



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Addressing Aggregation Bias in Zonal Recreation Models

Klaus Moeltner

Department of Applied Economics and Statistics
University of Nevada, Reno

Presented at the Annual Meeting of Western Regional Project W-133
Miami, Florida
February 26-28, 2001

The author thanks Jeffrey E. Englin, W. Douglass Shaw, and J. Scott Shonkwiler for insightful comments. Address correspondence to Klaus Moeltner, Department of Applied Economics and Statistics, University of Nevada / Reno, Mail Stop 204 / Reno, NV 89557-0105, Phone: (775) 784-6701 (work), (775) 827-8949 (home), Fax: (775) 784-1342, e-mail: moeltner@unr.edu

Abstract: Models of recreation demand are often based on zonal data. Results from such models are susceptible to aggregation bias. We propose a zonal model of recreation that captures some of the underlying heterogeneity of individual visitors by incorporating distributional information on per-capita income from Census sources into the aggregate demand function. This adjustment eliminates the unrealistic constraint of constant income across zonal residents, and thus reduces the risk of aggregation bias in estimated parameters. In addition, the corrected aggregate specification reinstates the applicability of generalized maximum likelihood methods, and increases model efficiency.

Key words: Aggregation bias; Count data models; Generalized maximum likelihood

Introduction

Resource managers and researchers are often interested in estimating welfare to visitors generated by an entire system of recreation sites. To aid in policy evaluation, budget allocation, and cross-system comparison, these measures are often most meaningful to planners if derived at an annual or seasonal level.

In recent years, researchers have begun to combine count data specifications with utility-theoretical frameworks, such as incomplete demand systems (LaFrance and Hanemann [15]), to estimate seasonal demand for multiple recreation sites (e.g. Englin et al. [4], Shonkwiler [20], Shonkwiler and Englin [21]). Most of these studies resort to recreation permits as their main source of data. Such permit data are commonly used in the economic analysis of recreation activities that requires visitors to register with local Public Land managers at each trip occasion. These data are routinely collected by authorities, and thus constitute a convenient and inexpensive source of information to the analyst. The main drawback of using permit data, however, is that visitors' characteristics, such as income, age group, or education level, are only available at an aggregate geographical level from public sources such as national census data. Depending on the scope and objective of a given study, the relevant geographical units chosen by researchers are frequently ZIP code areas (e.g. Englin et al. [4], Hilger and Englin [12], Lutz et al. [17], Shonkwiler and Englin [21]), counties (e.g. Cicchetti et al. [3], Hellerstein [9]), or other census-type units.

As is well known in econometric estimation, the use of such aggregate data in lieu of individual-specific information will generally not yield parameter estimates that accurately reflect individual behavior. This shortcoming is commonly referred to as "aggregation bias" (e.g. Stoker [24], Zellner [26]). The aggregation problem has, to date, not found much attention in the recreation literature. Notable exceptions are Hellerstein [9] and [11], although these studies focus more on the benefits of using aggregate data in combination with robust estimation methods rather than on the nature and problems related to aggregation bias.

The aim of this paper is to examine the econometric issues associated with the use of aggregate data within the context of multi-site recreational demand and welfare estimation. It will be shown that aggregation problems can be alleviated by incorporating additional zonal information into count data specifications. This remedial technique also increases model efficiency.

The remainder of this text is structured as follows: The following section briefly discusses some relevant econometric theory associated with the aggregation problem, as well as econometric issues that arise when aggregate data is used in the context of count data models. The next segment delineates a "corrected" count data model that reduces the undesired effects of aggregation bias. The empirical part of this study then discusses data, estimation results, and welfare measures. It is followed by a sensitivity analysis of welfare results. Concluding remarks and a summary of key findings are given in the last section.

Using Aggregate Information in Count Data Models – Econometric Considerations

Econometricians have been aware for a long time that econometric models using aggregate information, such as averages or totals over sub-groups of individual observations, generally fail in systematic ways to take account of underlying heterogeneous behavioral reactions. This makes it often difficult to assign a meaningful interpretation to estimated model parameters. Stoker [24] elaborates on this notion using numerous examples from the empirical

literature. He refers to the condition under which an aggregate specification unambiguously identifies an underlying micro function as "completeness", and the ability of retrieving individual-specific parameters from an aggregate model as "recoverability". In Stoker [23] he proposes the interpretation of the aggregate relationship as a reduced-form model, and the recoverability of individual- level, structural coefficients as an identification issue.

In the context of recreation demand, an unobserved micro-function could take the following general form:

$$d_{ij} = f(x_{ij}; \beta), \quad (1)$$

where d_{ij} is an individual's demand (in visits) for site j , x_{ij} is a vector of site and individual characteristics, and β is a vector of parameters. The corresponding aggregate or macro function could then be derived from (1) as

$$d_{ij}(\beta, \theta) = E_x(d_{ij}) = \int f(x_{ij}; \beta) p(x_{ij}; \theta) dx, \quad (2)$$

where the capital I in d_{ij} indicates that the elements of x_{ij} have been aggregated over some subgroup I (e.g. a ZIP code area, or some other geographical unit). The x -subscript to the expectation sign indicates that mathematical expectations are taken over regressors x_{ij} , which follow a probability density function p with parameter vector θ . Thus, d_{ij} can be regarded as demand by a representative agent from zone I .

The core segment of Stoker [23] identifies restrictions on (1) and $p(x; \theta)$ for completeness to hold. The key result for this study is that if $p(\cdot)$ belongs to the family of exponential distributions, completeness is ascertained for any functional form of (1). In addition, micro-level coefficients can generally be recovered from (2) for such models if information on θ is available.

Hellerstein [11] illustrates this notion within the framework of a simulated count data model for recreation demand by assuming a joint-normal distribution for his two right hand side variables, travel cost and income. He shows how this additional information can be incorporated into the specification of an aggregate Poisson model. In a variety of simulation settings, he demonstrates that this "corrected" zonal model yields more accurate welfare estimates than generic aggregate specifications. He also finds that for many of his visitation scenarios the corrected and, in some cases, even the standard aggregate model perform better than their counterparts based on limited individual observations.

