference in estimated parameters, CS/Y estimates from the respective endogenously stratified and non-stratified models do not essentially differ. Also, even if the violation of the mean–variance equality resulted in a poor fit for the Poisson models, the coefficients and related benefit estimates do not differ much from the Negbin. For the TSPOIS this is expected, since it was estimated as a standard Poisson which is consistent even in the presence of overdispersion if the sample size is large enough for the asymptotic unbiasedness to realize. In line with Grogger and Carson (1991), however, ignoring over-dispersion could result in serious error of inference: the uncorrected t-statistics were drastically inflated as compared to the reported values based on corrected standard errors.

---

[4] The travel cost coefficient now represents the slope for the reference case while the interaction terms indicate how the price coefficients for other sites differ from that. Based on these two the site specific price slopes were obtained, and their weighted average was eventually used in the computation of the average CS/Y. Consider the full TNB, for example. Using the individual sites' shares of observations (0.4985, 0.3125 and 0.1890) as their weights, the calculation becomes CS/Y = – 1 / [0.4985*(–0.035476) + 0.3125*(–0.035476 + 0.032422) + 0.1890*(–0.035476 + 0.017531)] = – 1 / (–0.02203) = FIM 45.39.

Instead, the OLS estimates differ significantly from the count data models due to its failure to take into account the properties of the data. OLS suggests an average CS/Y of FIM 67.68 which is roughly 50 per cent higher than the full TNB. This confirms earlier findings (e.g., Creel & Loomis 1990, 1991, Hellerstein 1991, Dobbs 1993) that the use of uncorrected estimators such as OLS could result in substantial overestimation of the benefit measures.

Estimates for individual sites

Besides the average figures considered above, the full models allow the computation of CS/Y estimates for individual sites by using the sum of the travel cost and interaction coefficients for the specific sites. This allows the value of a site to vary with differences in benefits per trip (not only with differences in the number of visits). Second, comparing these estimates with CS/Y measures computed from models estimated separately for each site can give some insight into the robustness of the results.

For Luukkaa and Pirttimäki, the common TNB model suggests per trip benefits of FIM 28.19 and FIM 55.71 (Table II, last line). Their separate models also have the expected significant price coefficients and imply very similar estimates, FIM 29.74 and 51.16. For Salmi, instead, both types of models fail to provide meaningful estimates. This is because the price coefficient in the separate model does not differ significantly from zero and the same is obviously true for the sum of the travel cost and interaction term for Salmi in the common model. As a whole, the results from both modeling approaches still seem quite consistent.[5]

---

[5] The prices for substitute sites were not included. The data did not contain any respondent based information such as self-reported closest substitutes. Second, while the sites considered could likely be assumed to be each others' closest substitutes, they are located so close to each other that the own and substitute prices would be strongly colinear. Further, LaFrance (1990) and Englin et al. (1997) point out that Marshallian uncompensated cross-price effects are not included in a utility theoretic semilog demand system.

Table II. Results for demand models for the individual sites (t-statistics in parentheses), the non-stratified truncated negative binomial model (TNB).

	Luukkaa	Salmi	Pirttimäki
constant	2.245	-1.093	2.4487
	(5.558)	(-0.773)	(2.093)
TC	-0.03363	-0.00504	-0.01955
	(-6.479)	(-0.820)	(-1.848)
MONEY	0.017	0.141	0.0005
	(0.429)	(2.119)	(0.006)
GEND	-0.070	0.124	-0.325
	(-0.570)	(0.394)	(-0.989)
INC	0.021	0.003	-0.221
	(0.388)	(0.030)	(-2.938)
AGE	0.014	0.042	0.022
	(2.192)	(3.093)	(1.192)
EQUIP	0.054	0.072	0.030
	(1.710)	(1.250)	(0.435)
α	1.252	2.052	2.133
	(5.846)	(3.355)	(2.142)
Log L	-929.64	-550.22	-321.90
Restricted (β=0) log L	-1536.44	-974.75	-564.91
LR index	0.39	0.44	0.43
CS/Y (FIM), separate model	29.74	(198.23)[b]	51.16
CS/Y (FIM), common model[a]	28.19	(327.87)[b]	55.71

[a] $CS/Y = -1/\beta_{Pk}$ where $\beta_{Pk} = \beta_{TC}$ for Luukkaa (reference case) and $\beta_{Pk} = \beta_{TC} + \beta_{TC*Dk}$ for Salmi and Pirttimäki.
[b] The price coefficient/the sum of price and interaction coefficients not significantly different from zero.

