5-1-2008

Effect on Recreation Benefit Estimates from Correcting for On-Site Sampling Biases and Heterogeneous Trip Overdispersion in Count Data Recreation Demand Models (STATA)

Roberto Martínez-Espiñeira
Memorial University of Newfoundland, rmartinezesp@mun.ca

Joseph M. Hilbe
Arizona State University

Effect on Recreation Benefit Estimates from Correcting for On-Site Sampling Biases and Heterogeneous Trip Overdispersion in Count Data Recreation Demand Models (STATA)

Roberto Martínez-Espiñeira
Memorial University of Newfoundland

Joseph M. Hilbe
Arizona State University
NASA/Jet Propulsion Laboratory
California Institute of Technology

Correction procedures (STATA commands NBSTRAT and GNBSTRAT) are applied to simultaneously account for zero-truncation, endogenous stratification, and overdispersion, and also consider heterogeneity in the overdispersion parameter. Their effect is shown on welfare estimates from previous studies, confirming that the routines perform the appropriate correction and only when endogenous stratification is expected.

Key words: on-site sampling, heterogeneous overdispersion, zero-truncation, endogenous stratification, count data

Introduction

When analysing and predicting individual demand and behaviour in a variety of settings, researchers often resort to count data models to handle the special characteristics of the dependent variable and they often collect the data on-site for reasons of cost-effectiveness. This is the case, for example, of many recent recreational demand studies based on the travel cost method (TCM). The TCM is used to value public areas used for recreational purposes that require most users to travel to the site (Braden & Kolstad., 1991; Freeman, 1993). The TCM

assumes that the costs individuals incur when travelling to the site can act as surrogate prices for their recreational experience and that the number of trips to the site should decrease with increases in distance travelled and other factors that increase the total travel cost. Exploiting this empirical relationship between increased travel cost and declining visitation rates makes it possible to estimate a demand relationship. This demand schedule can be used to estimate the total benefits derived by visitors (e. g. consumer surplus). A similar approach can be applied in a variety of settings related to individual demand and behaviour analysis, but we will focus here on empirical applications of the individual single-site travel cost method. In a sense, the single-site TCM could be seen as a gravity model restricted to one destination from which no departures originate.

Gravity models are popular among geographers and transportation analysts and have also been used by recreation planners/economists to distribute regional recreation use across sites. However, they are somewhat less popular with economists. Economists typically work with the visitation-origin data to predict visitation and value at a given site. Multiple sites can be included in the models and visitation and value summed across sites to reflect an entire region. Gravity models work in the opposite direction: total visitation

Roberto Martínez-Espiñeira is an Associate Professor of Economics. His research focuses on the econometric analysis of travel cost and contingent valuation data used for environmental valuation. E-mail him at rmartinezesp@mun.ca. Joseph M. Hilbe is a Solar System Ambassador with NASA/Jet Propulsion Laboratory, an Adjunct Professor of Statistics at Arizona State University, and an Emeritus Professor at the University of Hawaii. He is an elected Fellow of the American Statistical Association and the International Statistical Institute.

for an entire region is first estimated, followed by use of the gravity concept where the total visitation is then allocated across sites based on relative attractiveness (Platt, 2000). The aggregate gravity model concept is perhaps more similar to the random utility allocation models under the individual TCM model, while in this paper we focus on single-site TCM studies.

When implementing the TCM in practice, data on visitors' behaviour are often collected on site, because, for sites frequented by only a small proportion of the general population, on-site sampling is much more cost-effective. However, it can lead to problems of endogenous stratification, because frequent users (or, in some cases, visitors who stay longer at the site) will be overrepresented in the sample (Shaw, 1988; Englin & Shonkwiler, 1995). Welfare measures based on the analysis of on-site samples will overstate the benefit derived from access to recreational site, unless the bias in the estimates is corrected. In on-site samples the dependent variable (visits to the site) is truncated at zero, because non-visitors are not observed, and often exhibits overdispersion (the variance of the visits variable exceeds the mean).

Shaw (1998) proposed a correction method for endogenous stratification, applied first to real data by Englin and Shonkwiler (1995). The correction proposed turns out to be very straightforward for equidispersed data which can be assumed to follow a Poisson distribution and has been frequently applied (e.g. Loomis 2003; Hagerty & Moeltner, 2005). However, appropriately correcting for endogenous stratification under overdispersion used to require custom programming by the software user. Only recenty, Hilbe and Martínez-Espiñeira (2005) packaged the *NBSTRAT* routine, applied to the analyses in this paper, to greatly facilitate this type of analysis using *STATA* (Statacorp, 2005). Achieving convergence is still much more difficult than in the Poisson case, where one simply needs to subtract 1 from the trip count and run a plain Poisson (see Shaw, 1988; or Haab & McConnell, 2002, 174-181, for details). Overdispersion is quite common, because typically the dependent variable takes a low value in many observations (for example many visitors make few trips to the site or stay only a few days) while it takes a high value for only a few observations (for example, only a few visitors make many trips or stay many days). This means that the variance of the dependent variable in the trip demand function is larger than the mean, making the Poisson model and its variants overly restrictive. Englin and Shonkwiler (1995), Ovaskainen, Mikkola and Pouta (2001), McKean, Johnson and Taylor (2005); and Martínez-Espiñeira and Amoako-Tuffour (2008) constitute examples of the few papers where both corrections were applied simultaneously.

By contrasting the effect on welfare measures of applying the correction procedure to different datasets we try in this paper to detect patterns and to identify causal factors behind substantial biases due to on-site sampling.

In the following sections we describe the comparison of estimates corrected for overdispersion, endogenous stratification, and truncation in several recreational demand datasets previously analyzed in the literature. These reanalyses show that zero-truncation accounts for most of the on-site sample bias, as is usually the case (Martínez-Espiñeira, Amoako-Tuffour & Hilbe, 2006) but the effect of correcting for endogenous stratification is nevertheless significant. The effect of endogenous stratification is, as the theory predicts, to exaggerate the value of access to a recreational site. However, the relative magnitude of the bias differs depending on the characteristics of the study. In some datasets the effect of accounting for heterogeneous overdispersion is also significant. Furthermore, we find that *NBSTRAT* and *GNBSTRAT* perform the downward correction of welfare estimates and improve goodness of fit only in those cases where we theoretically expect there to be endogenous stratification. Therefore, they can be used not only as a correction mechanism, but also as a diagnosis tool for this bias.