The latter result indicates that there are inherent potential benefits of using aggregate regressor information in count data models. Unlike specifications based on individual information zonal models generally do not need to be corrected for truncation and / or endogenous stratification (e.g. Hellerstein [10], Shaw [19]), since information on non-participants in aggregate form is readily available from census data at any desired geographical level. Thus, they do not require additional distributional assumptions on the dependent variable associated with such adjustments, and avoid reliance on estimators that are highly sensitive to model misspecification. Furthermore, as discussed in [9] and [11], non-limited count data models lend themselves to estimation through generalized maximum likelihood (ML) techniques, such as pseudo- and quasi-generalized pseudo ML methods (PML and QGPML, respectively - Gouriéroux et al. [6] and [7]).¹ These robust estimators offer additional insurance

against misspecification of the distribution of the dependent variable, and are often computationally less burdensome than standard ML methods. However, the requirement for PML and QGPML to generate consistent parameter estimates is the availability of a consistent estimate of the moments of the dependent variable. These will not be available if aggregation bias is present, which effectively renders these useful estimation techniques inapplicable for most zonal models.

From the discussion above it is clear that in order to reduce the risk of aggregation bias, and to fully exhaust the benefits in zonal models, additional information on the distribution of explanatory variables is needed. The following section shows how estimates of distribution moments for some commonly used regressors in recreation demand models can be extracted from Census data and incorporated into a count data framework.

A Corrected Aggregate Count Data Model of Recreation Demand

We apply the Incomplete Demand System framework discussed in LaFrance [14], and follow Hilger and Englin [12], Shonkwiler [20], and Shonkwiler and Englin [21] by choosing a semi-log form for the expected demand equations:²

$$E(d_{ij}) = \lambda_{ij} = \exp(\alpha_i + \beta'_{q,j} \cdot q_j + \beta_{p,ij} \cdot p_{ij} + \beta'_{p,ik} \cdot p_{ik} + \beta_{m,i} \cdot m_i), \quad (3)$$

where d_{ij} is the actual demand by individual i for site j , α_i is an individual-specific intercept, q_j is a vector of trail features, p_{ij} is the price of site j to individual i measured as travel cost from i 's residence to trail j , p_{ik} is a vector of prices to all other sites, m_i denotes individual income, and the β -terms are coefficients. Specifically, $\beta_{p,ij}$ is the own-price coefficient for site j , and $\beta_{p,ik}$ is a vector containing all cross-price coefficients. Imposing the cross-equation restrictions given in LaFrance [14], and the additional simplifying constraints of origin- and destination-invariant travel cost coefficients and preferences for trail features, (3) reduces to

$$E(d_{ij}) = \lambda_{ij} = \exp(\alpha_i + \beta'_q \cdot q_j + \beta_p \cdot p_{ij} + \beta_m \cdot m_i) \quad (4)$$

Based on this specification, the IDS system yields the following expressions for expected CV and EV, assuming positive income effects (LaFrance [14]):

$$\begin{aligned} E(CV_i) &= \left(\frac{1}{\beta_m} \right) \cdot \ln \left[1 + \frac{\beta_m}{\beta_p} \cdot \sum_{j=1}^J (a_{ij} \cdot \exp(\beta_p \cdot p^{\circ}_{ij}) - a_{ij} \cdot \exp(\beta_p \cdot \tilde{p}_{ij})) \right] \\ E(EV_i) &= - \left(\frac{1}{\beta_m} \right) \cdot \ln \left[1 - \frac{\beta_m}{\beta_p} \cdot \sum_{j=1}^J (a_{ij} \cdot \exp(\beta_p \cdot p^{\circ}_{ij}) - a_{ij} \cdot \exp(\beta_p \cdot \tilde{p}_{ij})) \right], \end{aligned} \quad (5a,b)$$

where p°_{ij} is the price of site j to individual i in the original state, and \tilde{p}_{ij} is the price in the new state. The two equations constitute individual, per-season welfare measures resulting from a price change for one or more destinations within the recreation system. To capture the entire area under the compensated demand curve associated with either CV or EV, one can set \tilde{p}_{ij} to infinity. In that case the last terms in (5a,b) vanish.

In the spirit of Stoker's terminology presented above, (4) constitutes the unobserved micro-function for this model. Only trail features (i.e. the elements of q_i) are known, while aggregate values need to be substituted for α_i , p_{ij} , and m_i . As discussed below in more detail, aggregate information was collected for 134 ZIP-codes for this study. In order to keep the number of parameters to be estimated within reasonable limits, these ZIP code areas are further grouped into 35 population zones (PZ) for some parts of this analysis.

Since the micro-function is non-linear, the recoverability of the coefficients in (4) hinges on assumptions regarding the distribution of the three variables in question within a given ZIP code. Since no such information can be gleaned from public data sources for α_i , we simply assume that all residents of a given PZ share the same preferences as reflected by the intercept term. In other words, we stipulate that $\alpha_i = \alpha_Z, \forall i \in Z, Z = 1 \dots 35$, where Z denotes a given population zone.