5 Discussion

Truncated count data models were employed to estimate recreation demand and benefits using on-site survey data from a set of three adjacent recreation sites near Helsinki. Based on the (non-stratified) truncated negative binomial model, which was the best suited for the data, the estimated consumer surplus was in the order of FIM 45–55 per trip. Using OLS lead to an overestimation of roughly 50 per cent. Similarly, the results from Poisson models confirmed earlier findings on its poor fit and potential error of inference given overdispersion in the data. Note, however, that expanded Poisson models (e.g., Cameron & Johansson 1997) are being developed that allow for under- as well as overdispersion.

For the magnitude of empirical benefit estimates, the results are tentative as the paper's main purpose was to consider the properties of different estimators. For actual valuation purposes the sensitivity of the results with respect to the specification and level of travel costs should be further tested. However, considering average on-site time the estimates seem quite reasonable when compared to the price of a movie ticket, for example. Also, besides the point estimates the confidence intervals for the benefit measures could be simulated (see Creel & Loomis 1991), but this was beyond the scope of the present paper.

The paper focused on the negative binomial model comparing truncated and truncated, stratified models to consider the importance of endogenous stratification. An interesting result is that although endogenous selection is *a priori* an apparent problem with on-site data, the related adjustment had no remarkable effect on the estimated coefficients and consumer surplus per trip, and also resulted in a slightly poorer fit.

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

REFERENCES

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

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

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

Clawson, M., and J. L. Knetsch (1966), Economics of Outdoor Recreation, The Johns Hopkins University Press, Baltimore, MD.

Creel, M. D., and J. B. Loomis (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., and J. B. Loomis (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., and J. S. Shonkwiler (1995), '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., P. Boxall, and D. Watson (1997), 'Modeling Recreation Demand in a Poisson System of Equations: An Analysis of the Impact of International Exchange Rates', American Journal of Agricultural Economics (forthcoming).

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

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

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

Greene, W. H. (1995), LIMDEP User's Manual, version 7.0, Econometric Software, Inc., NY.

Grogger. J. T., and R. T. Carson (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., and R. Mendelsohn (1993), 'A Theoretical Foundation for Count Data Models', American Journal of Agricultural Economics 75: 604–611.

LaFrance, J. (1990), 'Incomplete Demand Systems and Semi-Logarithmic Demand Models', Australian Journal of Agricultural Economics 34: 118–131.

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

Pouta, E. (1990), Ulkoilualueen virkistyshyötyjen taloudellinen arviointi [Economic evaluation of the benefits of a recreation site], M.Sc. thesis, University of Helsinki, Department of Forest Economics.

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

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

Smith, V. K. (1993), 'Nonmarket Valuation of Environmental Resources: An Interpretive Appraisal', Land Economics 69(1): 1–26.

Smith, V. K., and Y. Kaoru (1990), 'Signals or Noise? Explaining the Variation in Recreation Benefit Estimates', American Journal of Agricultural Economics 72: 419–433.

Walsh, R. G., D. M. Johnson, and J. R. McKean (1992), 'Benefit Transfer of Outdoor Recreation Demand Studies, 1968–1988', Water Resources Research 28(3): 707–713.

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

APPENDIX I: DESCRIPTIVE STATISTICS FOR THE VARIABLES USED

(TRIPS: number of trips per year, the dependent variable; STAY: time spent on-site, not included in the model)

Entire sample (n=656)

Variable	Valid N	Mean	Median	Min	Max	Variance
TRIPS	656	6.8811	4.0000	1.0000	50.0000	73.6286
TC	656	49.2464	48.0000	8.0000	174.0000	449.7417
MONEY	656	3.4955	3.0000	1.0000	7.0000	2.6229
GEND	656	1.4771	1.0000	1.0000	2.0000	0.2499
INC	656	4.1099	4.0000	1.0000	7.0000	2.2841
AGE	656	38.2119	38.0000	16.0000	80.0000	128.9520
EQUIP	656	4.6753	5.0000	0.0000	10.0000	3.7677
STAY	656	8.0983	3.0000	0.0000	300.0000	567.9018

Luukkaa (n=327)

Variable	Valid N	Mean	Median	Min	Max	Variance
TRIPS	327	7.4862	4.0000	1.0000	50.0000	77.4898
TC	327	39.8810	38.0000	8.0000	86.0000	174.5398
MONEY	327	3.2385	3.0000	1.0000	7.0000	2.3908
GEND	327	1.5260	2.0000	1.0000	2.0000	0.2501
INC	327	3.9755	4.0000	1.0000	7.0000	2.3246
AGE	327	36.4954	36.0000	16.0000	80.0000	139.6557
EQUIP	327	4.4771	4.0000	0.0000	10.0000	3.6551
STAY	327	7.3492	3.0000	0.0000	300.0000	772.1876