Background

Many travel cost method studies are based on on-site sampling (Englin & Shonkwiler, 1995; Ovaskainen, Mikkola & Pouta, 2001; Bhat, 2003; Shaw, Fadali & Lupi.,

2003; Loomis, 2003, McKean et al., 2003; McKean, Johnson, Taylor & Johnson, 2005). Many recreational activities often attract only a small proportion of the population and users are rarely listed anywhere, so drawing a random sample is very costly. However, as described in further detail by Martínez-Espiñeira et al. (2006), this sampling strategy, which can be seen as a particular type of choice-based sampling, can lead to endogenous stratification. Uncorrected estimates will erroneously overstate the benefit derived from a certain site.

Shaw (1988) considered a correction for endogenous stratification count data estimators in the context of a single recreational site, and derived a correction procedure for the Poisson. Englin and Shonkwiler (1995) extended this correction procedure to the negative binomial model and applied it to real data.

Loomis (2003) compared benefit estimates calculated from a household survey data and data collected on-site, in order to measure the effect of correcting the on-site sample estimates for endogenous stratification. He showed that accounting for the truncated and endogenously stratified nature of the data collected on-site substantially reduced consumer surplus estimates, as theoretically expected, and brought them very close to those unbiased estimates obtained from the household survey. Martínez-Espiñeira et al. (2006) showed in their reanalysis that most of the bias in the uncorrected estimates obtained from Loomis (2003)'s on-site sample was due to the truncation, not the endogenous stratification. Both studies assumed equidispersion in the dependent variable as required by the Poisson. However, Martínez-Espiñeira, Loomis, Amaoko-Tuffour, and Hilbe (2008) reanalyzed the dataset accounting also for overdispersion (with an adjusted negative binomial model) and confirmed the main insights offered by previous comparative analyses.

Apart from those cited above, other studies, such as Ovaskainen et al. (2001), McKean et al. (2003), and McKean et al. (2005) analyzed on-site samples accounting for both overdispersion and endogenous stratification. However, with the exception of Englin and Shonkwiler (1995) and Martínez-Espiñeira and Amaoko-Tuffour (2008), previous analyses

apply a negative binomial model that assumes a constant overdispersion parameter for all visitors, while McKean, Johnson, and Taylor (2003) parameterise the overdispersion parameter on an artificially generated variable only. The assumption that the overdispersion parameter is constant across observations is often violated. In the case of conventional count data samples, this prompts the use of a generalized, or heterogeneous (see Hilbe, 2007), negative binomial model that account for this extra source of heterogeneity. This strategy provides information about which predictors contribute to overdispersion, which can be useful when evaluating the model and attempting to determine the worth of each predictor to the model.

For truncated and endogenously stratified data samples, *STATA* 9.1 (Statacorp, 2005) routine *GNBSTRAT* (Hilbe, 2005) can be applied to allow the parameterisation of the overdispersion parameter as a function of visitor characteristics. *NBSTRAT* (Hilbe & Martínez-Espiñeira, 2005) simply optimizes the value of a common overdispersion parameter.

Data

In this article, some analyses are replicated based on household and (mainly) on-site samples available in the literature and extend them to include corrections for overdispersion and/or heterogenous overdispersion. The reader is referred to the individual source for details about the individual data sets and the results of the analyses conducted in each paper. Here, we will focus on the nature of the dependent variable and the fashion in which the data were collected. We introduce the different datasets in chronological order, based on publication dates.

McConnell et al. (1986)

This dataset, also in Haab and McConnell (2002, pages 156-171) dealt with recreational trips to Fort Phoenix Beach (New Bedford, Massachusetts). There were originally 499 observations in this dataset on five variables, including the round-trip travel costs plus monetary value of time to Fort Phoenix Beach. The data were collected through a household survey, so they contain many zero

values for the dependent variable, but we only use the 168 observations for which the number of trips equals at least one, in order to artificially truncate the sample.

Gurmu and Trivedi (1996)

This is a count data set originally used by Gurmu and Trivedi (1996) with a focus on the modelling of excess zeros, for recreational boating trips collected through a household survey. Discussion of the variables can be found in Sellar, Stoll and Chavas (1985) and Ozuna and Gomaz (1995). These data are also used in Chapter 6.4-6.5 and Chapter 12.6 of Cameron and Trivedi (1998). The dataset includes 659 observations on the number of boating trips to Lake Somerville, East Texas, in 1980 and a series of variables that includes the travel cost to the Lake Somerville, income, and travel costs to substitute lakes. These data were collected through a household survey, so they contain many zero values for the dependent variable. We artificially truncated the sample by restricting our analysis to those 242 observations for which the dependent variable is at least equal to one.

Sohngen, Lichtkoppler and Bielen (2000)

These data were collected to study the value of day trips to Lake Erie beaches. Subsamples of this dataset have also been used by Alberini and Reppas (2005) and Parsons (2003). We use the 223-observation subset (on visits to Maumee Bay State Park beach in Ohio) used by Haab and McConnell (2002 pp. 179-180). This simplified dataset contains only four variables, including number of trips and round-trip travel costs plus monetary value of travel time to that site, round-trip travel costs plus monetary value of travel to nearest substitute beach, and household income.

Ovaskainen et al. (2001)

This dataset contains 656 observations from an on-site survey of visitors conducted on several adjacent recreation sites in the Nuuksio Lake Plain, Finland. The dependent variable is the count of trips taken to the site during the last 12 months. A potential anomaly, however, results from the way in which the relevant question was asked ("How many times did you visit this site during the last year?"). Because it

did not explicitly specify whether the current trip should be included or not, there are a non-negligible amount of zeros. This suggests that respondents excluded the current trip, so one trip was added by Ovaskainen et al. (2001) to all observations below 20 trips.

