TRUNCATION, OVERDISPERSION AND ENDOGENOUS STRATIFICATION IN THE RECREATION DEMAND MODELS

Jan Melichar

Abstract

In this paper, we investigate and compare several truncated count data models of travel cost recreation demand. Using on-site data from Jizerske hory Landscape Protection Area in the Czech Republic, we derive the single site recreation demand model with a Poisson and a Negative Binominal specification. The models account for the truncated and overdispersed nature of the data, we also treat the problem of endogenous stratification due to the oversampling of more frequent recreation users. We also use the more flexible approach that allows the overdispersion parameter to vary according to the characteristics of the visitors and we compare it with the more restrictive approach. The results show that the negative binomial model that accounts for truncation and endogenous stratification provides the better goodness-of-fit. If don't correct overdispersion by relying on Poisson estimates, the consumer surplus will be underestimated. If we account for endogenous stratification, the consumer surplus will be adjusted appropriately upwards.

Key words: travel cost model, on-site sampling, Poisson, negative binomial, overdispersion

JEL Code: C24, D12, Q51

Introduction

The travel cost model (TCM) is a demand-based model for use of a recreation site[1]. The basic idea behind TCM stems from the need to travel to a recreational site in order to enjoy its services and amenities. Visitors who choose to visit a certain recreation site must incur the cost of overcoming the distance. The travel costs reflect both visitor's distances from their place of living to the recreation site and their opportunity costs of time. TCM rely on the insight that differences in the travel costs cause differences in quantity demanded. Currently,

[1] All models that estimate the demand for recreational services using cost of access as the price are also termed recreation demand models in the literature.

travel cost models range from the most complex corner solution models, multiple site models based on McFadden's random utility theory to much simpler single site models.

Single site model (SSM) that is developed in this paper works like a conventional downward sloping demand function. The quantity demanded for a visitor is the number of trips realized to a recreation site in a season and the price is the travel costs of reaching the site. Variation in the travel costs is generating by observing visitors living at different distances from the site. SSM is widely used to forecast the number of visits to a recreation site and to measure the welfare effects of the elimination of a recreation site, i.e. access value of a site. This value is the total consumer surplus under the single-side demand function, i.e. difference between a visitor's willingness-to-pay for a trip and the actual travel costs.

Recreation demand curve could be derived using behavioral data of the general population. More often, the data are collected "on-site", directly among visitors in the recreation site. Unfortunately, several data problems must be addressed when the recreation demand parameters are estimated from on-site sampling data. First, on-site sampling means that all members of the sample have taken at least one trip to the recreation site, i.e. the data are truncated at zero. There is no information about those individuals who have not visited the site. Second, on-site sampling leads to endogenous stratification, i.e. visitors who visit the recreation site frequently are more likely to be sampled than occasional visitors. Third, the recreation data could exhibit overdispersion, meaning the variance is greater than the mean, because a few visitors make many trips and most visitors make only a few. Fourth, the dependent variable, the number of visits to the site, is a count, it can only take on values that are non-negative integers.

While much of the early literature on TCM employed ordinary least square (OLS) regression, it has been recognized that count regression models are more appropriate to model individual recreation behavior. Shaw (1988) was the first to apply count data models to recreation demand. Shaw developed a Poisson model that corrected for truncation and endogenous stratification problems associated with on-site sampling. An empirical application of truncated Poisson and negative binomial models, with confidence intervals for the welfare measures, was provided by Creel and Loomis (1990). They found that count data models were more appropriate for estimating and predicting the demand for deer hunting in California than OLS and nonlinear least-squares estimator. Hellerstein and Mendelsohn (1993) provided a theoretical foundation for count data TCM. One of their results is algebraic formulation of consumer surplus estimation from count data models.

Englin and Shonkwiler (1995) completed the set of models by developing a truncated, endogenously stratified negative binomial model and used it to estimate long-run demand for overnight hikes in the Cascade Mountains. Ovaskainen et al. (2001) used the Englin-Shonkwiler estimator to model the demand for forest recreation trips in Finland. They also showed that overdispersion in the data invalidates the Poisson model and suggested negative binomial model instead. A further contribution was done by Martínez-Espiñeira and Amoako-Tuffour (2007) that followed work of Englin and Shonkwiler (1995) by parameterising the overdispersion parameter of negative binomial specification so that it varied according to visitor characteristics. They concluded that this flexible specification dominates the model with restricted overdispersion parameter.