The computation of an aggregate value for p_{ij} requires a closer look at the exact definition of this variable. To facilitate the comparison of results generated by this model with those found in other studies, we follow the conventional approach of defining travel cost as a combination of mile cost and time cost, using a third of the hourly wage rate to capture the opportunity cost of time. Specifically,

$$\begin{aligned}
 p_{ij} &= ml_{ij} + f_{ij} + \gamma \cdot m_i \cdot t_{ij}, \quad \text{with} \\
 ml_{ij} &= \frac{(2 \cdot \delta_{ij} \cdot \$0.25)}{g_{ij}} \\
 \gamma &= \frac{0.33}{2080} \\
 t_{ij} &= \left(\frac{\delta_{hwy,ij}}{60} \right) + \left(\frac{\delta_{acc,ij}}{30} \right),
 \end{aligned} \tag{6a-d}$$

where ml_{ij} are mile costs based on a vehicle operating expense of \$0.25/mile,³ δ_{ij} is the distance, in miles, from i 's residence to site j , and g_{ij} is the number of passengers traveling in the same vehicle. Time cost, t_{ij} , is divided into a highway and an access-road component, where an access road is defined as the trip segment between highway exit and final destination. As reflected in (6d), we allow for twice as high a speed level on highways than on access roads. For some origins, visitors also incur ferry fees on their way to the recreation system. For these ZIP areas, an appropriate cost term, f_{ij} , was added to (6a). Since information on the exact location of a visitor's residence is not available, we make the simplifying assumption that all residents of a given ZIP area face the same distance in miles to a specific site within the recreation system of interest, i.e. $\delta_{ij} = \delta_{j}, \forall i \in I, I = 1 \dots 134$. Also, we substitute the average group size for a given ZIP as captured by permit data for g_{ij} , i.e. $g_{ij} = \bar{g}_j, \forall i \in I, I = 1 \dots 134$.

To this point, all components of the micro-function (4) are assumed to be equal for all residents of a specific ZIP area. In a simple aggregate model, one would extend this notion also to personal income, m_i , and estimate the following equation for the expected demand of a representative individual from zone I for destination j :

$$\tilde{\lambda}_{ij} = \exp\left(\tilde{\alpha}_z + \tilde{\beta}'_q \cdot q_j + \tilde{\beta}_p \cdot \bar{p}_{ij} + \tilde{\beta}_m \cdot \bar{m}_I\right), \quad (7)$$

where \bar{m}_I is the mean per-capita income for ZIP I , and \bar{p}_{ij} is given by (6a) with implementation of distance and group size constraints, and m_i replaced by \bar{m}_I . Since the assumption of zero variance for income is unrealistic, such a "naïve" model will most likely be flawed by aggregation bias. The resulting parameter values and predictions for site demand will not be consistent estimates for the ones given in (4). This is reflected by the "~" superscripts for coefficients and expected demand in (7).

However, information on the distribution of m_i for a given ZIP area can be extracted from common Census data sources. This additional information allows for the specification of an improved aggregate model. Based on a graphical inspection of income histograms, we stipulate that, for each ZIP area, income is distributed normally within two population groups: white ("w") and non-white ("nw"). Since the normal distribution belongs to the exponential family, this specification satisfies Stoker's [23] completeness condition mentioned above. Income, through the hourly wage rate, also enters the definition of travel cost (see 6a), which implies that the two regressors are distributed joint-normally for each race group, i.e.

$$\begin{aligned} \begin{bmatrix} p_{ij,r} \\ m_{i,r} \end{bmatrix} &\sim mvn(\mu_{ij,r}, \Omega_{ij,r}), \quad \text{where} \\ \mu_{ij,r} &= \begin{bmatrix} nt_{ij} + \gamma \cdot t_{ij} \cdot \bar{m}_{I,r} \\ \bar{m}_{I,r} \end{bmatrix}, \quad \Omega_{ij,r} = \begin{bmatrix} (\gamma \cdot t_{ij})^2 \cdot \sigma_{I,r}^2 & \gamma \cdot t_{ij} \cdot \sigma_{I,r}^2 \\ \gamma \cdot t_{ij} \cdot \sigma_{I,r}^2 & \sigma_{I,r}^2 \end{bmatrix} \\ nt_{ij} &= ml_{ij} + f_{ij} \\ r &= w, nw \end{aligned} \quad (8)$$

The notation is as follows: mvn denotes the multivariate normal distribution, nt_{ij} captures non-time costs (mile and ferry expenses), $\bar{m}_{I,r}$ and $\sigma_{I,r}^2$ are the mean and variance of income for ZIP I and ethnic group r , and the remaining terms are defined in (6c,d). As discussed in Hellerstein [11] the exponents of travel cost and income for each race group will be jointly log-normally distributed with the following mean and variance:

$$\begin{aligned} E\left(\exp\left[\begin{bmatrix} p_{ij,r} \\ m_{i,r} \end{bmatrix}\right]\right) &= \exp\left(\mu_{ij,r} + \frac{1}{2} \cdot \Omega_{ij,r}\right), \\ V\left(\exp\left[\begin{bmatrix} p_{ij,r} \\ m_{i,r} \end{bmatrix}\right]\right) &= \exp(2 \cdot \mu_{ij,r} + 2 \cdot \Omega_{ij,r}) - \exp(2 \cdot \mu_{ij,r} + \Omega_{ij,r}) \\ r &= w, nw \end{aligned} \quad (9)$$

The census also reveals the population share for each ethnic segment. Using this information in combination with (8) and (9), we can define a corrected macro-function for expected individual site demand as

$$\begin{aligned}\lambda_{ij} &= E_x(\lambda_{ij}) = pw_I \cdot \lambda_{ij,w} + (1 - pw_I) \cdot \lambda_{ij,nw} \\ &= pw_I \cdot \exp\left(\alpha_Z + \beta'_q \cdot q_j + \beta'_s \cdot \mu_{ij,w} + \frac{1}{2} \cdot \beta'_s \cdot \Omega_{ij,w} \cdot \beta_s\right) + \\ &+ (1 - pw_I) \cdot \left(\alpha_Z + \beta'_q \cdot q_j + \beta'_s \cdot \mu_{ij,nw} + \frac{1}{2} \cdot \beta'_s \cdot \Omega_{ij,nw} \cdot \beta_s\right), \text{ with} \\ \beta_s &= \begin{bmatrix} \beta_p \\ \beta_m \end{bmatrix},\end{aligned}\tag{10}$$

where pw_I indicates the share of white residents in ZIP area I .⁴ The exact derivation of income moments for a given ZIP and ethnic group is shown in Appendix A.