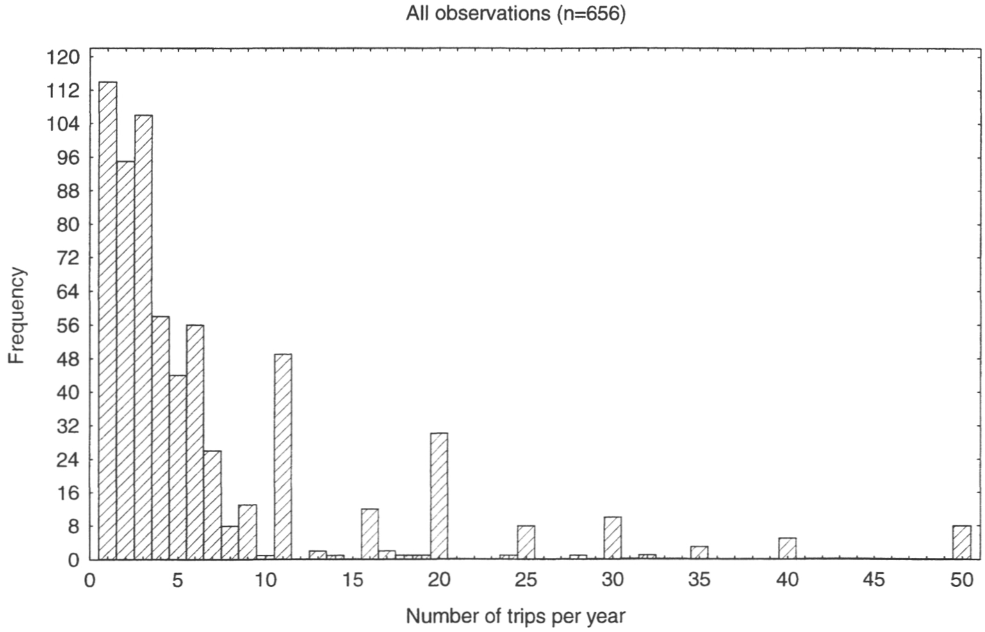
Salmi (n=205)

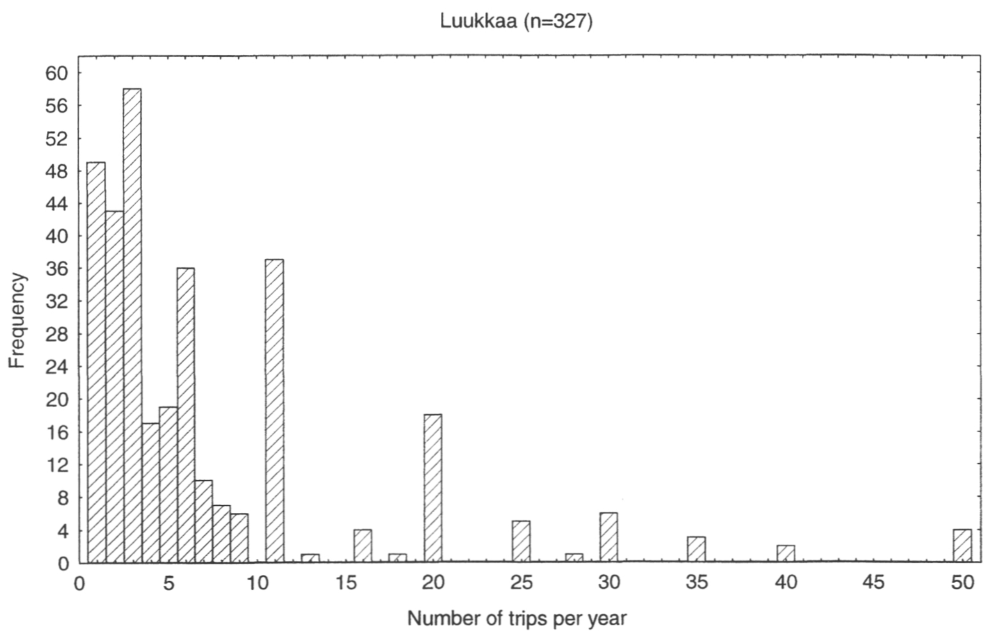
Variable	Valid N	Mean	Median	Min	Max	Variance
TRIPS	205	6.5122	4.0000	1.0000	50.0000	72.4962
TC	205	69.9525	72.0000	8.0000	174.0000	412.0991
MONEY	205	3.8195	4.0000	1.0000	7.0000	2.7075
GEND	205	1.4634	1.0000	1.0000	2.0000	0.2499
INC	205	4.1854	4.0000	1.0000	7.0000	2.0439
AGE	205	41.4829	40.0000	16.0000	71.0000	103.8490
EQUIP	205	4.6732	5.0000	0.0000	9.0000	3.9466
STAY	205	12.0293	3.5000	0.2000	160.0000	542.7102

Pirttimäki (n=124)

