Nonparametric and Semi-Nonparametric Recreational Demand Analysis

Cooper, Joseph C.

Economic Research Service, USDA

29 June 1999

Nonparametric and Semi-Nonparametric Recreational Demand Analysis

By

Joseph C. Cooper

June 29, 1999

The author is with the Economic Research Service, US Department of Agriculture, 1800 M St., NW, Washington, DC, 20036. The author thanks the anonymous referees for their helpful suggestions. The views expressed herein are the author's and do not necessarily represent the views of the US Department of Agriculture.

jcooper@econ.ag.gov ; 202-694-5482 (voice) / 202-694-5778 (fax)

Nonparametric and Semi-Nonparametric Recreational Demand Analysis

Abstract

This paper addresses issues of specification testing for the travel cost method (TCM). Two nonparametric approaches to TCM analysis are presented. In addition, semi-nonparametric count models for TCM are developed. A numerical illustration is provided in which the three methods are applied to an actual TCM data set on waterfowl hunting and the results compared to those from a parametric analysis.

Keywords: bootstrap, count model, Fourier, kernel, nonparametric, PAVA, semi-nonparametric, travel cost method

Nonparametric and Semi-Nonparametric Recreational Demand Analysis

Nonparametric smoothing techniques represent a set of flexible tools for analyzing unknown regression relationships. In the nonparametric (NP) methods of interest in this paper, neither the error distribution nor the functional form of the regression relationship is pre-specified. Because of this flexibility, NP econometric methods are especially useful as specification checks on parametric methods. This paper explores the application of NP methods in nonmarket valuation. While the published nonmarket valuation literature has addressed the subject of nonparametric and semi-nonparametric estimation of discrete choice models for the contingent valuation method, it has not addressed application of NP techniques to the travel cost method (TCM).

Although NP techniques have a number of advantages over parametric methods, they are somewhat unwieldy because, compared to parametric methods, they cannot easily include a large number of explanatory variables. Alternatives to NP are expanding parameter space, or semi-nonparametric (SNP), methods, which are halfway between parametric and nonparametric inference procedures (e.g., Fenton and Gallant). SNP methods allow the researcher to reduce the potential for misspecification bias associated with parametric techniques, while at the same time accounting for explanatory variables more easily than NP. Given this, SNP would appear to make NP methods redundant. In practice, however, this is not the case. While SNP methods reduce the potential for misspecification bias, they do so at the cost of increased complexity over the parametric (PARA) approaches and require careful fitting to the data. Creel's (1997) simulation-based analysis sets the stage for application of SNP methods to TCM, but practical issues of estimation and interpretation of the results remain open. Hence, in addition to the presentation of two NP approaches to TCM estimation, this paper addresses issues in applying SNP methods to actual TCM

data and in interpreting the results. A numerical illustration is provided in which the three methods are applied to an actual TCM data set on waterfowl hunting and the results compared to those from a parametric analysis. The SNP and NP methods act as specification checks on each other. If the results are comparable, the researcher can have some reason to believe the benefit estimates are robust to the choice of estimator.

A Variable Partition Histogram Approach to Nonparametic Estimation

A traditionally popular nonparametric technique is the histogram, in which the data are divided into partitions on the basis of some smoothing parameter and cell frequencies estimated based on these partitions (see e.g., Delgado and Robinson for a survey of nonparametric techniques). The model proposed in this section falls into the general category of variable partition histogram approaches (VPHA), which allow a locally adaptive smoothing (Anderson; Van Ryzin). The specific form uses the Pool Adjacent Violators Approach (PAVA), which can be considered a variation on a VPHA approach in which each partition is of different width. The PAVA approach to generating empirical Bernoulli trials has been around a relatively long time (e.g., Ayer, Brunk, Ewing, Reid and Silverman; Turnbull) and has been applied to discrete choice contingent valuation data (Kristom; Haab and McConnell). For discrete choice data, the goal of PAVA is to insure that the estimated cumulative densities are strictly increasing in the bid offer, that is, F_j = prob($WTP \leq bid_j$) = $N_j/(N_j+Y_j)$, where N_j = the number of no responses to the bid offer bid_j and Y_j the number of yes responses to that bid. Given the initial J empirical properties, the PAVA algorithm takes cases where $F_{j+1} \leq F_j$ and pools F_{j+1} and F_j as $(N_j + N_{j+1})/(Y_j + N_j + Y_{j+1} + N_{j+1})$, where this pooled value is associated with bid_j , i.e., cell boundaries are bid_j and bid_{j+2}. The pooling is continued until the F's are strictly increasing in the bids. While this method is appealing on intuitive grounds, it also

yields the maximum likelihood estimates of the desired probabilities (Ayer, Brunk, Ewing, Reid and Silverman).

This section presents a modification of this approach for the trips variable, T, which is a count data variable censored at 0, and the travel cost variable C, where $T = f(C)$. The modification is analogous in the discrete choice contingent valuation case, where PAVA insures that the empirical density is monotonic with respect to the price variable. The TCM application is designed to insure that the integral under $f(C)$ is well-behaved, i.e., that the price-quantity points used to estimate CS satisfy the condition that T decreases when C increases ($\Delta C \Delta T < 0$). Given the original set of data points $\{T_1,C_1\}$, $\{T_2,C_2\},\ldots,\{T_k,C_k\}$, the goal of the PAVA algorithm is to reconstruct the data points to produce a monotonic function representing the minimum loss of data points, and hence, of the greatest accuracy in the CS estimate. [i]

Since without great loss of generality the density in most binary choice cases can be represented nonparametrically by sets of Bernoulli trials, PAVA for binary choice yields MLE estimates. However, because the usual count distributions such as the Poisson or the negative binomial each imply different decision making processes by the recreationist, it would appear that PAVA for count models cannot find an analogy in MLE. Hence, the appeal of PAVA for count data must be based on intuitive grounds. The following procedure is simple and does not require sophisticated programming, although a fast compiler is useful in bootstrap applications:

1) Sort $\{T_j,C_j), j = 1,\ldots,k$ in ascending order with respect to C_i, where C_1 = minimum observed round trip travel cost and C_k = maximum observed round trip travel cost, and where T_j are the sum of trips across the sample that are associated with that C_j.