Shrestha, Seidl and Moraes (2002)

Data on recreational fishing were collected from the Brazilian Pantanal over the 3-month period from August through November, 1994. Visitors were queried about their travel costs of the visit, reasons for choosing the Pantanal as a travel destination, aspects of their experiences, and some demographic information. The number of trips taken within the previous year is the dependent variable and it is regressed on several explanatory variables that include the round trip travel cost of the respondent for the current fishing trip.

Loomis (2003)

These data, also used by Martínez-Espiñeira et al (2006) and Martínez-Espiñeira et al. (2006), consist of two sets: one collected through an on-site intercept survey of visitors to the Snake River in Jackson Hole, Wyoming, and a second one collected through a household survey about visits to that same site. Details about the data and the collection process are available in Loomis (2003).

McKean et al. (2003)

McKean et al. (2003) conducted an on-site survey of flat water recreationists on reservoirs at the impounded lower Snake River. The variables used include information on available free time and income, monetary and time costs of travel, outdoor recreation, and other activities on and off the recreation area. The dependent variable is annual trips to the site. McKean et al. (2003) apply a truncated negative binomial regression with an adjustment for endogenous stratification that allows the overdispersion parameter to vary across observations as a function of a randomly generated value. In the appendix they transcribed the code for LIMDEP 7 (Greene, 1995) used to obtain the truncated negative binomial model adjusted for endogenous

stratification and describe their difficulty to achieve convergence with this approach.

Martínez-Espiñeira and Amoako-Tuffour (2008)

This is a subset (N=413) of a larger dataset collected on-site at Gros Morne National Park in Newfoundland (Canada). The product of the number of trips to the park in the previous five years times the number of people travelling together in the current trip is regressed against several explanatory variables, including the combined travel cost (money costs and the value of travel time) spent reaching the park and days spent on-site during the current trip. The data were not collected randomly. Visitors were oversampled from rare origins, so the analysis uses sampling weights to correct for this. However, no correction was possible for oversampling of visitors who stayed longer at the park or who visited more locations within the park (so they would have a higher likelihood of being interviewed).

Mendes and Proença (2005)

This is an on-site survey at the Peneda-Gerês National Park (Portugal). The dependent variable is not the number of visits, but rather the number of days on-site during the current visit. In this case, a concern would be the problem of oversampled visitors who stayed longer at the park, since interviewers intercepting visitors within the park would be more likely to find visitors whose visit was longer (a problem described in detail by Lucas, 1963). Crucially, the authors note that, in order to avoid this type of endogenous stratification, visitors were instead interrogated only at the time they addressed themselves to the camping reception centre for camping inscription. For this reason, their reported results do not include a model that corrects for endogenous stratification. The price variable is the on-site and travel out-of-pocket costs, as well as travel and on-site time opportunity costs, and not only travel costs.

Methodology

Count data models are now routinely applied in single-site recreation demand models (Creel & Loomis, 1990; Englin & Shonkwiler, 1995;

Gurmu & Trivedi, 1996; Shrestha et al., 2002). These models account for the fact that the dependent takes only nonnegative integer values. These distributions exhibit a concentration of values on a few, small discrete values (e.g., $0 - 2$), skewness to the left, and intrinsic heteroskedasticity with variance increasing with the mean (Cameron & Trivedi, 1998 and 2001). Hellerstein and Mendelsohn (1993) theoretically justified the use of count data to model recreational demand: on any choice occasion, the decision to take a trip is modelled with a binomial distribution. As the number of choices increases the binomial asymptotically converges to a Poisson distribution. The first two moments of the Poisson distribution equal each other, a property known as equidispersion. The model can be extended to a regression framework by parameterizing the relation between the mean parameter and a set of regressors using an exponential mean parametrization.

Overdispersion

However, data on the number of trips are often overdispersed, making the Poisson model overly restrictive. The Poisson maximum likelihood estimator with overdispersion is still consistent, but it underestimates the standard errors and inflates the t-statistics in the usual maximum-likelihood output. If the overdispersion problem is severe, the negative binomial model should be applied. This is commonly obtained by adding an additional parameter that reflects the unobserved heterogeneity that the Poisson fails to capture. This parameter (usually denoted α) determines the degree of dispersion in the predictions (see e. g. Cameron and Trivedi, 1990; Cameron and Trivedi, 2001, p. 336).

Truncation

In on-site samples, the distribution of the dependent variable is also truncated at zero. Ignoring this leads to biased and inconsistent estimates, because the conditional mean is misspecified (Shaw, 1988; Creel & Loomis, 1990; Grogger & Carson, 1991; Yen & Adamowicz, 1993; Englin & Shonkwiler, 1995). In that case, the truncated negative binomial is in order. Examples of applications of this model

include Bowker, English and Donovan (1996); Liston-Heyes and Heyes (1999); and Shrestha et al. (2002). Yen and Adamowicz (1993) compare welfare measures obtained from truncated and untruncated regressions.

Endogenous stratification

Finally, on-site data are affected by endogenous stratification, because a visitors' likelihood of being sampled is positively related to the number of trips they made to the site (or the number of days they spent at the site). If the assumption of equidispersion holds, standard regression packages can be used to estimate a Poisson model adjusted for both truncation and endogenous stratification, as shown by Shaw (1988), by simply running a plain Poisson regression on the dependent variable modified by subtracting 1 from each of its values (Haab & McConnell, 2002, p. 174-181). This model has been used in several applied studies under the assumption of no significant overdispersion (Fix & Loomis, 1997; Hesseln et al., 2003; Loomis, 2003; Hagerty & Moeltner, 2005; Martínez Espiñeira et al., 2006).

For the case where overdispersion is significant, the density of the negative binomial distribution truncated at zero and adjusted for endogenous stratification, derived by Englin and Shonkwiler (1995), cannot be rearranged into an easily estimable form, so it used to require custom programming as a maximum likelihood routine, with the associated increase in computational burden. Englin and Shonkwiler (1995) provide an empirical application of this specification. Englin, Holmes and Sills (2003) and Ovaskainen et al. (2001) also used this model and found that correcting for endogenous stratification on top of zero-truncation does not make much difference in estimates.