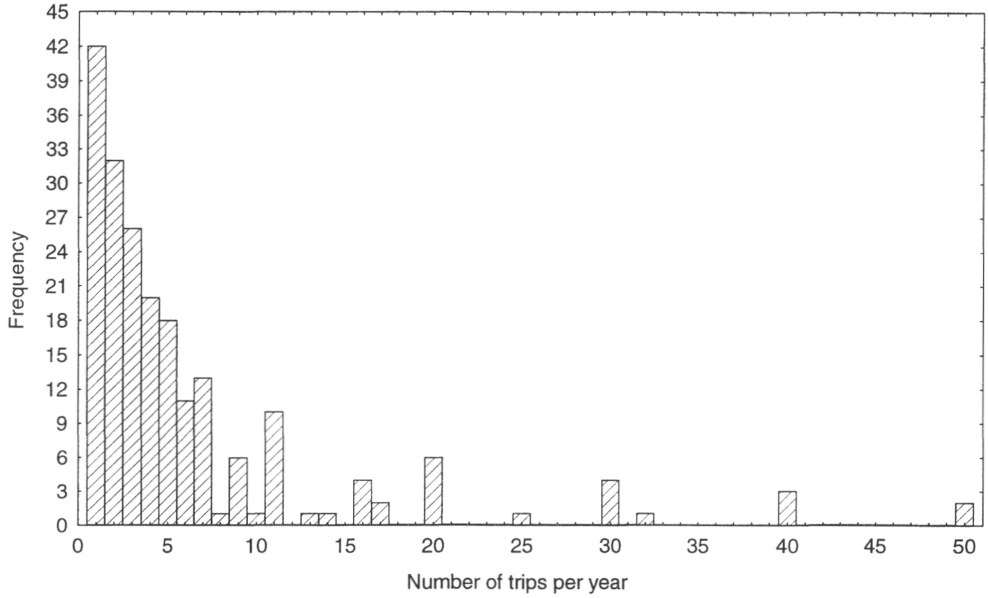
Variable	Valid N	Mean	Median	Min	Max	Variance
TRIPS	124	5.8952	3.0000	1.0000	50.0000	64.2897
TC	124	39.7123	39.4700	12.0000	72.0000	209.4907
MONEY	124	3.6371	3.0000	1.0000	7.0000	2.7697
GEND	124	1.3710	1.0000	1.0000	2.0000	0.2352
INC	124	4.3387	5.0000	1.0000	7.0000	2.5022
AGE	124	37.3307	36.5000	16.0000	72.0000	117.8654
EQUIP	124	5.2016	5.0000	1.0000	9.0000	3.4468
STAY	124	3.5750	2.5000	0.0000	48.0000	29.6043

APPENDIX II: SAMPLE DISTRIBUTIONS OF THE DEPENDENT VARIABLE (TRIPS, number of trips taken over the last 12 months)

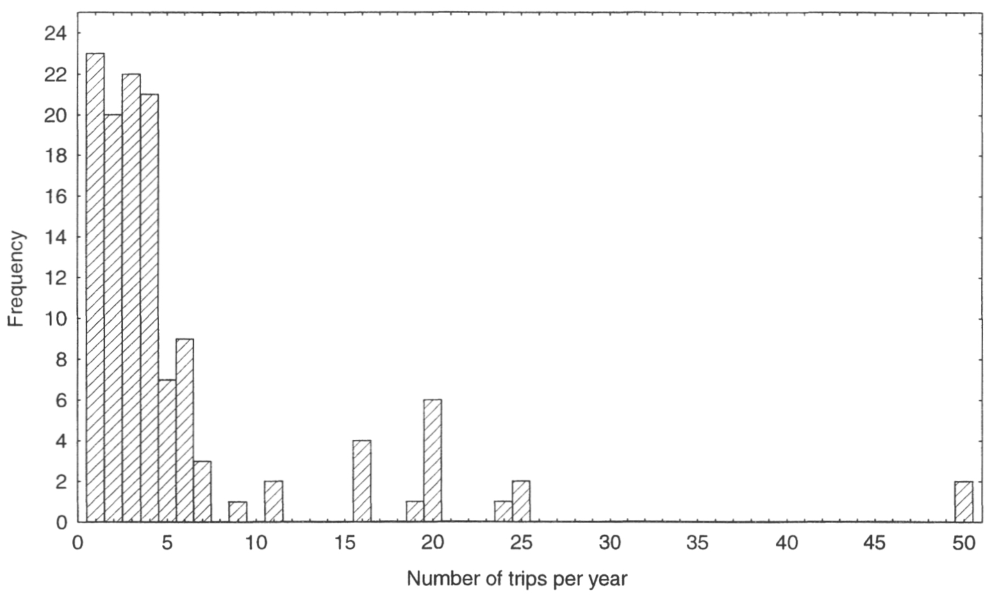




Salmi (n=205)



Pirttimäki (n=124)



ISBN 951-40-1623-8
ISSN 0358-4283
Yliopistopaino
Helsinki 1998