2) Starting with $j = 1$, compare T_j and T_{j+1}.

3) If $T_{j+1} < T_j$, continue.

4) If $T_{j+1} \geq T_j$, then pool T_j and T_{j+1} into a cell whose boundaries are C_j and C_{j+2}, i.e., for pooled trip cell $T_j + T_{i+1}$, pooled trip cell cost is C_j. The required assumption is that users who paid C_{j+1} would be willing to pay C_j, which is reasonable for a normal good given that $C_{j+1} > C_j$.

5) The pooling loop is continued until the T_j*'s* are strictly decreasing in C_j. The pooled data pairs are denoted $\{T^*_j, C^*_j\}$, $j = 1,\ldots m$, where $m \leq k$.

In sum, the goal of steps 1 through 5 is to maximize the number of data points m subject to $\Delta T^* \Delta C^* < 0, \forall \{T^*_i, C^*_i\}, j = 1,\ldots, m$. Of course, numerous variables other than C, such as income, site quality, differences in sampling rates among origins, and unobserved variables, influence the number of trips. Hence, the stronger the relationship between $\boldsymbol{T}$ and $\boldsymbol{C}$, and the lower the influence of other variables on $\boldsymbol{T}$, the greater the number of cells, or histograms, in the set $\{\boldsymbol{T}^*, \boldsymbol{C}^*\}$.

Given the set of points $\{T^*_j, C^*_j\}$, $j = 1,\ldots, m$, the approximation of the integral CS = $\int_{C_0}^{C_m} f(C)dC$ is estimated using the trapezoidal rule as follows:

(1) $$CS_{PAVA} = \sum_{j=2}^{m} (C_j - C_{j-1}) T_j + 0.5 \sum_{j=2}^{m} (C_j - C_{j-1})(T_{j-1} - T_j) \quad .$$

where, to simplify the notation, $T = T^*$ and $C = C^*$ for the rest of this section. Since no condition except $\Delta C \Delta T < 0$ is set for the demand function drawn between the available points, the lower bound of CS_{PAVA} deletes the lower triangles, giving $CS^L_{PAVA} = \sum_{j=2}^{m} (C_j - C_{j-1}) T_j$. The upper bound CS estimator includes the upper triangle and is thus

$CS^{U}_{PAVA} = \sum_{j=2}^{m}\left(C_j - C_{j-1}\right)T_j + \sum_{j=2}^{m}\left(C_j - C_{j-1}\right)\left(T_{j-1} - T_j\right)$. Since ΔT converges on 1 in the limit, the limit in the difference between CS_{PAVA} on either bound is $\left|0.5\sum_{j=2}^{m}\left(C_j - C_{j-1}\right)\right|$.

A potentially important question involving the CS measure regards the extent of the impact of the truncation resulting from the exclusion of the approximation of $\int_{C_m}^{C_p} f(C)dC$ from the estimation of CS, where C_p is the choke price. In parametric TCM models, closed form solutions to the continuous integral usually imply a choke price that requires an extrapolation outside the range of the available travel cost data. For the NP, SNP, and PARA cases, given that this choke price is usually unobserved, truncating the upper limit of the CS measure at C_m may be preferable if a conservative measure is desired,. The data set used in the numerical illustration below is exceptional in that a $C_m \geq C_p$ is observed. Given that this price is less than 9% greater than the highest observed C associated with nonzero trips, its use as the choke price does not seem unreasonable.

Choke price issues aside, because the integral in the PAVA procedure is a line operator, it gives a consistent estimator of mean consumer surplus in the population: expected population consumer surplus is the expectation of consumer surplus conditional on a set of characteristics. PAVA switches the order of integration to give the integral over demand price unconditional on the characteristics.

Note that the nonparametric recreational demand model can be modified to handle discrete/continuous processes. It is common in the TCM literature to model the trip participation process in two stages. In the first stage the recreationist decides which of the $j = 1,\ldots,L$ alternative

sites to visit. In the second stage, she decides how many trips to take to the chosen site. One way to deal with the first stage is to use a multinomial logit approach to estimate probabilities for selection among the alternatives (Feather and Hellerstein; Feather, Hellerstein, and Tomasi). These probabilities are then used to construct the expected trip cost, E(C_g), $g = 1,\ldots, N$ individuals, and in multivariate cases, the expected values of the other explanatory variables. The expected trip costs can then be used along with the dependent variable T_g (total trips taken by individual g) in a continuous regression of a trip demand function or in a hurdle model application (Feather and Hellerstein). The results are used to calculate total CS for each individual. While the first stage has been only modeled parametrically in the literature, it is possible to model it nonparametrically in a simple procedure, at least for the case where the number of elemental alternatives L is the same for each individual. The nonparametric probability that individual g chooses elemental alternative l is

(2) $$P_{gl} = \Big/ \sum_{j=1}^{L} T_{gj}, \; g = 1,\ldots,N; j = 1,\ldots,L \quad .$$

Given estimates of P_{gl}, expected cost from the first stage RUM is $E(C_g) = \sum_{j=1}^{L} P_{gj} C_{gj}$. The data set used in the nonparametric estimation is then $\{T_g, \text{E}(C_g)\}$. The nonparametric estimate of P_{gl} can be used as a specification test on the estimate of P_{gl} from a multinomial regression.

Kernel Approach to Nonparametic TCM Estimation

While the PAVA approach in the previous section is simple to compute, the discontinuities inherent in the histograms do not allow estimation of derivatives (a minor concern here). In addition, asymptotic convergence of the PAVA to the true density may be slower than for the kernel approach, at least for smooth densities. The kernel is a continuous function that describes the shape

of a weight function, or local averaging procedure, that is used to represent a regression relationship $T_i = m(C_i)+\varepsilon_i$, $i = 1,\ldots n$. The kernel imposes greater form on the demand function than does the PAVA approach through the selection of a bandwidth, which controls the level of smoothing of the function $m(C_i)$. The higher the bandwidth, the higher the amount of smoothing.