Based on preliminary specification tests and fit with the underlying data, we choose Cameron and Trivedi's [1] Negative Binomial II (Negbin II) specification as the specific count data framework to be used for this analysis. Using the result that the sum of a Negbin random variable follows the same distribution with first moment given by the sum of the individual expectations, the aggregate model to be estimated emerges as

$$\begin{aligned}p(D_{ij} = Y_{ij}) &= \frac{\Gamma(Y_{ij} + N_I \cdot v)}{\Gamma(Y_{ij} + 1) \cdot \Gamma(N_I \cdot v)} \cdot \left(\frac{v}{\lambda_{ij} + v}\right)^{N_I \cdot v} \cdot \left(\frac{\lambda_{ij}}{\lambda_{ij} + v}\right)^{Y_{ij}}, \text{ with} \\ D_{ij} &= \sum_{i=1}^{N_I} d_{ij},\end{aligned}\tag{11}$$

where N_I indicates the total population of ZIP area I , λ_{ij} is the representative individual's demand defined in (10), D_{ij} is the (unobserved) total micro-demand from ZIP I to destination j , Γ denotes the mathematical gamma function, and v is the index or precision parameter. The integer value Y_{ij} indicates a realization of total visits per season from ZIP I to site j . The model in (11) is then estimated via QGPML. Appendix B shows the details of this process.

Application

The system demand model developed in the previous section is implemented using information on day trips to the Alpine Lakes Wilderness (ALW), which covers 393,000 acres in Washington's Mt. Baker-Snoqualmie National Forest. The area's proximity to some of the State's largest population centers makes it a popular hiking destination, with over 100,000 visitors per year. The numerous trails in the Wilderness are accessible through 51 trailheads, located around the Wilderness boundaries. Forty-nine of these "hiking zones" (HZs) are included in this analysis.

Permit data were provided by the National Forest Service.⁵ The original set includes valid information on 14087 hiking groups, collected for the entire year of 1995. After filtering out multiple-day hikers and observations from ZIP codes that could not be matched with those included in the 1990 Census, 8750 valid observations on day-use group-visits from 134 different

ZIP code origins were retained for this analysis. The total number of residents in these zones is 3.01 million, roughly 55% of the State's population. Aggregating visits to a given HZ over ZIPs and introducing zero-visit values for the relevant ZIP / trail combination not represented by the original data yields a rectangular set of $(134 \times 49) = 6566$ observations. These data are then combined with trail and access road information from an ALW guidebook (Spring et al. [22]), highway distances from a road atlas, and ZIP-specific information from the 1990 U.S. Census, as discussed in the previous section.

Estimation Results

We estimate both a simple aggregate Negbin (SANB) and a corrected aggregate Negbin (CANB) model. The simple model uses the naïve specification for λ_{ij} (7) in the estimation of (11) while the corrected model applies (10) instead. Both models are estimated using QGPMML.⁶ Robust variance-covariance matrices for parameters are constructed applying the method suggested by White [25], and Gourieroux et al. [6]. This implies that SANB yields consistent estimates for coefficients and standard errors if (7) is the correct aggregate specification. The same is true for CANB with respect to (10). Naturally, since (7) is incorrect by assumption, the simple model will be flawed by aggregation bias, and bias resulting from using inconsistent estimates for λ_{ij} in the Gourieroux et al. procedures. It is retained in this analysis merely for comparison purposes.

Table 1 presents parameter estimates for the two models. The components of q are trailhead elevation (measured in 1000 ft-units), a dummy taking the value of "1" if a lake can be reached within 5 miles of hiking from a given trailhead, and a dummy equal to "1" if the access road as defined above is longer than 5 miles. Travel cost are measured in dollars, and income in \$1000 units. The "PZ-" dummies correspond to α_z described above. Both models fit the underlying data fairly well, with about two thirds of coefficient estimates significant at the 1 percent or 5 percent level. In general, the differences in parameter estimates and standard errors between the corrected and simple model are rather subtle. However, the corrected model produces a significantly lower and clearly more efficient parameter estimate for income, the variable subjected to the largest adjustments in the corrected version. To a smaller extent, this also holds for travel cost, which is a function of income as discussed above.

For both models, easy access to an alpine lake increases visitation rates by over 200 percent,⁷ *ceteris paribus*. Also, visitors seem to have a slight preference for high-elevation trails based on the positive sign and significance of the corresponding coefficients. This result may be rooted in better views associated with higher elevations, and an increased chance for encountering alpine meadows, a generally positively valued trail feature (Englin and Shonkwiler [5]). In contrast, lengthy access roads are not as strong a deterrent as expected, based on the lacking significance of this coefficient. Perhaps this reflects the increased popularity of Sport-Utility Vehicles ("SUVs") and associated reduced problems in maneuvering these mostly unpaved secondary roads. The estimates for the different population zone dummies generally indicate that smaller, more rural communities have stronger preferences for day hiking in the ALW than larger population centers.

Beyond an inspection of coefficient estimates, the comparison of these models is somewhat problematic. Likelihood ratio tests are inapplicable since different regressors are used for the two specifications. Instead, we use goodness-of-fit statistics based on Pearson residuals,

and deviance. These statistics are discussed in detail in Cameron and Windmeijer [2]. Based on these two measures, the corrected model performs slightly better than its simple counterpart.⁸

Estimates for elasticities and welfare measures for the two Negbin models are summarized in table 2. Elasticity measures are based on sample averages for travel cost and per-capita income, following e.g. Lutz et al. [17]. For CANB, they also reflect the average over the two race groups, weighted by population share. The own-price elasticity of demand for a "prototypical site" with price \bar{p} is slightly greater than 1 for both models. The difference in elasticity estimates is more pronounced for income elasticity, which SANB estimates at 1.14, thirty percent higher than the values generated by CANB. The CANB figures of 0.85 to 0.88 seem to be more reasonable estimates considering the general finding in the literature of inelastic demand for recreation activities with respect to income (Loomis and Walsh [16]). For both elasticity groups, the corrected Nb model generates a tighter 95 percent confidence interval, as measured by the spread between lower and upper bound, divided by the mean estimate. This difference is especially pronounced for income elasticity, where the weighted spread for SANB is more than twice as large as the CANB counterpart for all residents (0.72 vs. 0.34).