However, these studies are based restrict the overdispersion parameter to a common value for all observations (so $\alpha_i = \alpha$). To our knowledge, only Englin and Shonkwiler (1995) have attempted to parameterize α (as $\alpha_i = \alpha 0/\lambda_i$). Ovaskainen, Mikkola, and Pouta (2001) also tried this specification but their keeping α constant at a value previously estimated using a nonlinear squares regression yielded better results in their study. McKean, Johnson, and Taylor (2003) allowed α to vary as a function of

a randomly generated parameter, not related to visitor characteristics. One of the main methodological contributions of the present paper is to use the more flexible approach that allows the overdispersion parameter to vary according to visitor characteristics and compare it with the more restrictive approach. The code is now available for *STATA* 9.1 (Statacorp, 2005) as downloadable commands *NBSTRAT* (Hilbe & Martínez-Espiñeira 2005) and *GNBSTRAT* (Hilbe, 2005). GNBSTRAT makes it possible to evaluate how visitors characteristics influence the individual degree of overdispersion and permit to more fully evaluate the effect of these characteristics on the number of trips in the main part of the trip prediction model.

Results

Replicated analyses and the reanalyses of the datasets described in Section Data are considered. In order to check consistency, for all the datasets replicated exactly the analyses conducted in the original works first. Then we ran a negative binomial (NBREG), zero-truncated negative binomial (ZTNB), a zero-truncated negative binomial adjusted for endogenous stratification (NBSTRAT), and a zero-truncated negative binomial adjusted for endogenous stratification and heterogenous overdispersion (GNBSTRAT). These four types of regression are reported in Table 1, summarising the characteristics of the datasets and the results concerning the travel cost coefficient. To maintain consistency, the same model specifications proposed by the original authors to run NBSTRAT and GNBSTRAT is used. For ease of comparison with the original works, the same number of significant decimal places is used to report results.

The focus is on the usefulness of using NBSTRAT and GNBSTRAT and their effects on welfare estimates obtained through count data models. We, therefore assume that the data collection processes and the specifications proposed by the original authors to model the number of trips are a sufficiently valid approximation to the requirements of the individual TCM. In this sense, we abstract, among others, from any potential problems related with additional sources of non-

randomness in the sample (although the idea of oversampling of visitors according to length of stay below is considered) or the fact that some of the datasets might be affected by problems of multi-purpose or multi-site visitation. It is likely that one or more of these internal problems other than those related to the issue of endogenous stratification affect one or more of the studies described below. Those issues are beyond the scope of this work, but the interested reader is directed to Parsons (2003) or Phaneuf and Smith (2006).

McConnell et al (1986)

Using McConnell et al. (1986)'s household sample of beach recreationists, we replicated the Poisson and Negative Binomial specifications reported by Haab and McConnell (2002), not reported, but available upon request, and then applied a zero-truncated model to the positive trip observations of the data set (ZTNB in Table 1). This is compared with the NBSTRAT specification, which not only takes into account truncation and overdispersion, but also endogenous stratification, which should not be expected to affect this dataset.

As expected, NBSTRAT correctly suggests that there is no problem with endogenous in this case, because the data were not collected on site. NBSTRAT yields a worse goodness of fit (log-likelihood) than ZTNB and also a smaller (in absolute value) estimate for the price coefficient, so the consumer surplus per trip, as shown in Table 2, would be higher ($5.32 while under ZTNB it would be $5.13). The standard negative binomial regression (NBREG) is also reported, which reveals that correcting for zero-truncation, even in the artificially truncated sample, would account for most of the correction over an inflated estimate of consumer surplus.

Gurmu and Trivedi (1996)

This is a count data set on recreational boating trips to Lake Somerville collected through a household survey. Gurmu and Trivedi (1996) focus on modelling excess zeros. As pointed out by Phaneuf and Smith (2006, p.57) the Poisson and negative binomial distributions typically do not place enough probability mass at zero to match the excess zeros found in many

recreation datasets. Hurdle models consider different data generating processes for explaining the likelihood of individuals being users and for the number of trips for those who are users. There are several types of hurdle and zero-inflated models (see Mullahy, 1986; Lambert, 1992; Cameron & Trivedi, 1998, pp. 123-125 for theoretical details, pp. 889-891; and Martínez-Espiñeira, 2007, for a recent application) and Gurmu and Trivedi (1996) report, among others, the results of a zero-truncated negative binomial as part of their hurdle model. They label this regression Negbin hurdle on Positives.

By restricting the current analysis to the positive values of the dependent variable, we managed to replicate this regression as ZTNB, reported in Table 1, together with the results for NBSTRAT on the positive values of the dependent variable. As expected, this model does not work well on this sample. There was substantial difficulty getting NBSTRAT to converge. Additionally, the log-likelihood worsens relative to ZTNB and the absolute value of the own travel cost coefficient is smaller under NBSTRAT, leading to a higher estimate of consumer surplus per trip (while a correction for endogenous stratification would adjust the consumer surplus downwards).

The command NBSTRAT performs in a satisfactory manner in this example, since even if the researcher had wrongly expected endogenous stratification to affect this household sample, NBSTRAT would have revealed ZTNB preferable to NBSTRAT. Of course the original sample also contains zeros, so the best models overall are either a negative binomial (with no truncation) or, as shown by Gurmu and Trivedi (1996), models that account for excess zeros. We tried to run GNBSTRAT but no choice of independent variables helped explain any additional variation of α across visitors, stressing the notion that, as expected, endogenous stratification is not a problem, so modelling the overdispersion more flexibly while accounting for the nonexistent endogenous stratification was not helpful either.