The Koning (1996) implementation of the Nadaraya-Watson (NW) approach (Härdle; Silverman) is used for the kernel TCM regression. For the regression of the (Nx1) travel cost vector trips $\boldsymbol{T}$ on $\boldsymbol{C}$, the NW function is

$$\textbf{(3)} \qquad \hat{T}(x_j) = \frac{n^{-1}\sum_{i=1}^{n} T_i K\left(\frac{x_j - c_i}{h}\right)}{n^{-1}\sum_{i=1}^{n} K\left(\frac{x_j - c_i}{h}\right)}, \quad j = 1,\ldots,V$$

where $\hat{T}(x_j)$ is the predicted value of trips evaluated at some travel cost x_j, $K(.)$ is the kernel function, h is the bandwidth, and V is the number of increments of x_j. Given the estimated bandwidth, a smooth function for $\hat{T}(x_j)$ can be found by setting V arbitrarily large. In this application, V is set equal to 1000, and $\boldsymbol{x} = \{x_1, x_2,\ldots, x_v\}$ represents a sequence starting from minimum observed travel cost to maximum observed travel cost in equal increments. Given the function above, the trapezoidal estimate of CS is generated as described in the previous section but with the data points $\{\hat{T}(x_j), x_j\}, j = 1, \ldots V$. Note that unlike the PAVA method described in the previous section, the count data nature of trips is not accounted for in estimation. However, it is not clear how a kernel count model can be constructed without making any potentially incorrect assumptions about the nature of the underlying count data distribution.

Many different kernel functions are available and the choice of which one to use is a combination of art and science. The criterion used here is to choose the kernel function that

minimizes the sum of the differences between predicted and total actual trips, subject to $\hat{T}(x_j)$ being strictly decreasing in x_j. Of several $K(u)$ functions examined, the Epanechnikov kernel best mets this criterion. Letting $u = (x_j - c_i)/h$, this kernel is

(4) $$K(u) = \frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}u^2\right)I_{|u|\le\sqrt{5}} \quad ,$$

where indicator $I = 1$ if $|u| \le \sqrt{5}$, and $I = 0$ otherwise. The starting bandwidth is found according to Silverman, using the rule

(5) $$h^* = 0.9 * \min\left(s, \frac{w/1.34}{n^{0.2}}\right)$$

where s is the standard deviation of $\boldsymbol{C}$, w = ($C_{p=0.75}$ - $C_{P=0.25}$), and $C_{p=0.75}$ and $C_{P=0.25}$ are the values of C_i corresponding (or closest) to the 75^{th} and 25^{th} percentiles of an empirical distribution of the observed travel cost values. Cross-validation (e.g., Härdle, Nason) can be used to find the optimal value of h. In this application, half the data is randomly omitted, and equation 5 is applied to the remaining data to produce the starting value of h. Given this starting value, a search is performed over h to find the value that minimizes the squared prediction error of the omitted half of the data.[ii] For the results here, simulated annealing (Goffe, Ferrier, and Rodgers) is used to minimize $L(h) = \sum_{i=1}^{N/2}\left[T_i^O - \hat{T}_i^O(h)_i\right]^2$ to find the optimal value of h, where T_i^O and $\hat{T}_i^O(h)_i$ are the omitted trips and estimated trips, respectively.[iii] Alternatively, a grid search around the starting value of h can be used to find the omitted sample squared prediction error minimizing value.

Semi-Nonparametric and Parametric Count Data Approaches

Creel (1997) presents the first discussion of using the Fourier form for recreational demand analysis. The Fourier form itself will be discussed only briefly here, with a focus on the application of the Fourier to count data models and on issues of calculation of the benefit estimate based on this form. The Fourier functional form is the only functional form known to have Sobolev flexibility, which means that the difference between the model $h(\boldsymbol{x},\theta)$ and the true function $f(\boldsymbol{x})$ can be made arbitrarily small for any value of $\boldsymbol{x}$ as the sample size becomes large (Gallant, 1987).[iv] The Fourier flexible functional form $h_k(\boldsymbol{x},\theta)$, which attaches linear and quadratic terms to the Fourier to help decrease the number of terms needed to model nonperiodic functions (Gallant, 1982), is specified as

$$\textbf{(6)} \qquad h_k(\mathbf{x},\theta_k) = U_0 + b'\mathbf{x} + 0.5x'Cx + \sum_{\alpha=1}^{A}\left\{ \sum_{j=1}^{J}\left(v_{j\alpha}\cos\left[j\mathbf{k}'_\alpha s(\mathbf{x})\right] - w_{j\alpha}\sin\left[j\mathbf{k}'_\alpha s(\mathbf{x})\right]\right)\right\}$$

where $U_0 = u_0 + \sum_{\alpha=1}^{A}\{u_{0\alpha}\}$, and $C = \sum_{\alpha=1}^{A} u_{0\alpha}k_\alpha{}' k_\alpha$,

the (k-A-J) x 1 row vector $\boldsymbol{x}$ is the vector of all arguments of the utility difference model, k is the dimension of θ, A (the length) and J (the order) are positive integers, and $\boldsymbol{k}_\alpha$ are vectors of positive and negative integers that form indices in the conditioning variables, after shifting and scaling of x by s(x). For example, if $\boldsymbol{x}$ contains 3 variables and length = order = 1, then the $\boldsymbol{k}_\alpha$ vectors are (1,0,0), (0,1,0), and (0,0,1). The function s($\boldsymbol{x}$) prevents periodicity in the model and is a function that shifts and scales the variable to lie in an interval less than 2π (Gallant, 1982). Specifically, the variable is scaled by subtracting its minimum value, then dividing by the maximum value and then multiplying the resulting value by (2π - 0.00001), which produces a final scaled variable in the interval [0, 2π - 0.000001]. If a variable has only three unique values, then only the v or w

transformation may be performed. With two values, none of the transformations can be performed. A formal criterion for choosing A and J is not well established. Chalfant and Gallant suggest a rule of thumb that the dimension of $\theta = N^{2/3}$, but this may be high. Asymptotic theory calls for $\theta = N^{1/4}$ (Andrews), but Fenton and Gallant note that $\theta = N^{1/2}$ is likely to be more representative of actual practice.