Welfare estimates in table 2 are measured by compensating variation (CV), based on (5a).⁹ Estimated per-trip welfare measures of \$31-33 are consistent with recent findings in other studies on hiking behavior in Washington's Cascade mountains (e.g. Englin and Shonkwiler [5]). The corrected aggregate Negbin model produces a value of \$1.83 million in total welfare generated by the ALW to the target population in 1995. The simple aggregate model overestimates total welfare by approximately \$90,000, or close to 5 percent. We use the simulation method suggested by Krinsky and Robb [13] to derive 95 percent confidence intervals for these values.¹⁰ The results of this procedure mirror the findings for elasticities discussed above: The more efficient corrected model produces much tighter intervals. As measured by the spread-over-mean statistic, CANB outperforms SANB by 20 percent.

Sensitivity Analysis for Corrected Model

An inspection of (7) and (10) shows that the difference in welfare estimates between the two models largely depends on income effect, population shares, and income variance. To investigate the sensitivity of our results to model specifications, we perform a series of simulations with the same trip data, but altered population characteristics.¹¹ The results of these tests are captured in table 3. In the first simulation sequence (models 1a – 1e), we gradually reduce the proportion of white residents from an original ZIP average of 90 percent by up to 50 percent for all areas. This has a somewhat ambiguous effect on the bias of the simple model (measured as ratio of CVs / CVc, where *s* and *c* subscripts denote the simple and corrected model, respectively), but leads to a monotonic increase in efficiency, i.e. a progressively tighter confidence interval for CV generated by the corrected model, compared to the simple specification (last column of table 3). The second series (models 2a – 2e) represents a gradual increase in income variance from 10 percent to 300 percent. This leads to an opposite effect of monotonically increasing bias (from 5 percent to 10 percent), but ambiguous efficiency gains. The last test series, models 3a – 3d, combines a 20 percent decrease in the share of white residents with increases in income variance. Now both bias and efficiency loss of the simple model increase monotonically. Doubling income variance, for example, (model 3b) leads to an increase in bias of CV point estimates from 5 percent to 8 percent, and an increase in the relative

width of the mean-calibrated confidence interval for CVs vs. CVc from 21 percent to a full 40 percent.

Such population proportions and income variances are not unrealistic, especially for larger urban population zones. This is illustrated in table 4, where the ZIP population and income characteristics of models 3a – 3d are compared to such information for five Metropolitan Statistical Areas (MSAs). As shown in column 3, the 20 percent reduction in whites across the 134 ZIPS used for this study leads to an average proportion of this group of around 72 percent, which compares well to analogous proportions for the MSAs. The remainder of the table shows averages and percentiles, over all ZIPs, of mean / variance ratios for income for both population groups. For example, a doubling of income variance for our data (model 3b) yields an average mean / variance ratio of 5 percent for white residents, and 11 percent for non-whites. Median values (at the 50th percentile) are similar for this model. In comparison, such ratios correspond roughly to the lower quartile for ZIPs in MSAs. The upshot of this analysis is that a strong representation of recreation participants from zonal origins with racial diversity and large income variances in a given research project for could well lead to substantial aggregation bias and efficiency losses for welfare estimates in uncorrected zonal models.

Conclusion

This study shows that problems stemming from the use of aggregate data in zonal recreation models can affect both the theoretical relevance and analytical accuracy of estimation results. At a theoretical level, the parameter and welfare estimates generated by a misspecified aggregate model carry little informative value regarding the underlying recreation preferences of heterogeneous individuals. From an econometric perspective, the main drawback of an incomplete macro-specification is the loss of applicability of robust estimation techniques. This, in turn, exposes such models to additional sources of misspecification error.

In this paper we propose an extension to existing zonal count data models of recreation that captures some of the underlying heterogeneity of individual visitors by incorporating distributional information on per-capita income into the aggregate demand function. This adjustment eliminates the unrealistic constraint of constant income across zonal residents, and thus reduces the risk of aggregation bias in estimated macro-parameters. The suggested technique is easy to implement, and does not require the collection of additional data.

While relative bias and efficiency losses of welfare estimates generated by the uncorrected model are noticeable for the data used in this study, a follow-up analysis shows that the benefits of using the corrected specification increase with participation of residents from origin zones characterized by racial heterogeneity and high income variances.

Naturally, the enhanced count data framework suggested in this study will not completely eliminate aggregation problems in any given application. We believe, however, that our proposed adjustment strengthens the linkage between individual preferences and aggregate estimation for a wide spectrum of underlying micro-functions, and thus constitutes a general improvement over existing recreation models that rely on zonal data. Full recoverability of parameters for a micro-specification with fewer restrictions will almost certainly require additional information at the individual level.

Appendix A: Derivation of Income Variance from Census Information

In the 1990 U.S. Census, households are grouped into discrete income categories for each ZIP code. This information is also available in a nested version over race groups, although the number of income categories for this exposition is smaller than for the general grouping (9 vs. 25). Naturally, the lower bound for this discrete income distribution is given by zero. An upper bound, on the other hand, is not reported. The population mean and total for household income, \bar{m}_{hh} and m_{tot} , are reported by the Census for any given race group. The remaining task is to estimate income variance.

As a first step, we estimate mean income in the highest category as

$$\bar{m}_K = \left(\frac{m_{tot} - \sum_{k=1}^{K-1} n_k \cdot \left(\frac{m_{\max,k} - m_{\min,k}}{2} \right)}{n_K} \right), \quad (A1)$$

where k is a category index, K indicates the highest income category, m_{tot} is total ZIP income, $m_{\min,k}$ and $m_{\max,k}$ are the lower and upper bounds of income category k , and n_k and n_K are the number of households in category k and the highest category, respectively. Race and ZIP-subscripts are omitted for convenience. The variance of household income can then be approximated by

$$\sigma_{hh}^2 = \frac{\left(\sum_{k=1}^{K-1} n_k \cdot \left(\left(\frac{m_{\max,k} - m_{\min,k}}{2} \right) - \bar{m}_{hh} \right)^2 \right) + n_K \cdot (\bar{m}_K - \bar{m}_{hh})^2}{N - 1}, \quad (A2)$$

i.e. the average, over all households, of the sum of the squared deviations of bin midpoints (bin mean for the highest category) from the overall mean, with each such deviation weighted by the number of households in a given category.¹² This leads to the expression of the variance for per-capita income:

$$\sigma_{I,r}^2 = \left(\frac{1}{\bar{h}_{I,r}} \right)^2 \cdot \sigma_{hh,I,r}^2, \quad r = (w, nw), \quad (A3)$$

where $\bar{h}_{I,r}$ is the average household size for race group r in ZIP area I .