Table 1. Results

Dataset		NBREG	ZTNB	NBSTRAT	GNBSTRAT
McConnell et al. (1986)	β_{TC}	-0.1666**	-0.1950*	-0.1880**	-0.2123**
N = 168 (trips>0 only)	LL	-578.8	-563.3	-564.8	-562.6
Household survey	AIC	1170	1139	1140	1141
Gurmu & Trivedi (1996)	β_{TC}	-0.054***	-0.078***	-0.072***	
N = 242 (trips>0 only)	LL	-644.9	-591.6	-594.3	
Household survey	AIC	1308	1201	1207	
Sohngen et al. (2000)	β_{TC}	-0.013***	-0.017***	-0.017***	-0.029***
N= 223	LL	-588.2	-562.2	-562.3	-549.5
On-site	AIC	1186	1134	1135	1111
Ovaskainen et al. (2001)	β_{TC}	-0.0117***	-0.01484***	-0.01397***	-0.01385***
N= 656	LL	-1928	-1822	-1835	-1834
On-site	AIC	3872	3659	3686	3689
Ovaskainen et al. (2001)	β_{TC}	-0.0098***	-0.01122***	-0.01137***	-0.01095***
N= 542 (trips>1 only)	LL	-1663.8	-1623	-1618	-1611
On-site	AIC	3344	3261	3253	3244
Shrestha et al. (2002)	β_{TC}	-0.0008**	-0.0019***	-0.0021***	-0.0018**
N = 286	LL	-354.5	-175.2	-175.1	-172.4
On-site	AIC	733.1	376.4	376.2	372.8
Loomis (2003)	β_{TC}	-0.02097***	-0.03874***	-0.04076***	-0.02598***
N = 172	LL	-674.5	-624.1	-626.4	-563.3
On-site	AIC	1365	1264	1269	1147
Loomis (2003)	β_{TC}	-0.04617***	-0.06987***	-0.06663***	
N=217	LL	-819.2	-774	-787.8	
Household survey	AIC	1654	1564	1592	
McKean et al. (2003)	β_{TC}	-3.342***	-3.368***	-3.405***	-2.276***
N= 388	LL	-1092.6	-994.4	-995.2	-916.4
On-site	AIC	2213	2017	2018	1865
Martínez-Espiñeira & Amoako-Tuffour (2008)	β_{TC}	-0.3855***	-0.5272***	-0.5701***	-0.4665***
N= 413 (persontrip)	LL	-1020.7	-969.0	-957.6	-940.6
On-site	AIC	2063	1960	1937	1907
Martínez-Espiñeira & Amoako-Tuffour (2008)	$\beta_{Cost/day}$	-0.5709***	-0.7762***	-0.9026***	-0.9051***
N= 413 (days spent on site)	LL	-947.3	-922.6	-908.8	-905.7
On-site	AIC	1915	1865	1838	1833
Mendes & Proença (2005)	$\beta_{Cost/day}$	-0.00526***	-0.00599***	-0.00666***	-0.00614***
N= 243 (days spent on site)	LL	-598.7	-589.5	-590.2	-582.3
On-site	AIC	1211	1193	1194	1185
*p<0.1; **p<0.05; *** p<.001; LL = log-likelihood; AIC = Akaike Information Criterion					

Table 2. Consumer surplus estimates.

Dataset		NBREG	ZTNB	NBSTRAT	GNBSTRAT
McConnell et al. (1986)	β_{TC}	-0.1666**	-0.1950*	-0.1880**	-0.2123**
Household survey (trips>0)	CS/trip	$6.00	$5.13	$5.32	$4.71
Gurmu & Trivedi (1996)	β_{TC}	-0.054***	-0.078***	-0.072***	
Household survey (trips>0)	CS/trip	$18.51	$12.90	$13.88	
Sohngen et al. (2000)	β_{TC}	-0.013***	-0.017***	-0.017***	-0.029***
On-site	CS/trip	$79.51	$59.03	$57.80	$34.12
Ovaskainen et al. (2001)	β_{TC}	-0.0117***	-0.01484***	-0.01397***	-0.01385***
On-site	CS/trip	$85.59	$67.38	$71.58	$72.20
Ovaskainen et al. (2001)	β_{TC}	-0.0098***	-0.01122***	-0.01137***	-0.01095***
On-site (trips>1 only)	CS/trip	$101.75	$89.12	$87.94	$91.32
Shrestha et al. (2002)	β_{TC}	-0.0008**	-0.0019***	-0.0021***	-0.0018**
On-site	CS/trip	$1250.00	$526.32	$476.19	$555.56
Loomis (2003)	β_{TC}	-0.02097***	-0.03874***	-0.04076***	-0.02598***
On-site	CS/trip	47.68	25.81	24.53	38.49
Loomis (2003)	β_{TC}	-0.04617***	-0.06987***	-0.06663***	
Household survey	CS/trip	$21.66	$14.31	$15.01	
McKean et al. (2003)	β_{TC}	-3.342***	-3.368***	-3.405***	-2.276***
On-site	CS/trip	$29.93	$29.69	$29.37	$43.94
Martínez-Espiñeira & Amoako-Tuffour (2008)	β_{TC}	-0.3855***	-0.5272***	-0.5701***	-0.4665***
On-site	CS/trip	$2,593	$1,897	$1,754	$2,143
Martínez-Espiñeira & Amoako-Tuffour (2008)	$\beta_{Cost/day}$	-0.5709***	-0.7762***	-0.9026***	-0.9051***
On-site	CS/day	$1,752	$1,288	$1,108	$1,105
Mendes & Proença (2005)	$\beta_{Cost/day}$	-0.00526	-0.00599***	-0.00666***	-0.00614***
On-site	CS/day	$190.11	$166.94	$150.15	$162.87
*p<0.1; ** p<0.05; *** p<.001					

Sohngen, et al. (2000)

A subset (N=223) of the original sample was used to successfully replicate the regressions reported by Haab and McConnell (2002, p. 180), who ran a Truncated Poisson model and a Truncated Poisson corrected for endogenous stratification. We report in Table 1 our ZTNB, NBSTRAT and GNBSTRAT results. As expected, endogenous stratification affects the dependent variable in this sample collected on-site. NBSTRAT, although the level of accuracy (3 decimal places) used for the coefficients by Haab and McConnell would not make it apparent, corrects downwards the estimated consumer surplus. Finally, GNBSTRAT was used to model the overdispersion parameter as a function of the travel cost to the site, finding that it significantly improves the goodness of fit. In this case, accounting for the heterogeneous nature of the overdispersion across visitors increases the value of the estimated consumer surplus to $34.48.