There are a number of concerns about the use of a semi-nonparametric OLS or GLS estimator. First, the functional form $T_i = h_k(\mathbf{x}, \hat{\theta}) + e_i$ does not account for the censored and count nature of trips. Second, while $\left\| \mathcal{E}(T|x) - h_k(x, \hat{\theta}) \right\| \xrightarrow{a.s.} 0$ in principle, in applied work this limit may not be achievable. If it is not, the GLS estimator is not particularly desirable. Third, GLS does not guarantee that estimated trips are nonnegative or that the sum of estimated trips does not equal actual trips. Because it is frequently of interest in TCM analysis to predict the new level of trips given a change in a policy relevant variable, it is a useful property for the sum of estimated trips to equal actual trips as a baseline. Placing the Fourier form in a count data MLE framework can adequately address these three concerns.

Given that GLS does not account for the censored (at zero) and the integer nature of the dependent variable and may produce biased, and almost certainly, inefficient coefficient estimates, count data regression models such as the Poisson or negative binomial have received extensive attention in parametric TCM applications (e.g., Hellerstein; Creel and Loomis, 1990).[v] Count data models assume a distribution over Prob(Trips = T; T = 0,1,2,...), as in Prob(Trips=T)= $\text{Exp}(-\lambda)((\lambda^T)/T!)$ for the Poisson case. The single parameter Poisson distribution has a rather restrictive assumption that the mean, E(Trips), and variance, σ^2(Trips), of the distribution are equal. The two parameter negative binomial relaxes this assumption and allows the variance to

vary. By doing this, the negative binomial can control for overdispersion of the dependent variable. The usual functional form chosen for the Poisson parameter and the negative binomial mean in these parametric applications is $\lambda = E(TRIPS) = \exp(f(\boldsymbol{x},\beta))$, where $\boldsymbol{x}$ is the vector of explanatory variables, and where the usual functional form for $f(\boldsymbol{x},\beta)$ is $\boldsymbol{x'\beta}$. The exponential form eliminates the possibility of a negative λ. If sampling weights are used in the MLE, such as a population variable with aggregate trip data, then $\exp(f(\boldsymbol{x},\beta))$ is multiplied by the weight (Hellerstein). Unlike with a log-linear GLS application, among the useful properties of the Poisson and negative binomial in TCM applications are that zero trip values are allowed and, if a constant is included, the sum of predicted trips across all observations is equal to total actual trips. However, nonnegativity of estimated trips is not needed for consistent estimation of the CS estimate. Hence, the numerical illustration below includes OLS results using the SNP functional form.

For the Poisson case, and similarly for the negative binomial, the parameter λ can be made highly flexible by setting $\lambda = \exp[h_k(\mathbf{x},\theta)]$. However, while a likelihood function utilizing this specification for the density is highly flexible, it maybe not be fully SNP. A fully SNP specification is not obtained because $\lambda = \exp[h_k(\mathbf{x},\theta)]$ presumes that λ belongs to a class of linear exponential models and because there are remaining cross-moment restrictions, e.g., the variance and mean are still equal. A count approach that may be fully flexible is that of Cameron and Johansson (1997), although the authors do not attempt to demonstrate this feature. This approach utilizes a polynomial series expansion from a baseline Poisson density. The major downside of this approach is that the likelihood function for this nonlinear model can have multiple optima, which makes traditional gradient methods difficult to apply. Because of the difficulty of applying this model in practice, it is not extensively discussed here, although results are given briefly in footnote ix.

Given the coefficient estimates, the total consumer surplus is the sum of each individual's or representative consumer's CS, or

$$\textbf{(7)} \qquad CS_j = \sum_{i=1}^{n} \left(w_i * \int_{TC_{ij}}^{TC_{ij}^{\max}} \exp\left[h_k(\mathbf{x}, \theta_k)\right] dTC \right),$$

where $TC_{ij}^{\max}$ is the choke price (or to be conservative, maximum observed trip cost), i.e. the travel cost that drives trips from origin i to site j to zero, and w_i is some weighting index, if required by the data. Evaluating the CS of each individual separately, as opposed to evaluating average CS/person at the sample means and then multiplying by n persons, is necessary since the SNP form can be highly nonlinear. Because it is not practical to evaluate this function analytically, it is estimated empirically utilizing the "second stage" approach (e.g., Cooper and Loomis).[vi]

Taken individually, Fourier coefficients do not have an economic interpretation, and there is little point reporting them, especially if the number of parameters is large. To give the Fourier regression results an economic interpretation, they must be re-expressed in terms of the base variables. One possible way to add economic content is to generate graphs of the relationship between the benefit measure and the explanatory values numerically (Creel and Loomis, 1997). Another way is to evaluate $\partial \exp[h_k(x,\theta)] / \partial x$ and use this expression to form elasticities or flexibilities, noting that

$$\textbf{(8)} \qquad \partial h_k(\mathbf{x}, \theta_k) \Big/ \partial x = b + Cx + 2\sum_{\alpha=1}^{A} \left\{ \sum_{j=1}^{J} j\left(v_{j\alpha} \cos\left[j\mathbf{k}'_\alpha s(\mathbf{x})\right] + w_{j\alpha} \sin\left[j\mathbf{k}'_\alpha s(\mathbf{x})\right]\right) k_\alpha \right\}.$$

In the numerical illustration below, the regression results are expressed in terms of flexibilities evaluated at total trips and the mean price.

Numerical Illustration

The data set used for the numerical illustration covers 1989-1990 waterfowl hunting trips for the six national wildlife refuges (NWR's) in California's San Joaquin Valley (SJV) and consists of the whole population of hunters to the SJV NWR's for the 1989-1990 hunting season. The waterfowl hunting trip data for the San Joaquin Valley refuges were obtained from the on-site sign-up sheets the hunters are required to sign before they enter the hunting area. Since total participation per sampling unit is known, it is possible to include nonparticipants in the data set by including T_{kj}'s that are zero, where k = the hunter's county of origin and j = the refuge (for a total of 396 observations). Full details of the data are in Cooper and Loomis (1993).