Appendix B: Model Estimation Through Generalized Maximum Likelihood

A correct specification of the aggregate demand function notwithstanding, parameter estimates can still be biased if the probability density function for d_{ij} is misspecified (Gourieroux et al. [6] and [7]). Specifically, if the distributional assumptions reflected in (11) do not hold for the Negbin model, ML estimates will be inconsistent. PML and QGPML methods guard against such specification error. For the Negbin model, either of the two estimation techniques can be used to derive consistent coefficients and robust standard errors, but QGPML estimators are generally more efficient. Both methods differ from ML by arbitrarily specifying a value for ν in (11) instead of treating it as an additional model parameter. This reduces the relevant part of the LLF to

$$l_{(Nb)} = \sum_{I=1}^{I_T} \sum_{j=1}^J \left[a \cdot (\ln(a) - \ln(\Lambda_{ij} + a)) + Y_{ij} \cdot (\ln(\Lambda_{ij}) - \ln(\Lambda_{ij} + a)) \right], \quad \text{where} \quad (B1)$$

$$\Lambda_{ij} = N_I \cdot \lambda_{ij}$$

Following the notation in Gourieroux et al. [7], a represents the substitute for $(N_I \cdot \nu)$. The QGPML process requires two additional estimation steps. In an interim procedure, a consistent estimate for $V = N_I \cdot \nu$ is derived by using the Negbin II expression for the variance of D_{ij} and the parameter estimates from maximization of (B1) in the following regression:

$$(Y_{ij} - \hat{\Lambda}_{ij})^2 - \hat{\Lambda}_{ij} = \frac{1}{V} \cdot \hat{\Lambda}_{ij}^2 \quad (B2)$$

The OLS estimate for V in (B2), \hat{V} , is then inserted into (B1) in lieu of a . A second round of optimization using this adjusted LLF generates the final results.

Notes

¹ The theoretical underpinnings to these concepts are discussed in Gourieroux et al. [6]. In a related study [7] the same authors show how PML and QGPML estimation can be used in the context of count data models. The general idea behind this approach is appealing: If the dependent variable, say trips to a recreation site, is characterized by a probability distribution that is a member of the linear exponential family, such as Poisson or Negbin, ML estimation yields consistent parameter estimates, even if the empirical model does not reflect the true distribution of site demand.

² Within the framework of a count data model actual demands by individual i for site j are assumed to be unobserved. Site demand d_{ij} is treated as a random variable with an expected value commonly labeled as λ_{ij} .

³ Source: U.S. DoT Web Site: <http://www.bts.gov/btsprod/nts/chp2/tbl2x18.html>. The \$0.25/mile figure constitutes a compromise between \$0.41 for total, and \$0.1 for variable costs (1995).

⁴ Naturally, this specification implies that the proportions of race groups among visitors are equal to racial shares in the general ZIP population at all destinations. If this assumption seems implausible in a given application, equation (10) can easily be modified to allow for population shares specific to a given origin – destination pair, if empirical estimates or “educated guesses” for these shares are available. This generalization would be captured in (10) by assigning “ I_j ” subscripts to pw , and pnw . All other population parameters would be still based on ZIP code characteristics, and remain unchanged.

⁵ We are indebted to Gary Paull, Wilderness trail coordinator, and his colleagues at the Mt. Baker-Snoqualmie NF headquarters in Mt. Lake Terrace, WA, for provision of data and other helpful information.

⁶ We use Matlab’s Trust Region Method with user-supplied gradients and Hessian matrices as optimization algorithm for this analysis. The full Matlab program is available from the author upon request. We thank Daniel Hellerstein for providing count data GAUSS code, which proved very useful in developing this program.

⁷ We use Halvorsen and Palmquist’s [8] formula of $(\exp(\beta) - 1)$ as an approximation for this marginal effect.

⁸ The R-squared values for CANB and SANB are 0.306 vs. 0.271 based on Pearson residuals, and 0.348 and 0.335 based on deviance.

⁹ We also computed estimates for consumer surplus (CS) and equivalent variation (EV). The three welfare measures differ only marginally from each other within a specific aggregate model. This is an expected result, given the small income effects mentioned earlier. The relatively insignificant magnitude of these divergences notwithstanding, the three measures are still ranked as required by utility-theory for an increase in prices, with $|CV| > |CS| > |EV|$ for both models.

¹⁰ Specifically, we draw 1000 sets of coefficients, based on parameter estimates and their variance-covariance matrix. For each set, we then derive a corresponding set of λ_{ij} for all I_j combinations. This leads to the computation of CV for each ZIP code area, using (5a) weighted by ZIP population. These values are added over all ZIPs to yield 1000 estimates of total CV. The reported confidence intervals are based on the empirical distribution of these estimates. Elasticities, on the other hand, are linear in parameters for our model, so the derivation of their standard deviation and corresponding confidence intervals is straightforward.

¹¹ These simulations are based on the assumption that the alterations in population characteristics would not change observed trip behavior. This assumption seems tenable, given that mean income remains unchanged for all ZIPs, and income effects are generally small.

¹² This variance estimator is slightly biased upwards, but generally performed well in Monte Carlo simulations. A possible extension for the computation of $\sigma^2_{I,r}$ would be to apply more advanced nonparametric techniques, such as the histospline - method proposed by Minnotte [18].