Basically, truncated and endogenously stratified count data models allow the unbiased estimation of the unconditional recreation demand curve and expected benefit (consumer surplus) per visit. Hence, the computation of aggregate social benefits of a recreation site, using non-normal count data from truncated and stratified samples collected on-site, should be based on truncated and stratified models.

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 Jizerské hory Mountains in the Czech Republic. Paper considers the relative performance of alternative truncated count data models. As the data are strongly overdispersed, we focused on truncated negative binomial models with special emphasis on the role of endogenous stratification due to the oversampling of avid users. Further, flexible specification of negative binomial model by parameterising the overdispersion parameter is also introduced.

The rest of this paper is organized as follows. Section 1 describes the economic foundation of the single site model and its welfare implications. Section 2 shortly describes the study area and outlines the sampling plan. Section 3 presents sample characteristics. Section 4 describes count data models. Results including welfare estimates are provided in Section 5. Section 6 concludes.

1 Economic foundation of a basic travel cost model

Applying the single site travel cost model, we suppose that the individual's utility depends on a consumption of market goods, x, the number of trips to the recreation site, v and the environmental quality of site, q (Kolstad, 2000). Higher q is better. We also assume a weak complementarity of the trips and the environmental quality of the recreation site, q. The individual's utility is not influenced by environmental quality if the individual does not visit the site ($\partial U/\partial q = 0$ when $v = 0$). Furthermore, v is increasing with q (Alberini and Longo, 2005). We also assume that the price of x is unity. The out-of-pocket expenses related to a single trip to the recreation site (fuel expenses, admission and parking charges) are denoted as p_0. The individual works for L hours at a wage rate w. Then, the individual's utility maximization problem can be recorded as follows:

$$\max_{x,v} U(x,v,q) \qquad (1a)$$

such that

$$wL = x + p_0 v \qquad (1b)$$

Out-of-pocket expenses are not the only cost of visiting the recreation site. The individual must take time to visit the recreation site. Thus, T denotes the total time expressed in hours that is available to the individual for leisure activities and work. The travel time associated with a single round trip is t_t and the on-site time associated with single trip is t_v. The individual then faces a time-budget constraint that we can be recorded as follows:

$$T = L + (t_t + t_v)v \qquad (1c)$$

Equation (1c) can be substituted into equation (1b) in order to eliminate L and thus reduce the maximization problem, so we get:

$$wT = x + [p_0 + w(t_t + t_v)]v \equiv x + p_v v \qquad (2a)$$

where

$$p_v = p_0 + w(t_t + t_v) \qquad (2b)$$

The result of the maximization problem that is specified in equation (2) will be a demand function for trips to the recreation site:

$$v = f(p_v, q, y) \qquad (3)$$

where y is income (wT). We can assume that the demand function is log-linear and therefore we can write the demand equation as follows (Alberini and Longo, 2005):

$$v = \exp(\beta_0 + \beta_1 p_v + \beta_2 y + \beta_3 q) \qquad (4)$$

Using the demand function specified in equation (3) we can measure willingness to pay for a small change in environmental quality of the site, *q*. In fact, this is exactly the problem determined in the context of restricted demand.

Once the demand function is estimated, we can assess the consumer surplus (CS). If we follow the demand equation defined in (4) the consumer surplus is equal to (Haab and McConnell, 2002):

$$CS(p_{v0}, q_0) = -\frac{1}{\beta_1} v_0 \qquad (5)$$

where v_0 is v estimated in equation (4) at the initial level of environmental quality ($q = 0$) and the price, p_v, is defined as in equation (2b).

2 Study area and sampling strategy

The Jizerske hory Mountains Protected Landscape Area was designated in 1968 and it is one of the oldest protected areas in the Czech Republic. The surface area of the PLA is approximately 368 km^2. The JH Mts. are situated in North Bohemia, close the Polish and German borders. Two large cities of Liberec and Jablonec nad Nisou, with a combined population of almost 190 thousand, lie near the JH Mts, to their southwest. Forest ecosystems cover almost 73% of the study area (that is 270 km^2). The most common wood in the JH Mts. is spruce, representing 67% of the forest ecosystems. The JH Mts. are a favorite destination for summer and winter recreational activities such as hiking, mountain biking, cross-country skiing, and downhill skiing.