The Pseudo-Maximum Likelihood (Gourieroux, Montfort, and Trognon) parametric and SNP count results, as well as OLS SNP regression results, are presented in the form of flexibilities in Table 1. Brief descriptions of the variables are in the footnotes to the table. The flexibilities presented in the table for variable $\boldsymbol{x}$ are constructed around base estimated total trips ($\Sigma_k \hat{T}$) and the mean of the variable $\boldsymbol{x}$, i.e., $\frac{\partial \Sigma_k \hat{T}}{\partial \bar{x}} \frac{\bar{x}}{\Sigma_k \hat{T}}$. Generating the travel cost variable E(C_k) from the results of the nonparametric first stage as discussed at the end of the first section is the same as C_{kj} for this data set, given that while the hunters come from all over the State, the refuges are clustered together.[vii] The base variables are used in linear form rather than in logarithmic form simply because the former fit the data better than the semi-log specification. In addition, for SNP regressions in which the base variables undergo a logarithmic transformation before undergoing the Fourier transformation, $\partial \hat{T} / \partial C$ was not strictly less than zero across the observed travel cost range (evaluated numerically). The relationship $\partial \hat{T} / \partial C < 0$ holds across the observed C for all the SNP regressions in table 1, and of course holds for the parametric model given its negative

coefficient on *C*. The adjusted R^2 and log-likelihood values show that the biggest gains in fit come with just the first series transformation, even though all the regressions are statistically different from each other on the basis of likelihood ratio tests. However, as is discussed further below, SP-III and SP-IV results are unstable due to high levels of multicollinearity. This is not surprising given that these two models have 25 and 45 coefficients each, respectively.[viii] The SNP specifications with order $J = 2$ (SP-II and SP-IV) produced flexibilities for H20DEL much smaller than the other models, although the flexibility was not significant for SP-II.

Because the Fourier form produces estimators that are inherently highly collinear, hypothesis tests derived from the covariance matrix are not trustworthy. Instead, confidence intervals for the regression results are produced with a bootstrap approach. Specifically, 1,000 simulated data sets are generated by randomly drawing (with replacement) observations from the real data set to create simulated data sets with the same sample size as the real data set. The bootstrap is the most general method for estimating confidence intervals since it uses the actual sample as the population for the choice sets. The confidence intervals presented in Table 1 are constructed from the regression results on each simulated data set and are of the bias corrected accelerated (BCa) type (Efron, 1987), which gives the bootstrap results an interpretation analogous to *t*-statistics by making the estimated confidence interval symmetric around the mean.

The α term in the table is a test of the equi-dispersion property of the Poisson form and was performed with the following regression (Cameron and Trivedi, 1990):

$$\mathbf{(9)} \qquad \left[T_i - \hat{\lambda}_i\right]^2 \Big/ \left[\hat{\lambda}_i - 1\right] \;=\; \alpha\hat{\lambda}_i + \varepsilon_i ,$$

where $\hat{\lambda}_i = \exp(h_k(\boldsymbol{x_i},\theta))$ for the SNP and $\exp(\boldsymbol{x_i}'b)$ for the parametric case. If Ho:$\alpha = 0$ is not rejected, then the Poisson condition E(*T*) = var(*T*) is not rejected. For the Poisson models in Table

1, α is both large and significant only in the parametric case. While it is significantly different from 0 at the 10% level for SP-II, given that α=0.04 for this model, the drawback of using Poisson over the negative binomial in this case cannot be great. The null hypothesis is not rejected for SP-I, III, and IV. The estimates of α for the Fourier Poisson models suggest that the Fourier form provides a substantial increase in flexibility over the parametric form, and even does so with only a few leading Fourier terms. However, cross-validation results below suggest that SNP-IV and perhaps SNP-III are highly over-fitted. If so, the dependent variable in equation 9 will be close to zero, and hence $\alpha = 0$ will hardly ever be rejected for SNP-III and IV.

The first two columns of Table 2 give the solution to the PAVA model from the first section. In the table, the value in the j^{th} row of "Trips" column is the total number of trips at travel cost C^* greater than or equal to C^*_j but less than C^*_{j+1}. As is evident from the first two columns, the relationship between trips and travel cost is highly nonlinear. If an econometric model predicted trips precisely, then these predicted values, when summed to correspond to the travel cost cells in the first column, should yield the same vector of trips in the second column. The value of this test over a single measure of fit such as the R^2 is that it can indicate across what travel cost range estimated trips may deviate from actual trips. This knowledge can be useful if tracking movements along the demand curve.. As Table 2 shows, the parametric count model ("Param") noticeably underpredicts trips for the lowest C^* cell and overpredicts trips for the next cell. The SNP models did notably better, although there is still room for improvement in the tail of the distribution.

Table 3 presents the consumer surplus (CS) point estimates and some descriptive statistics for each of the models. Confidence intervals (CIs), standard errors, and coefficients of variation for the CS measures are created using the bootstrap method described earlier. The BCa confidence intervals are shown for each point estimate. For cases where the empirical bootstrap CS distribution

is biased (relative to the point estimate), the CI from this distribution is shown as well. Presented on a per-trip basis, estimated CS ranges from $11.39 for SP-I to $32.26 for SP-II. A comparison among the SNP count model results in table 2 shows that the greatest percentage difference between the model predictions occurs in the tails of the distribution, where SP-I predicts one trip in the travel cost range $171.69 ≤ C < $188.38, while SP-II and SP-IV predict 25.4 and 27.5 trips, respectively, and where the true value is 19. Most likely, even though they represent only a small fraction of total trips, it is these differences in the tails that cause the SP-I and SP-II CS estimates to differ by almost a magnitude of 3.

As Table 3 shows, the empirical CI is biased enough in the cases of SP-III and IV that BCa lower bounds are negative. Because PAVA imposes the least structure on the data, the statistics for it should provide the lower bound on efficiency relative to the parametric and semi-nonparametric methods. Both SNP-III and IV have higher coefficients of variation than the PAVA point estimate, which suggests that these two models are overfitted and that collinearity makes the results unstable. The SNP-I OLS model produces CS results similar to the SNP-I count model. However, as Table 2 shows, the SNP-I OLS underpredicts trips by around 50%, so it is probably not as useful as the count models for predicting changes in trips associated with policy shocks. While they are not reported here, OLS versions of SNP-II, III, and IV were also tried. Again, the CS/trip values were relatively close to the count data versions, although trip predictions were off. That the CS/trip results for the count data SNP models are close to those for the fully flexible OLS SNP model speaks well for the flexibility of the count SNP model. [ix] For the kernel model, the estimated bandwidth *h* was 2.84.[x] Estimated CS/trip was within several cents of the PAVA estimate, and as would be expected, its coefficient of variation is lower than that for PAVA but higher than the parametric one.