References

- Cameron, A. Colin, and Pravin K. Trivedi, Econometric models based on count data: Comparison and applications of some estimators and tests, *Journal of Applied Econometrics* **1**, 29-53 (1986).
- Cameron, A. Colin, and Frank A. G. Windmeijer, R-Squared measures for count data regression models with applications to health-care utilization, *Journal of Business & Economic Statistics* **14**, 209-220 (1996).
- Cicchetti, Charles J., Anthony C. Fisher, and V. Kerry Smith, An econometric evaluation of a generalized consumer surplus measure: The Mineral King controversy, *Econometrica* **44**, 1259-1276 (1976).
- Englin, Jeffrey, Peter Boxall, and David Watson, Modeling recreation demand in a Poisson system of equations: an analysis of the impact of international exchange rates, *American Journal of Agricultural Economics* **80**, 255-263 (1998).
- Englin, Jeffrey, and J. S. Shonkwiler, Estimating social welfare using count data models: An application to the long-run recreation demand under conditions of endogenous stratification and truncation., *The Review of Economics and Statistics* **77**, 104-12 (1995).
- Gourieroux, C., A. Monfort, and A. Trognon, Pseudo maximum likelihood methods: theory, *Econometrica* **52**, 681-700 (1984a).
- Gourieroux, C., A. Monfort, and A. Trognon, Pseudo maximum likelihood methods: application to Poisson models, *Econometrica* **52**, 701-720 (1984b).
- Halvorsen, Robert, and Raymond Palmquist, The interpretation of dummy variables in semi-logarithmic equations, *American Economic Review* **70**, 474-475 (1980).
- Hellerstein, Daniel, Using count data models in travel cost analysis with aggregate data, *American Journal of Agricultural Economics* **73**, 860-867 (1991).
- Hellerstein, Daniel, The treatment of nonparticipants in travel cost analysis and other demand models, *Water Resources Research* **28**, 1999-2004 (1992).
- Hellerstein, Daniel, Welfare estimation using aggregate and individual-observation models: a comparison using Monte Carlo techniques, *American Journal of Agricultural Economics* **77**, 620-630 (1995).
- Hilger, James, and Jeffrey Englin, Utility theoretic linear exponential demand systems: an analysis of forest fire impacts on recreational values and uses, *Working paper*, Department of Applied Economics and Statistics, University of Nevada, Reno (2000).
- Krinsky, Itzhak, and A. Leslie Robb, On approximating the statistical properties of elasticities, *Review of Economics and Statistics* **68**, 715-719 (1986).
- LaFrance, Jeffrey T., Incomplete demand systems and semilogarithmic demand models, *Australian Journal of Agricultural Economics* **34**, 118-131 (1990).
- LaFrance, Jeffrey T., and W. Michael Hanemann, The dual structure of incomplete demand systems, *American Journal of Agricultural Economics* **71**, 262-274 (1989).
- Loomis, John B., and Richard G. Walsh. *Recreation economics decisions - comparing benefits and costs*. 2nd edition ed. State College, PA: Venture Publishing, Inc. (1997).
- Lutz, Janet, Jeffrey Englin, and J. Scott Shonkwiler, On the aggregate value of recreational activities - A nested price index approach using Poisson demand systems, *Environmental and Resource Economics* **15**, 217-226 (2000).
- Minnotte, Michael C., Achieving higher-order convergence rates for density estimation with binned data, *Journal of the American Statistical Association* **93**, 663-672 (1998).

- Shaw, Daigee, On-site samples' regression - problems of non-negative integers, truncation, and endogenous stratification, *Journal of Econometrics* **37**, 211-223 (1988).
- Shonkwiler, J. S. Recreation demand systems for multiple site count data travel cost models, in "Valuing recreation and the environment: revealed preference methods in theory and practice", (J. A. Herriges and C. L. Kling, Ed.). Cheltenham, U.K., and Northampton MA: Edward Elgar (1999).
- Shonkwiler, J. S., and Jeffrey Englin, Welfare losses due to livestock grazing on public lands: some evidence from the Hoover Wilderness, *Paper presented at the W-133 Meetings in Tucson, Arizona* (1999).
- Spring, Vicky, Ira Spring, and Harvey Manning. *100 Hikes in Washington's Alpine Lakes Wilderness*. 2nd ed. Seattle: The Mountaineers (1993).
- Stoker, Thomas M., Completeness, distribution restrictions, and the form of aggregate functions, *Econometrica* **52**, 887-908 (1984).
- Stoker, Thomas M., Empirical approaches to the problem of aggregation over individuals, *Journal of Economic Literature* **31**, 1827-1874 (1993).
- White, Halbert, Maximum likelihood estimation of misspecified models, *Econometrica* **50**, 1-25 (1982).
- Zellner, Arnold, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *American Statistical Association Journal* 351-368 (1962).