Ovaskainen et al. (2001)

Ovaskainen et al. (2001) reported the results of running in LIMDEP (Greene, 1995) a series of count data models that include zero-truncated models and also models that correct for endogenous stratification. The replication with NBSTRAT in STATA yields slightly different results than what the original authors report as their zero-truncated endogenously stratified negative binomial. This is likely due to the fact that they had to fix the value of alpha to a constant estimated from a separate regression based on nonlinear least squares. It is noteworthy that NBSTRAT achieved a much higher log-likelihood (-1835) than the original procedure used by the original authors (-1891). Additionally, note that the zero-truncated endogenously stratified negative binomial yields a price coefficient (-0.01397) that is smaller in absolute value than the one obtained without correcting for endogenous stratification. This is in line with the results obtained in Ovaskainen et al. (2001). Because the data were collected on-site, we would expect a bias from endogenous stratification in the opposite direction.

It is possible that this puzzling result has to do with the anomaly in the data described in Section Data. The original authors added one trip to each observation with trips less than 20, being unsure of whether respondents had included the current trip in their response or not. This possibility seems more likely when we analyze only those 541 observations for which the `manipulated' number of trips is more than one. If that is done NBSTRAT performs as expected. Although not reported here, further regressions on smaller samples (for observations with only more than 2 trips, more than 3 trips, etc) confirmed in an increasingly reassuring way that the endogenous stratification correction performed by NBSTRAT would have worked in the expected direction if the data collection had not suffered from this unfortunate wording of the question about the number of trips.

It can also be shown that, for the trimmed samples, GNBSTRAT also slightly overperforms the previous models by making the overdispersion parameter a function of the age, equipment ownership, and income of the visitors. GNBSTRAT results are reported, although they do not offer much improvement over NBSTRAT.

Shrestha, et al. (2002)

A model equivalent to ZTNB was reported in the original paper. We failed to replicate its results exactly, but they are similar. The original authors claimed that no significant bias due to endogenous stratification was expected, "mainly because of the one-time survey of the anglers, rather than using annual visitor-data in the analysis." NBSTRAT shows (Table 1) that the correction would clearly reduce the estimates of consumer surplus per trip (from $526.32 to $476.1). However, the improvement in terms of log-likelihood is not substantial. GNBSTRAT improves the fit somewhat by making the overdispersion parameter a function of income.

Loomis (2003)

Income was rescaled into $10,000 units, but otherwise the same 172 observations and variables were used when applying different count data specifications to the on-site sample used in Loomis (2003). NBSTRAT performs the appropriate type of correction on ZTNB. However, NBSTRAT does not improve the fit much. A GNBSTRAT specification that makes

the overdispersion parameter α a function of the number of trips and income does improve the fit substantially.

When it comes to the reanalysis of the household sample collected by Loomis (2003), it can be seen in Table 1 that NBSTRAT would, as expected, show no improvement over ZTNB on an artificially truncated sample. The log-likelihood worsens and the estimated consumer surplus per trip increases, while a correction for endogenous stratification on a sample collected on-site would of course lead to a measure of consumer surplus revised downwards.

McKean et al (2003)

Using the code provided by McKean et al. (2003) in an appendix, the results were replicated using the maximizing commands in LIMDEP (Greene, 1995). However, when GNBSTRAT in STATA was used to try to replicate them, we failed to obtain the same results. This is probably because of two reasons. First, McKean et al. (2003) parameterise their overdispersion parameter as a function of a randomly generated variable (zz in their own notation) which takes a different value in each estimation. Second, it is likely that STATA maximum likelihood routine can obtain a more finely improved log-likelihood.

In any case, the results are similar in regards to the correction for endogenous stratification. In Table 1 we show the results of several specifications using the same data set used in the original. However, note that the values were rescaled (dividing by 100) for some of the variables to improve the presentation. For example the estimate found by McKean et al. (2003) for the travel cost under the equivalent of NBSTRAT was equal to -0.0337, while we obtained -3.405. The goodness of fit improves as we allow for a more flexible specification that accounts for on-site sample biases. NBSTRAT performs the expected type of correction on the estimates of consumer surplus per trip. In this case the magnitude of the bias caused by on-site sampling is not substantial in terms of consumer surplus per trip.

One off-pattern feature of the analysis of this dataset is that the correction for zero-truncation in itself does not seem to account in this case for much of the correction of the bias

due on-site sampling. As suspected, this appears to be related to the high value of the average number of trips (8.448). This value is larger than in the other studies analyzed, but closest to the equivalent value in Ovaskainen et al. (2001). The results from the latter show that, most of all when only trip values above one are used, the zero-truncation correction is, by itself, not substantial in relative terms. This is in line with the intuition that this type of correction is more necessary when the average value of the count (trips in the illustrations used here) is very low, as is typical in count data analyses.

Martínez-Espiñeira and Amoako-Tuffour (2008)

Martínez-Espiñeira and Amoako-Tuffour (2008) considered the dependent variable persontrip (number of trips times size of visitor party) as a function of travel costs and other characteristics of the trip and the visitors. Here we use a subsample of their data set to illustrate the effect of correcting for endogenous stratification. As expected, NBSTRAT performs a downward correction on the estimates of consumer surplus per trip. Allowing the overdisperion parameter to vary according to variables related to income and the age composition of the visitor party would improve the goodness of fit.

With this dataset it can also be considered how the issue of on-site sampling can affect welfare estimates when the travel cost model is based on the length of stay as the dependent variable (Lucas, 1963; Mendes & Proença 2005). In this case visitors were intercepted at several locations within the park, with no specific strategy for avoiding oversampling those visitors who stayed longer at the park. Therefore, those visitors who spend more days at the park had a higher likelihood of being intercepted than those who spent fewer days. Correcting for the resulting endogenous stratification would reduce the estimates of consumer surplus in a model that relates the length of stay to the cost of reaching the park. For this reanalysis, combined travel and on-site cost per day was constructed analogous to the one used by Mendes and Proença (2005). The associated coefficient are labeled $\beta_{Cost/day}$ in Tables 1 and 2. Those visitors who face a higher combined travel and stay cost are expected to

spend less time at the site. They likely use part of their available recreational time to visit other sites adjacent to the site of interest or on the way to it from their home. Correcting for endogenous stratification in this case also works as expected, decreasing the estimate of consumer surplus per day spent at the park.