Cross-validation can be used to better assess at what point over-fitting begins in the SNP count models. The cross-validation procedure involves removing one observation $\{T_i, X_i\}$, where T_i is trips and X_i is a vector of explanatory variables, doing the regression on the remaining data, and then using the estimated coefficients to predict $\hat{T}_i = f(X_i)$. This procedure is carried out for each $i = 1,\ldots,N$ observations to create the (N x1) vector of predictions $\hat{\boldsymbol{T}}$. Out-of sample squared prediction error is $\left(\boldsymbol{T} - \hat{\boldsymbol{T}}\right)'\left(\boldsymbol{T} - \hat{\boldsymbol{T}}\right)$. For the SNP-I, SNP-II, SNP-III, and SNP-IV count models, the square roots of this statistic are 0.891, 0.832, 0.830, and 3.792, respectively (normalizing on the parametric count error by setting it to 1 to make comparisons more obvious). The results show a gain in out-of-sample predictive power in moving from the parametric model through SNP-I to SNP-II. However, SNP-III has hardly any gain in this test over SNP-II, and SNP-IV does several times worse than the parametric count model. In fact, a look at the results in Table 3 shows that the coefficient of variation of the SNP-IV consumer surplus estimate is much higher than that for the PAVA model, which appears to corroborate with the results of this cross-validation test. Finally, the coefficient of variation for SNP-III and SNP-IV are larger than that for PAVA, which again suggests that SNP-II represents the maximum amount of fitting advisable for these data.

Table 4 presents the results of the hypothesis test Ho: $CS_i - CS_j = 0$, where the subscripts reference the approaches. Each of the 1,000 simulated data sets was saved, thereby allowing each data set to be analyzed by each approach and hence, allowing pairwise comparisons of the CS estimates across these approaches. The tests are based on the 90% CIs from the empirical interval as well as 90% CIs from the BCa transformation. In most cases, the null hypothesis is not rejected at the 5% level of significance. Of all the models, SNP-I is the one for which equality is most often rejected. The general conclusion drawn from the results from the various approaches is that for this

data set the true CS value is somewhere within the range of the values reported in Table 3. Note that the parametric-based value falls in this range.

Conclusions

Nonparametric (NP) methods are especially useful approaches for TCM modeling at the exploratory level. The two NP approaches used here gave quite similar benefit estimates, although the PAVA is less efficient than the kernel given that it imposes less structure on the data. The semi-nonparametric (SNP) models presented here provide both a means to deal with multiple explanatory variables more easily than with NP methods, and unlike NP, embed the parametric model, thereby providing a direct link to economic models. However, while the SNP approach should converge on the correct function as the sample becomes large, in practice, samples may not be large enough to ensure convergence. The estimation results show that with real world data, the SNP approach requires careful modeling, as is demonstrated in the increasing instability in the econometric results for the SNP models as the number of parameters increases. For the data set examined here using the SNP, only a few series transformation terms were needed to improve the fit substantially, and maintain a balance in the tradeoff between efficiency and bias.

The PAVA approach imposes the fewest assumptions on the data and the parametric count model imposes the most. The empirical comparisons of the consumer surplus estimates indicate that the parametric count model appears trustworthy in estimating this value. At the same time, one should note the relatively short length of the confidence interval around the parametric estimate compared with those of the NP and SNP models. This relatively precise fit makes the parametric results appear better than what the data themselves suggest. An implication of this result for econometric analysis is that, in the absence of specification tests of parametric models against NP or

SNP models, the researcher should be explicit in noting that precision of the parametric point estimate is attributable more to the imposition of specific assumptions used in estimation, such as functional form, than would be explained by the data themselves.

With finite samples, the regression results for the higher parameter SNP models show that flexible regression tools can produce dubious results. Therefore, a risk-reducing research strategy is to apply several different flexible tools to analyzing the data set, as is done here. If equality of the parametric to the NP and SNP results is rejected in each case, then the "truth" of the parametric model is certainly questionable.

References

Anderson, T. "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," *Multivariate Analysis I*, P. Krishnaiah, ed., 1965.

Andrews, D. "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Model." *Econometrica* 59(March 1991): 307-345.

Ayer, M., H Brunk, G. Ewing, W. Reid, and E. Silverman. "An Empirical Distribution Function for Sampling with Incomplete Information." *Ann. Math. and Statist.* 26(1955):641-647.

Cameron, A., and P. Johansson. "Count Data Regression Using Series Expansions: with Applications." *J. Appl. Econometrics* 12(May-June,1997):203-223.

Chalfant, J. and R. Gallant, "Estimating Substitution Elasticities with the Fourier Cost Function." *J. Econometrics* 28(May 1985):205-222.

Cooper, J., and J. Loomis. "Testing Whether Waterfowl Hunting Benefits Increase with Greater Water Deliveries to Wetlands." *Environ. and Resour. Econ.* 3(December 1993):545-561.

Creel, M. "Welfare Estimation Using the Fourier Form: Simulation Evidence for the Recreation Demand Case. " *Rev. Econ. and Statist.* 78(February 1997):88-94.

Creel, M. and J. Loomis. "Semi-nonparametric Distribution-Free Dichotomous Choice Contingent Valuation." *J. Environ. Econ. and Manage.* 32(March 1997):341-358.

Creel, M. and J. Loomis. "Theoretical and Empirical Advantages of Truncated Count Data Estimators for Analysis of Deer Hunting in California." *Amer. J. Agr. Econ.* 72(May 1990):434-45.

Delgado, M. "Nonparametric and Semi-parametric Methods for Economics Research." *J. Econ. Surveys* 6(1992):201-249.

Efron, B. “Better Bootstrap Confidence Intervals.” *J. Amer. Statist. Assoc.* 82(1987):171-185.