Table 1: Estimation Results

a (b) = sign. at 1% (5%)		SANB		CANB	
Variable		coeff.	s.e	coeff.	s.e
Intercept		-9.630	0.278 ^a	-9.427	0.194 ^a
Trailhead elevation		0.126	0.039 ^a	0.144	0.051 ^a
Lake w/in 5 miles		1.152	0.075 ^a	1.119	0.078 ^a
Acces road > 5 miles		-0.060	0.072	-0.001	0.083
Travel cost		-0.031	0.004 ^a	-0.032	0.003 ^a
Income		0.060	0.011 ^a	0.045	0.004 ^a
PZ 2		-0.217	0.172	-0.217	0.168
PZ 3		0.815	0.188 ^a	0.827	0.184 ^a
PZ 4		0.149	0.206	0.159	0.204
PZ 5		1.146	0.266 ^a	1.137	0.251 ^a
PZ 6		1.566	0.289 ^a	1.528	0.280 ^a
PZ 7		0.614	0.203 ^a	0.593	0.175 ^a
PZ 8		0.326	0.184	0.368	0.185 ^b
PZ 9		-0.590	0.190 ^a	-0.619	0.204 ^a
PZ 10		2.360	0.321 ^a	2.226	0.309 ^a
PZ 11		1.250	0.388 ^a	1.098	0.380 ^a
PZ 12		0.806	0.273 ^a	0.685	0.273 ^b
PZ 13		2.203	0.410 ^a	2.094	0.390 ^a
PZ 14		1.970	0.415 ^a	1.828	0.397 ^a
PZ 15		-0.212	0.269	-0.141	0.266
PZ 16		-0.791	0.266 ^a	-0.845	0.261 ^a
PZ 17		0.344	0.388	0.448	0.408
PZ 18		0.304	0.267	0.196	0.259
PZ 19		2.123	0.289 ^a	2.081	0.297 ^a
PZ 20		-0.297	0.278	-0.225	0.372
PZ 21		0.719	0.369 ^b	0.669	0.347
PZ 22		-1.355	0.370 ^a	-1.334	0.388 ^a
PZ 23		0.886	0.366 ^b	0.905	0.364 ^b
PZ 24		1.187	0.549 ^b	1.228	0.609 ^b
PZ 25		-0.372	0.264	-0.378	0.276
PZ 26		0.909	0.356 ^b	0.944	0.345 ^a
PZ 27		-0.978	0.514	-0.926	0.524
PZ 28		2.144	0.467 ^a	1.996	0.425 ^a
PZ 29		0.125	0.511	0.086	0.438
PZ 30		-1.446	0.495 ^a	-1.443	0.501 ^a
PZ 31		0.047	0.389	-0.082	0.381
PZ 32		1.241	0.161 ^a	1.198	0.156 ^a
PZ 33		0.991	0.160 ^a	0.853	0.154 ^a
PZ 34		0.340	0.159 ^b	0.329	0.155 ^b
PZ 35		0.752	0.187 ^a	0.751	0.180 ^a

Table 2: Elasticity and Welfare Estimates

Statistic	SANB	CANB
Elasticities		
Price		
L.B.	-0.91	-1.06
Estimate	-1.23	-1.29
U.B.	-1.54	-1.53
<i>Spread over mean</i>	<i>0.51</i>	<i>0.37</i>
Income		
L.B.	0.73	0.73
Estimate	1.14	0.88
U.B.	1.55	1.03
<i>Spread over mean</i>	<i>0.72</i>	<i>0.34</i>
CV		
Per trip (ZIP average)	32.77	31.07
<u>Total, all ZIPs</u>		
L.B.	1,587,808	1,599,931
Estimate	1,922,749	1,833,440
U.B.	2,510,312	2,324,822
<i>Spread over mean</i>	<i>0.48</i>	<i>0.40</i>

L.B. (U.B.) = lower bound (upper bound) for 95 percent confidence interval
 For computation of confidence intervals see footnote 10.

Table 3: Model Simulations with Different Population Shares and Income Variances

Model	Simulations		CVs	CVc	Ratio CVs / CVc	Ratio SOMs /SOMc
	Redution in share of white population	Increase in income variance (both groups)				
original	0%	0%	1,922,774	1,833,373	1.049	1.213
1a	10%	0%	1,927,901	1,828,717	1.054	1.256
1b	20%	0%	1,920,959	1,828,384	1.051	1.276
1c	30%	0%	1,942,608	1,830,410	1.061	1.305
1d	40%	0%	1,954,637	1,834,008	1.066	1.307
1e	50%	0%	1,962,951	1,839,131	1.067	1.330
original	0%	0%	1,922,774	1,833,373	1.049	1.213
2a	0%	10%	1,922,774	1,828,744	1.051	1.235
2b	0%	30%	1,922,774	1,819,740	1.057	1.091
2c	0%	50%	1,922,774	1,810,651	1.062	1.191
2d	0%	200%	1,922,774	1,789,074	1.075	1.346
2e	0%	300%	1,922,774	1,743,421	1.103	1.472
original	0%	0%	1,922,774	1,833,373	1.049	1.213
3a	20%	50%	1,920,959	1,801,645	1.066	1.377
3b	20%	100%	1,920,959	1,776,244	1.081	1.400
3c	20%	150%	1,920,959	1,750,379	1.097	1.577
3d	20%	200%	1,920,959	1,725,935	1.113	1.650

CVs = CV for simple model

CVc = CV for corrected model

SoMs = Spread of confidence interval of CV over mean of CV for simple model

SoMc = Spread of confidence interval of CV over mean of CV for corrected model

Table 4: Distribution of Mean-Variance Ratios for Income over ZIPs

Model / Area	No. of zips	pw	Mean-Variance Ratio for Income*									
			Average		Percentiles							
			w	nw	10		25		50		75	
					w	nw	w	nw	w	nw	w	nw
original	134	0.90	0.10	0.21	0.05	0.07	0.08	0.12	0.10	0.15	0.13	0.22
3a	134	0.72	0.07	0.14	0.03	0.05	0.05	0.08	0.07	0.10	0.09	0.14
3b	134	0.72	0.05	0.11	0.02	0.04	0.04	0.06	0.05	0.08	0.06	0.11
3c	134	0.72	0.04	0.08	0.02	0.03	0.03	0.05	0.04	0.06	0.05	0.09
3d	134	0.72	0.03	0.07	0.02	0.02	0.03	0.04	0.03	0.05	0.04	0.07
L.A.	273	0.63	0.13	0.19	0.03	0.06	0.05	0.10	0.09	0.17	0.16	0.27
Bay Area	228	0.72	0.10	0.17	0.03	0.07	0.05	0.10	0.08	0.14	0.13	0.21
California	1495	0.77	0.14	0.26	0.04	0.07	0.07	0.13	0.11	0.23	0.19	0.35
Chicago	202	0.78	0.13	0.19	0.04	0.07	0.07	0.10	0.12	0.18	0.17	0.24
Miami	78	0.75	0.14	0.26	0.03	0.095	0.05	0.16	0.12	0.26	0.20	0.34

pw = percentage of white population (ZIP average)

w = white, nw = non-white

*Income measured in \$000