It is noteworthy that in this case, the number of trips made to the park has no significant effect on the length of stay during the current trip. Only when GNBSTRAT makes the overdispersion parameter a function of that variable does the number of trips become significant and does it take the expected negative sign. It was expected to find that those who live closer to the park make more frequent but shorter visits to the park, once every other influence on the length of stay (particularly the travel cost) has been controlled for.

Mendes and Proença (2005)

Contrary to the case of Martínez-Espiñeira and Amoako-Tuffour (2008), Mendes and Proença (2005) did adopt a specific strategy to avoid oversampling those visitors who stayed longer at the park. They only interviewed them when signing in at the camping reception centre. This strategy is expected to successfully avoid the problem of endogenous stratification, so NBSTRAT was used to test if this was indeed true.

Table 1 shows that, although the estimated coefficient of the price variable (minimum recreation cost of each day of stay at the site, including travel cost) is slightly larger in absolute terms under NBSTRAT than under ZTNB, there is no improvement in goodness of fit due to correcting for endogenous stratification under the negative binomial models. Once again, this confirms that NBSTRAT can be relied upon to diagnose problems of endogenous stratification, since it does not spuriously improve the goodness of fit, relative to the uncorrected ZTNB for samples that are not affected by the problem.

Conclusion

The reanalyses above show that the newly developed commands NBSTRAT and GNBSTRAT perform appropriately when correcting for the simultaneous problems of zero-truncation, overdispersion, and endogenous stratification. These commands illustrate the effect of endogenous stratification on the estimates obtained from samples of recreation data obtained on site and allow the researcher to easily correct the resulting bias. Furthermore, by applying them to datasets obtained by artificially truncating at zero a sample collected from the general population, we show that the commands will not reduce the estimates of consumer surplus and will not improve the goodness of fit of the regression when they are applied to datasets that are not actually affected by endogenous stratification. That is, it is not a sledgehammer solution: it only works well when the problem is actually there. In this sense, we can safely suggest the use of NBSTRAT and GNBSTRAT as both a diagnostic tool, useful when the researcher does not know how serious the problem of oversampling of avid users is, and as a correction tool for the bias. NBSTRAT helped us confirm that, in some cases, on-site sampling is not subject to endogenous stratification if the sampling strategy is carefully designed to avoid it.

We have confirmed for several datasets that most of the overall bias caused by sampling on site is due to the truncation at zero of the dependent variable. This is a result that appears to apply regardless of the idiosyncrasies of each particular example, although it is less apparent in those datasets with a high average value of the dependent variable. However, the problem of endogenous stratification contributes to inflate uncorrected welfare estimates.

We expect that these two newly developed commands will help applied researchers with average computing abilities to properly analyze recreational datasets obtained through on-site surveys. By applying them, while enjoying the advantages of on-site sampling, researchers no longer need to worry about endogenous stratification or the computational burden associated to alternative ways to handle it.

Note that in the analysis we have assumed that the only problems affecting the welfare estimates had to do with on-site sampling. In particular, we assumed that the assumptions needed for a meaningful travel cost method analysis were met and that the correct

set of variables was included in the model specification in each case. Further research efforts should be directed at addressing these issues and analysing the influence of different types of misspecification and measurement problems on the magnitude of biases due to on-site sampling.

Finally, we should note that although the corrections showcased in this paper focused on the effects on consumer surplus measures in the context of the travel cost method, the analysis extends to any other type of count data analysis where obtaining unbiased estimates of the relevant coefficients was an issue.

References

Alberini, A. & D. Reppas (2005). Model misspecification and endogenous on-site sampling in the travel cost method. presented at EAERE 2005 Bremen.

Bhat, M. G. (2003). Application of non-market valuation to the Florida Keys Marine Reserve Management. *Journal of Environmental Management* 67 (4), 315-325.

Bowker, J. M., D. B. K. English & J. A. Donovan (1996). Toward a value for guided rafting on southern rivers. *Journal of Agricultural and Applied Economics* 28 (2), 423-432.

Braden, J. B. & C. Kolstad (1991). *Measuring the Demand for Environmental Quality*. Amsterdam: Elsevier.

Cameron, A. C. & P. K. Trivedi (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics* 46 (3), 347-364.

Cameron, A. C. & P. K. Trivedi (2001). Essentials of count data regression. In B. H. Baltagi (Ed.), A *Companion to Theoretical Econometrics*, pp. 331.348. Oxford, U.K.: Blackwell.

Cameron, C. & P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Creel, M. & 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.

Englin, J. & J. Shonkwiler (1995). Estimating social welfare using count data models: An application under conditions of endogenous stratification and truncation. *Review of Economics and Statistics* 77, 104-112.

Englin, J. E., T. P. Holmes & E. O. Sills (2003). Estimating forest recreation demand using count data models. In E. O. Sills (Ed.), *Forests in a Market Economy*, Chapter 19, pp. 341.359. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fix, P. & J. Loomis (1997). The economic benefits of mountain biking at one of its meccas: An application of the travel cost method to mountain biking in Moab, Utah. *Journal of Leisure Research* 29 (3), 342-352.

Freeman III, A. M. (1993). *The Measurement of Environmental and Resource Values: Theory and Methods*. Washington D.C.: Resources for the Future.

Greene, W. H. (1995). *LIMDEP Version 7.0 User.s Manual*. Bellport, NY: Econometric Software, Inc.

Grogger, J. T. & R. T. Carson (1991). Models for truncated counts. J*ournal of Applied Econometrics* 6 (3), 225-238.

Gurmu, S. & P. Trivedi (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics* 14, 469-477.