Feather, P. and D. Hellerstein. “Calibrating Benefit Function Transfer to Assess the Conservation Reserve Program.” *Amer. J. Agr. Econ.* 79(February 1997):151-162.

Feather, P., D. Hellerstein, and T. Tomasi. “A Discrete-Count Model of Recreational Demand.” *J. Environ. Econ. and Manage.* 29 (September 1993):214-27.

Fenton, V. and Gallant, A. “Qualitative and Asymptotic Performance of SNP Density Estimators,” *J. Econometrics* 74 (1996):77-118.

Gallant, A. "Unbiased Determination of Production Technologies. *J. Econometrics* 20(1982):285-323.

Gallant, A. “Identification and Consistency in Semi-Nonparametric Regression,” in *Advances in Econometrics Vol. I*, Truman F. Bewley, ed., Cambridge University Press, New York, 1987, pp.145-169.

Gallant, A. and G. Tauchen. “Seminonparametric Estimation of Conditionally Constrained Heterogeneous Processes: Asset Pricing Applications.” *Econometrica* 57(September 1989):1091-1120.

Gallant, A. and G. Souza. “On the Asymptotic Normality of Fourier Flexible Form Estimates.” *J. Econometrics* 50(December 1991):329-353.

Goffe, W., G. Ferrier, and J. Rodgers. “Global Optimization of Statistical functions with simulated Annealing.” *J. Econometrics* 60(January 1994):65-99.

Gourieroux, C., A. Montfort, and A. Trognon. “Pseudo Maximum Likelihood Methods: Applications.” *Econometrica*. 52(May 1984):701-20.

Greene, W. *Limdep: User's Manual and Reference Guide, Version 6.0.* Bellport NY: Econometric Software, Inc., 1992.

Haab, T. and K. McConnell. "Referendum Models and Negative Willingness to Pay: Alternative Solutions." *J. Environ. Econ. and Manage.* 32(February 1997):251-270.

Hellerstein, D. "Using Count Data Models in Travel Cost Analysis with Aggregate Data." *Amer. J. Agr. Econ.* 73(August 1991):860-866.

Härdle, W. *Applied Nonparametric Regression.* Cambridge, UK: Cambridge University Press,1990.

Koning, R. H. "Kernel: a Gauss Library for Kernel Estimation." http://www.rhkoning.com/gauss/ June, 1999.

Kristom, B. "A Non-parametric Approach to the Estimation of Welfare Measures in Discrete Response Valuation Studies." *Land Econ.* 66 (May 1990):135-139.

Nason, G.P. "Wavelet Shrinkage Using Cross-Validation." *J. Royal Statist. Soc., Series B* 58(1996): 463-479.

Van Ryzin, J. "A Histogram Method of Density Estimation." *Communications in Statistics* 2(1982): 493-506.

Silverman, B.W. *Density Estimation*. London: Chapman and Hall, 1986.

Turnbull, B. "The Empirical Distribution Function with Arbitarily Grouped, Censored, and Truncated Data." *J. Royal Statist. Soc., Series B* 38(1976):290-295.

Table 1. Parametric and Semi-Nonparametric (SNP) Regression Results.

Results Presented in Form of Flexibilities (90% BCa Confidence Intervals Shown in Parentheses)

	PML Poisson Count Model					
Variable	Parametric	SNP-I	SNP-II	SNP-III	SNP-IV	SNP-I (OLS)
TC(RT)	-3.10	-6.98	-2.64	-5.54	-3.31	-6.63
	(-4.24, -1.99)	(-9.2, -4.82)	(-4.24, -1.09)	(-7.41, -3.73)	(-5.55, -1.15)	(-8.50, -4.81)
INC89	0.16	-3.36	2.34	-1.18	14.02	-10.21
	(-0.74, 1.08)	(-8.0, 1.12)	(-0.74, 20.54)	(-6.1, 3.66)	(-12.67, 41.83)	(-27.76, 6.56)
PSBAG	0.29	0.15	0.12	0.06	0.14	0.22
	(0.04, 0.55)	(-0.05, 0.36)	(0.04, 0.36)	(-0.18, 0.32)	(-0.22, 0.50)	(-0.48, 0.96)
H20DEL	1.59	0.91	-130.60	1.30	-133.58	-3.72
	(0.67, 2.54)	(-0.10, 1.97)	(-76.45, 0.67)	(0.25, 2.40)	(-191.3, -77.73)	(-7.93, -0.34)
#coef	5	13	21	25	45	13
α	1.34 (10.3)	0.09 (0.12)	0.06 (1.51)	0.04 (1.99)	0.02 (0.76)	--

Table 1. – continued

LnL	-8754	-4063	-2987	-3617	-2491	--
Adj R2	0.511	0.899	0.949	0.933	0.977	0.313

Notes:

BCa 90% confidence interval applies the bias corrected accelerated approach (Efron) to 1000 bootstrap runs.

SNP-I: order = length = 1; SNP-II; order = 2; length = 1; SNP-III: order = 1; length = 2; SNP-IV: order=length =2.

OLS-I: ordinary least squares with SNP transformation of data, order = length = 1.

#Coef is number is coefficients in the regression; The coefficient α is a test of overdispersion.

h^2 = 1 - RSS/TSS, where RSS is residual sum of squares and TSS is total sum of squares. For OLS (with a constant term), h^2 equals ESS/TSS, though this is not necessarily the case for nonlinear models (Peterson and Stynes, 1986).

The variables are (see Cooper and Loomis, 1993, for details): TC(RT) is round trip travel cost, including variable vehicle operating cost and the opportunity cost of travel time, from county *i* to refuge *j*. H20DEL is water deliveries to refuge *j*. INCOME is average income in county *i*. PSBAG is the price of substitutes index, weighted by bag at the destination to reduce multicollinearity.