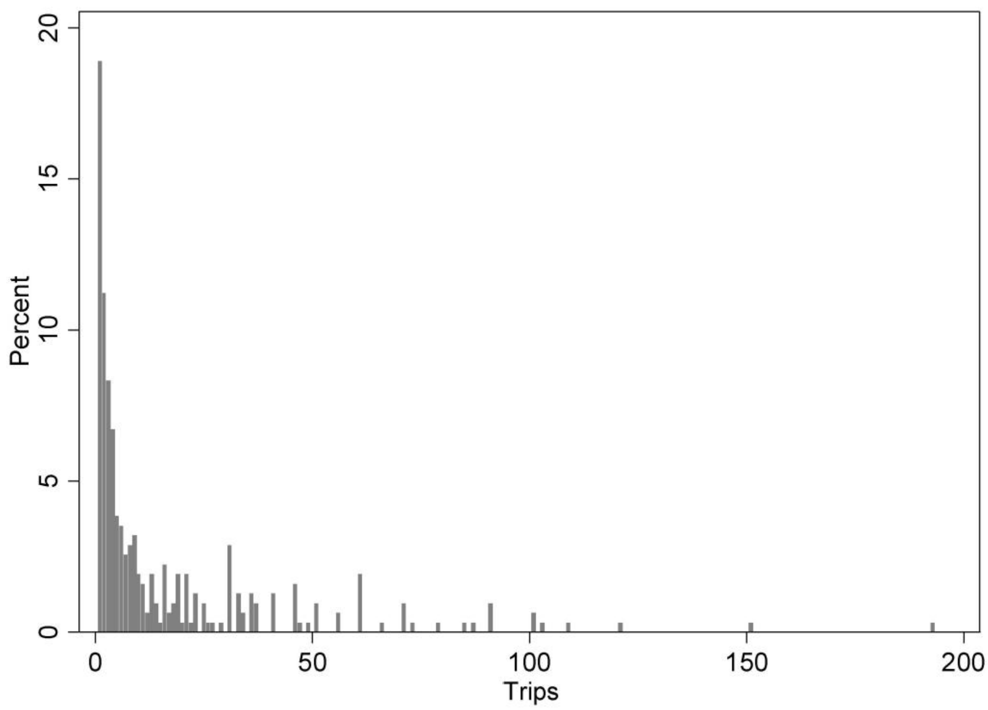
In order to apply the single site travel cost model, which relies on observed behavior of individuals, relevant information has to be obtained from visitors to the recreation site. Individual data are usually obtained by administering a survey. Therefore, a questionnaire that queried respondents about their current visit to the JH Mts., travel mode and attitudes was constructed. The survey on the JH Mts. was conducted from May to October 2005. Visitors doing summer recreational activities such as hiking and mountain biking in the central part of the JH Mts. were the target population of the survey.

The questionnaire was administrated on-site to visitors at four sites located in the central part of the Jizerske hory Mts. Respondents were intercepted randomly and interviewed by trained interviewers face-to-face on each of these four sites. Interviewing began early on the day, and respondents were selected randomly throughout the day. The survey resulted in a total of 719 completed questionnaires.

3 Sample characteristics

6% of the respondent sample had come to the JH Mts. for the first time. The average number of trips in summer season (spring to autumn 2005) to the JH Mts. is 21.07 with a median value of 7. More than 11% of the sample made only one trip over the past 12 months. More than 31% of the sample made 1, 2, 3, or 4 trips to the JH Mts. A histogram of the numbers of trips is shown in Figure 1. Approximately 60% of the respondents were on a one-day trip to the JH Mts. when interviewed. More-day trips composed the rest of the sample.

Fig. 1: Histogram of the number of trips realized to the JH Mts., N = 719



The total average cost spent on a trip per person was CZK 447. The costs included transport costs, accommodation costs[2], and opportunity costs of time. The median value is CZK 108, and the maximum is CZK 5,820. The majority of the respondents (50%) were

[2] The costs of transport and accommodation are subjective costs that were stated directly by the respondent in the questionnaire.

residents of the Liberecký Region, in which the JH Mts. are situated. Almost 26% of the sample came from Prague, the capital of the Czech Republic. The JH Mts. can be reached very easily from the capital by approximately a one hour's drive on a motorway.