Haab, T. & K. McConnell (2002). *Valuing Environmental and Natural Resources: Econometrics of Non-Market Valuation*. Cheltenham, UK: Edward Elgar.

Hagerty, D. & K. Moeltner (2005). Specification of driving costs in models of recreation demand. *Land Economics* 81 (1), 127-143.

Hellerstein, D. & R. Mendelsohn (1993). A theoretical foundation for count data models. *American Journal of Agricultural Economics* 75 (3), 604-611.

Hesseln, H., J. B. Loomis, A. González-Cabán & S. Alexander (2003). Wildfire effects on hiking and biking demand in New Mexico: A travel cost study. *Journal of Environmental Management*, 69 (4), 359-368.

Hilbe, J. M. (2005). GNBSTRAT: Stata module to estimate generalized negative binomial with endogenous stratification. Statistical Software Components, Boston College Dept. of Economics: http://ideas.repec.org/c/boc/bocode/s456413.html.

Hilbe, J. M. & R. Martínez-Espiñeira (2005). NBSTRAT: Stata module to estimate negative binomial with endogenous stratification. Statistical Software Components, Boston College Department of Economics. http://econpapers.repec.org/software/bocbocode/s456414.htm.

Hilbe, J.M (2007). *Negative Binomial Regression*, Cambridge: Cambridge University Press.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1-14.

Liston-Heyes, C. & A. Heyes (1999). Recreational benefits from the Dartmoor National Park. *Journal of Environmental Management* 55 (2), 69-80.

Loomis, J. (2003). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure. *Water Resources Research* 39 (4), 1105.

Lucas, R. C. (1963) Bias in estimating recreationists' length of stay from sample interviews. *Journal of Forestry*, 61(2), 912-913

Martínez-Espiñeira, R. (2007) "'Adopt a Hypothetical Pup': A Count Data Approach to the Valuation of Wildlife" *Environmental and Resource Economics,* 37(2), 335-360.

Martínez-Espiñeira, R. & J. Amoako-Tuffour (2008). Recreation Demand Analysis under Truncation, Overdispersion, and Endogenous Stratification: An Application to Gros Morne National Park, *Journal of Environmental Management, 88(4)*, 1320-1332 http://ideas.repec.org/p/wpa/wuwpem/0511007.html.

Martínez-Espiñeira, R., J. Amoako-Tuffour & J. M. Hilbe (2006a). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure: Reevaluation. Water Resources Research 42 (W10418). DOI:10.1029/2005WR004798.

Martínez-Espiñeira, R., J. B. Loomis, J. Amoako-Tuffour & J. M.Hilbe (2008). Comparing recreation benefits from on-site versus household surveys in count data travel cost demand models with overdispersion, *Tourism Economics* 14(3), 567-576

McConnell, K. E. & I.E.C. Inc. (1986). The damages to recreational activities from PCBs in New Bedford Harbor. Unpublished report prepared for NOAA Ocean Assessment Division. Rockville, MD: December.

McKean, J. R., D. Johnson & R. G. Taylor (2003). Measuring demand for .at water recreation using a Two-Stage/Disequilibrium travel cost model with adjustment for overdispersion and self-selection. *Water Resources Research* 39 (4), 1107.

McKean, J. R., D. Johnson, R. G. Taylor & R. L. Johnson (2005). Willingness to pay for non angler recreation at the Lower Snake River reservoirs. *Journal of Leisure Research* 37 (2), 178-191.

Mendes, I. & I. M. Proença (2005). Estimating the recreation value of ecosystems by using a travel cost method approach. WP 2005/08 DE/CIRIUS, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa. http://ideas.repec.org/p/ise/isegwp/wp82005.html

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* 33, 341-365.

Ovaskainen, V., J. Mikkola & E. Pouta (2001). Estimating recreation demand with on-site data: An application of truncated and endogenously stratified count data models. *Journal of Forest Economics* 7 (2), 125-144.

Ozuna, T. & I. A. Gomez (1995). Specification and testing of count data recreation demand functions. *Empirical Economics* 20 (3), 543-550.

Parsons, G. R. (2003). The travel cost model. In P. A. Champ, K. J. Boyle & T. C. Brown (Eds.), *A Primer on Nonmarket Valuation*, Chapter 9. London: Kluwer Academic Publishing.

Phaneuf, D. J. & V. K. Smith (2006). Recreation Demand Models. In K. G. Mäler & J. R. Vincent (Eds.) *Handbook of Environmental Economics*, Chapter 15. North-Holland Elsevier Vol. 2, Number 2.

Platt, J. 2000. "Impact of Fluctuating Reservoir Elevation on Recreation Use and Value." Technical Memorandum #: EC-2000-02, U. S. Bureau of Reclamation, Technical Service Center, Economics Group, Denver CO.

Sellar, C., J. Stoll & J. Chavas (1985). Validation of empirical measures of welfare change: A comparison of nonmarket techniques. *Land Economics* 61, 156-175.

Shaw, D. (1988). On-site sample regression: Problems of non-negative integers, truncation, and endogenous stratification. *Journal of Econometrics* 37, 211-223.

Shaw, W. D., E. Fadali & F. Lupi (2003). Comparing consumer´s surplus estimates calculated from intercept and general survey data. Proceedings of the W-133 (U.S.D.A.) Regional Economics Group, compiled by J. Scott Shonkwiler. Las Vegas, Nevada, February.

Shrestha, K. R., A. F. Seidl & A. S. Moraes (2002). Value of recreational fishing in the Brazilian Pantanal: A travel cost analysis using count data models. *Ecological Economics* 42 (1-2), 289-299.

Sohngen, B., F. Lichtkoppler & M. Bielen (2000). The value of day trips to Lake Erie beaches. Technical Report TB-039, Ohio Sea Grant Extension, Columbus OH.

StataCorp (2005). *Stata Statistical Software: Release 9.1.* College Station, TX: StataCorp LP.

Yen, S. T. & W. L. Adamowicz (1993). Statistical properties of welfare measures from count-data models of recreation demand. *Review of Agricultural Economics* 15, 203-215.