Table 2. Comparison of Actual, Count Data, and OLS Estimated Trips Aggregated According to the Trip Cost Cells From the PAVA Model

		Estimated trips -- Count Models					
TC(RT)	Trips	Param	SNP-I	SNP-II	SNP-III	SNP-IV	SNP-I (OLS)
1.07	8727	6632.9	8458.2	8622.7	8762.0	8750.4	9656.3
42.13	2496	4701.7	2697.8	2494.9	2375.9	2405.3	381.9
55.73	2419	2369.9	2188.4	2409.4	2310.7	2411.2	215.6
117.62	334	211.5	417.5	364.1	394.7	397.8	39.2
130.38	237	319.3	476.6	356.8	369.2	277.8	2.2
137.34	92	87.9	144.3	84.8	135.5	76.8	2.4
157.76	41	58.6	9.7	31.8	36.3	38.1	6.3
168.90	22	1.8	0.8	1.3	5.3	4.7	0.0
171.68	19	12.1	1.0	25.4	5.2	27.5	0.1
188.38	6	0.4	0.0	1.2	0.2	0.8	0.0
206.02	2	0.4	0.0	1.7	0.0	1.7	0.0
225.50	0	0.0	0.0	0.0	0.0	0.0	0.0

Table 3. Consumer Surplus per Trip Estimates and Related Statistics

	PAVA	Kernel	SNP-I (OLS)
CS/Trip	24.72	24.87	12.38
BCa 90%	(16.67 , 58.49)	(10.78 , 39.45)	(8.15 , 16.72)
Empirical 90%	(3.02 , 47.25)	(17.62 , 45.84)	(9.93 , 17.72)
Standard Error	13.44	8.71	2.61
Coef. Variation	0.41	0.31	0.20
	Count Data Models		
	Parametric	SNP-I	SNP-II
CS/Trip	26.09	11.39	32.26
BCa 90%	(18.19 , 34.24)	(6.61 , 16.33)	(17.5 , 52.57)
Empirical 90%	--	--	--
Standard Error	4.88	2.95	11.49
Coef. Variation	0.19	0.23	0.37
	SNP-III	SNP-IV	
CS/Trip	14.59	26.37	
BCa 90%	(-0.86 , 32.17)	(-49929 , 88570)	
Empirical 90%	(10.27 , 20.63)	(15.04 , 431.85)	
Standard Error	10.02	40177	
Coef. Variation	0.66	15.05	

Table 4. Hypothesis tests comparing benefit estimates from each model

(1 = equality not accepted; 0 = equality accepted)

	PAVA	Kernel	Param	SNP-I	SNP-II	SNP-III	SNP-IV
PAVA	--	0	0	1	0	0	0
Kernel	0	--	0	1	0	1	0
Param	0	0	--	1	0	0	0
SNP-I	0	0	1	--	1	0	0
SNP-II	0	0	0	1	--	1	0
SNP-III	0	0	0	0	0	--	0
SNP-IV	0	0	0	0	0	0	--

Notes:

Lower Triangle is the hypothesis test based on BCa transformations of the bootstrap results.

Upper Triangle is the hypothesis tests based on empirical confidence intervals of the bootstrap results.

SNP results are for the count models.

Endnotes

[i] The subscript k refers to the maximum number of unique values of C, where $k = n$, the sample size. For each C_j, j=1,…,k, T_j is the total number of trips associated with C_j. Hence, with this horizontal summation of individual demand points, $\{T_j,C_j\}$ represent points along a total demand function.

[ii] Cross-validation can be of the "leave one out" or "leave half out" variety (Nason). The former is slow in this case and may be overkill in this application.

[iii] Simulated annealing differs from gradient methods in that it permits movements that increase the objective function in addition to ones that decrease it (we are minimizing here), so that the program can move out of local optimum if need be. A version of the annealing routine by E. Tsionas was used and is available from http://gurukul.ucc.american.edu/econ/gaussres/optimize/optimize.htm.

[iv] Note that asymptotically, SNP results can be considered nonparametric (Gallant and Tauchen), making the distinction between the two somewhat artificial, at least in theory if not in practice.

[v] Given the flexibility of the Fourier, the fact that GLS does not account for the censored and integer nature of the dependent variable in its Fourier application is probability not of major concern in and of itself, but the lack of a nonnegativety constraint is.

[vi] The SNP, Kernel, and PAVA computer programs written in Gauss are available directly from the author at jcooper@econ.ag.gov. Similar programs for CVM are available as well.

[vii] To maintain degrees of freedom, the nonparametric first stage model is not used to create expected values of the other variables for the regressions utilizing the data set $\{T_k, C_k, \boldsymbol{X}_{k,}\}$, where $\boldsymbol{X_k}$ is the matrix of explanatory variables, and k = 1,…,N denotes individuals (or sampling aggregations, as is the case here). If the goal of the study is to recover properly the total consumer surplus of each individual across all the choices, then this data set should be used instead of $\{T_{kj}, C_{kj}, \boldsymbol{X}_{kj,}\}$, k =

1,…N, j = 1,…,L. However, in this study the econometric comparisons are of interest, and reliability of the parametric and semi-parametric results are increased by maintaining the larger sample size.

[viii] The experiment here suggests that inclusion of the quadratic terms as well as the Fourier series terms in the regressions had little impact on the CS estimates. Hence, they are left out for the sake of efficiency.

[ix] While not reported in the tables, the Cameron and Johansson model, which takes a polynomial series expansion around a baseline Poisson density, was also tried. Because traditional gradient methods are difficult to use with this model, I used the method of annealing, as recommended by the authors. The CS/trip estimates for a PP2 (Poisson polynomial of order 2) and PP3 are $26.52 and $21.13, respectively. However, convergence could not be achieved with PP1, and for both PP2 and PP3 the covariance matrix was not positive definite.

[x] Not accounting for population differences across the centroids (observation i) in this dataset with trips aggregated by centroid can result in heteroskedasticity – predicted trips for high (low) population centriods will be higher (lower) than expected from their travel costs alone. To account for the impact of population-related heteroskedascity in the kernel regression, C_i is weighted by the ratio of total sample population size to population size at that observation. The weights are multiplied by the same size and divided by the sum of the weights so that the sum of the weights across the observations is the sample size (e.g., Greene, 1992). The weight is

$$w_i = n\left(\sum_{i=1}^{n} pop_i \Big/ pop_i\right) \Big/ \sum_{i=1}^{n}\left(\sum_{i=1}^{n} pop_i \Big/ pop_i\right).$$