More than 59% of the interviewed were male. The average age was 40 years, which is close to the medial value of 39 years. The average household size was 2.9 persons, the median value was 3 persons. The number of children per household was very low: the average was 0.5 child per family. The sample is highly educated as almost 52% of the respondents have secondary education, and 38% have a university degree. The majority of the respondents were married (51%), 37% of them were single. The economic status of the respondents is as follows. The majority, 59%, has full-time jobs, 18% are businesspeople, 8% are retired, and 6% are students. The average net individual income is 18.7 thousand CZK per month and the average net household income is 31,6 thousand CZK per month.

4 Specification of count data models

4.1 Poisson model

As one can see in Figure 1, the number of trips to the JH Mts. is proportionate to a model using a Poisson distribution. The number of trips is a count data variable which can be denoted as *Y*. If we follow Haab and McConnell (2002), then the probability function for *Y* could be expressed as:

$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{6}$$

where the parameter λ is the expected number of trips and is a function of independent variables specified in the model. The expected value and the variance of *Y* are equal to λ. The number of trips is the non-negative integer variable and therefore λ usually takes a log linear form:

$$\lambda_{ij} = \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}_1 + p_{ij}\beta_2) \tag{7}$$

where $\boldsymbol{x}$ is a vector of socio-economic variables and other variables determining the trip to the JH Mts. p_{ij} are the travel cost spent by the respondent ($i = 1, 2, \ldots, n$) on the trip. $\boldsymbol{\beta}_1$, and β_2 are unknown parameters.

The parameters in equation (7) are estimated using a maximum likelihood method. Using equation (6) and (7), the probability of observing the number of trips is estimated for each person in the sample. As Parsons (2003) suggests, the likelihood function becomes:

$$L = \prod_{n=1}^{n} \frac{e^{-\lambda_n} \lambda_n^{y_n}}{y_n!} \tag{8}$$

The on-site sample which was realized in the JH Mts. is truncated to one trip, and also the more frequent users occur in the sample (endogenous stratification). To correct the probability function we replace y_n by y_n-*1* in the basic Poisson function (6). Then the function assumes the following form (Shaw, 1988):

$$\Pr(y_n \mid y_n > 0) = \frac{e^{-\lambda_n} \lambda^{y_n - 1}}{(y_n - 1)!} \tag{9}$$

Then equation (9), instead of (6), enters the likelihood function for each individual.

4.2 Negative binomial model

When using the Poisson distribution, we assume that the expected value and the variance of *Y* are equal to *λ*. For recreational trip data, the variance is usually higher than the conditional mean, causing overdispertion in the data. The consequence of overdispersion is the fact that the standard errors in the case of the Poisson model are underestimated. The negative binominal regression model addresses the failure of the Poisson model by adding a parameter, *α*, that reflects unobserved heterogeneity among observation. The negative binominal distribution assumes the following form (Haab and McConnell, 2002):

$$\Pr(y \mid x) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda} \right)^{\alpha^{-1}} \left(\frac{\lambda}{\alpha^{-1} + \lambda} \right)^{y} \tag{10}$$

Where *Γ()* is the gamma function. The expected value of the negative binominal distribution is equal to *λ*. However, the variance of the dependent variable is $V = \lambda\ (1 + \alpha\ \lambda)$. The parameter is the overdispersion parameter. If $\alpha = 0$, no overdispersion exists. But if $\alpha > 0$, then overdispersion exists and the Poisson model is rejected in favor of the negative binominal distribution.

For the case where overdispersion is significant, the density of the negative binomial distribution truncated at zero and adjusted for endogenous stratification for the count (y) was derived by Englin and Shonkwiler (1995) as:

$$\Pr(y_i \mid y_i > 0) = \frac{y_i \Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \alpha^{y_i} \lambda_i^{y_i - 1} (1 + \alpha\lambda_i)^{-(y_i + \alpha^{-1})} \qquad (11)$$

The application (10) restricts the overdispersion parameter α to a common value for all observation, so $\alpha_i = \alpha$. We use more flexible approach that allows the overdispersion parameter to vary according to the characteristics of the visitors. Then, the parameterization of α *is* $\alpha_i = \alpha_0 / \lambda_i$ (Englin and Shonkwiler, 1995; Martínez-Espiñeira and Amoako-Tuffour, 2007).

5 Results

5.1 Model specification

Within the framework of the individual TCM, the single site demand function for the i^{th} visitors is:

$$v_i = f(tc_i, s_i, p_i, d_i, n_i, a_i) \qquad (12)$$

where v_i is the estimated number of trips of the i^{th} visitor under current environmental conditions in JH Mts. The recreation demand is a function of individual travel costs, tc_i, characteristics of substitution site, s_i, individual income, p_i, and his/her other socio-demographic characteristics, d_i, characteristics of the current visit, n_i.

5.2 Estimation results

The specified count data models were estimated using a maximum likelihood method by means of STATA 11 software, and they are reported in Table 1. The coefficient of the travels cost variable is significant in these models and negative according to the economic theory. Its magnitude is from -0.0003 (TSNB) to -0.010 (TP). The numbers of trips also increase with the respondents' age (AGE) and the distance to the substitute recreation site (SUB). The visitation is higher for one-day trips compared to more-day trips (TYPE). The numbers of trips tend to be greater among visitors with higher income (INCOME), secondary degree (SECOND) and if they are single (SINGLE). The length of the trip (LENGTH) also has a positive influence on the number of trips to the JH Mts. The number of people living in the household decreases the visitation rate to the JH Mts.

Likelihood-ratio tests confirmed that the econometric models based on the Poisson and negative binomial distribution which accounts for truncation (TP and TNB) dominates more restrictive count data models (P and NB). There is significant evidence of overdispersion, we can reject H_0: $\alpha = 0$ in all negative binomial models. The negative binominal regression model is preferred to the Poisson model. Therefore, the choice of the best model rests among the models based on the negative binomial distribution, which correct the overestimation of t-ratios and the underestimation of consumer surplus in the Poisson.

Tab. 1: Parameter estimates for on-site count data model in JH Mts.

Variable	P	TP	NB	TNB	TSNB	GNB
TC	-0.0008***	-0.0010***	-0.0004***	-0.0005***	-0.0003***	-0.0004***
SUB	0.0003***	0.0003***	0.0003*	0.0004**	0.0000	0.0002
TYPE	0.2302***	0.2020***	0.2106*	0.2373	0.3957***	0.1574
PEOPLE	-0.0447***	-0.0457***	-0.0298	-0.0254	0.0295	-0.0345
AGE	0.0125***	0.0126***	0.0102**	0.0111**	0.0026	0.0092**
SINGLE	0.0575*	0.0618**	0.0267	0.0132	0.0154	0.0500
SECOND	0.0840***	0.0856***	0.0810	0.1125	0.1012	0.0447
MALE	0.0960***	0.0952***	0.1664*	0.2297**	0.0378	0.1589*
INCOME	0.0000***	0.0000***	0.0000	0.0000	0.0000*	0.0000
LENGHT	0.0177***	0.0178***	0.0142***	0.0146***	0.0003	0.0138***
_cons	0.2811***	0.3034***	0.3433	-0.2222	-0.4997*	0.4817
alpha			0.8970***	1.5322***	0.9818***	1.9011***
CHILDREN						-0.2357***
AGE						-0.0099**
UNIVER						-0.5162**
BIKE						-0.2176*
N	692	692	692	692	692	692
L-likelihood	-7 507	-7 465	-2 525	-2 430	-2 159	-2 515
Pseudo R^2	0.418	0.421	0.083	0.067	-	0.086
$chi^2(18)$	10 789	10 873	458	348	1 270	476
AIC	15 054	14 970	5 092	4 903	4 361	5 087
BIC	15 145	15 060	5 187	4 998	4 461	5 219
Notes:	P – Poisson, TP – Truncated Poisson, NB – Negative binomial, TNB - Truncated NB, TSNB – Truncated and endogenously stratified NB, GNB – Generalized NB * = p < 0.1; ** = p < 0.05; *** = p < 0.01					

When comparing the negative binomial models (NB, TNB, GNB and TSNB), it can be seen that the log-likelihood increases under the model that account for endogenous stratification (TSNB). The correction for endogenous stratification results in a decrease in the absolute size of the price coefficient (TC) and, as a result, in an appropriate increase of the estimates of consumer surplus.

Surprisingly, model GNB doesn't dominates TSNB, by allowing the overdispersion parameter to vary according to variables that reflect visitor's characteristics: age, university degree, number of children in the household, and mountain bicycling as the main outdoor activity. The coefficients of all the covariates in the equation are highly significant, revealing that using the same overdispersion parameter for all observations would be overrestrictive. However according to AIC and BIC criteria, the econometric model based on negative binomial distribution that account both truncation and endogenous stratification provides the best fit of the data collected on-site.

5.3 Welfare estimates

We use the travel cost coefficients reported in Table 1 to calculate welfare measures in terms of the consumer surplus users derive from having access to the park. In all the count data models reported in Table 1 the consumer surplus per visit is calculated as $-1/\beta_{TC}$. If this expression is multiplied by $\hat{Y}$ (the predicted number of person-trips), we obtain the predicted CS per season for the typical visitor in the sample results.

As shown in Table 2, the average consumer surplus associated with an access to the JH Mts. ranges from CZK 11,622 for Truncated Poisson model to CZK 34,610 for negative binomial model. Expressed in US dollars the average consumer surplus per trip is about USD 40 for PM and USD 124 for TSNB[3].

Tab. 2: Welfare estimates per person-season and per person-trip (in CZK and USD)

Model	β_{tc}	$\hat{Y}$	CZK/trip	CZK/season	USD/trip	USD/season
P	-0.0008	12.25	1 256	15 390	50	616
NB	-0.0004	13.34	2 595	34 610	104	1 384
GNB	-0.0004	13.36	2 527	33 749	101	1 350
TP	-0.0010	11.64	999	11 622	40	465
TNB	-0.0005	10.62	2 019	21 427	81	857
TSNB	-0.0003	9.18	3 097	28 423	124	1 137

By decreasing the coefficient for TC, the correction for endogenous stratification yields a higher consumer surplus estimate per person-trip (under TSNB) than under TNB. In particular, TNB would yield USD 81 while TSNB would yield USD 124.

[3] The exchange rate at the time of survey was CZK 22 per US dollar.

6 Discussion and conclusions

We have used on-site survey data from Jizerské hory Mountains to estimate and compare a set of truncated count data models of recreation demand. The paper's main objective was on the performance of different estimators. Our results confirm, in line with earlier works based on recreational sites in the US and Europe that a model that corrects simultaneously for overdispersion, truncation, and endogenous stratification dominates more restrictive models in terms of goodness of fit.

The theoretical implications for the estimation of welfare estimates, i.e. consumer surplus, are as expected. The coefficient of the travels cost variable was significant in all specified count data models and negative according to the economic theory. Not correcting for overdispersion (by relying on Poisson estimates) substantially understates true consumer surplus, while accounting for endogenous stratification appropriately adjusts the consumer surplus upwards.

Acknowledgment

The research on this study was supported by the grant of the Czech Ministry of Agriculture: 1R56014 - Monetary valuation of recreational and aesthetical function of forest in the Czech Republic within the program "Krajina a Budoucnost". This support is gratefully acknowledged.

References

Alberini, A., and Longo, A. *The value of cultural heritage sites in Armenia and sustainable tourism: evidence from a travel cost study,* EAERE conference, 2005.

Creel, M. D., and Loomis, J. B. „Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California." *American Journal of Agricultural Economics*, 72, 1990: 434-441.

Englin, J., and Shonkwiler, J. „Estimating social welfare using count data models: an application to long-run recreation demand under conditions of endogenous stratification and truncation." *Review of Economics and Statistics*, 77, 1995: 104-112.

Haab, T., and McConnell, K. E. *Valuing Environmental and Natural Resources: the econometrics of non-market valuation*. Cheltenham UK: Edward Elgar, 2002.

Hellerstein, D., and Mendelsohn, R. „A teoretical foundation for count data models.“ *American Journal of Agricultural Economics*, 75, 1993: 604-611.

Kolstad, C. D. *Environmental Economics*. Oxford University Press. New York, 2000.

Martínez-Espineira, R., and Amoako-Tuffour, J. *Recreation Demand Analysis under Truncation, Overdispersion, and Endogenous Stratification: An Application to Gros Morne National Park, conference paper*, EAERE conference: Thessaloniky, 2007.

Ovaskainen, V., Mikola, J., and Pouta, E. „Estimating recreation demand with on-site data: An application of truncated and endogenously stratified count data models.“ *Journal of Forest Economics*, 7, 2001: 125-144.

Parsons, G. R. The Travel Cost Method. In: Champ, P. A., Boyle, K. J., Brown, T. C. (eds), *A Primer on Nonmarket Valuation.* London: Kluwer Academic Publishers, 2003.

Shaw, D. „On-site samples regression: problems of non-negative integers, truncation, and endogenous stratification.“ *Journal of Economics*, 37, 1988: 211-223.

Contact

Jan Melichar

Charles University Environment Center

José Martího 2, Praha, Czech Republic

jan.melichar@czp.cuni.cz