CHAPTER 5- TECHNICAL NOTE 29: MEASUREMENT OF SUPPLY USING NATIONAL INTERVIEW DATA ON PARTICIPATION IN OUTDOOR RECREATION ACTIVITIES

By J. Beaman, S. Smith, and Y. Kim

ABSTRACT

A recurring theme in outdoor recreation research is the need to develop models to forecast future use levels for outdoor recreation -facilities. However, most aggregate demand models describe how demand is related to socio-economic characteristics of populations. Consequently such models are calibrated from data on observed participation. Their weakness is that the "observed use levels" are not only a function of demand but also of supply.

This paper presents an approach to defining the effects of supply on the rate of participation in an activity. The effect of supply is measured by finding out the average errors in predictions of participation in activities that took place in given areas. It was expected that by using a model that did not take into account the distribution of supply there would, on the average, be over-predictions of the amount of participation in areas with "low supply" for an activity and vice versa for areas with a high level of supply. In other words residuals from a regression in which supply was not considered were expected to be significantly positive on the average in some areas and significantly negative in other areas.

Statistically very significant residuals were found and do appear to be associated with availability of supply. These were found using 1972 Secretary of State leisure time use data for 25,500 independent individuals, people from different households.

One result of the study was to show that a significant effect of supply could not be determined from 1972 CORDS National Survey data. Reasons for this result are discussed in a special appendix to the paper where sample size formulae and the accuracy of supply factors are discussed. It is concluded that to get results that are good enough to study how supply factors are determined by the supply on the ground 800,000 interviews would be necessary from 60 geographic areas. However, it is also indicated that a well designed analysis with interviews from a few areas (about 8) could be successful with only about 80,000 interviews.

PURPOSE

The purpose of this note is to present research work on the development of a quantitative, behaviour-based measure of the effects of supply of facilities for participation in some activity on participation in those activities.

INTRODUCTION

A central concern in planning for outdoor recreation involves the development of models to forecast future use levels for facilities that are built. Most forecasting models used in recreation research have been destination models based on demand factors describing socio-economic characteristics of target populations; cost (usually as distance); attractiveness or quality of existing and proposed sites; and occasionally measures of the numbers of alternative facilities. Such models are calibrated from observed participation figures (see Chapters III, VII and IX of this volume in which relevant notes are included). There have been origin models developed but in these it is almost invariably only socio-economic variables that are considered. However, a critical weakness of this approach is that the models developed do not reflect supply. They make participation only a function of demand factors. The 1969 study by Cicchetti, Seneca and Davidson (or see Cicchetti 1973 ch. 3) is one of the rare cases where a "supply factor" has been incorporated into an analysis so that the level of supply does influence the amount of participation predicted.

The reason that other quantitative studies reported on in this volume such as those on the use of ANOVA and AID models (TN 12, TN 20) do not relate people’s, total behaviour to the amount of supply around them is not simple. A major problem is the problem encountered by Cicchetti, Seneca and Davidson. That was that information on supply which was at all adequate for use in a model was rarely available. The research reported on in TN 16 and the Facility Inventory, reported on in Volume show that CORDS research has focused on how to inventory supply that is on the ground. The lack of success of the Canadian Outdoor Recreation Demand Study Facility Inventory (as reported on in Volume III) meant that inventory data were not available for building the kind of model proposed by Cicchetti, Seneca and Davidson (1969), and suggested by Knetsch for the CORDS demand analysis (see Chapter I).

In the way of introduction one final point seems important. A traditional approach to modelling supply effects on participation has been to intuitively decide on some physical resource as a limiting factor for participation in a given activity at a given location. For example, one may choose as a measure of "camping supply" the number of campsites in a ring around a population centre, divided by their distance from that center, summed over a number of concentric rings. This number can then be used in a model to adjust for probable expansion or contraction of supply. A concern raised by this approach, however, is just why one measure of supply is used rather than another one. In modelling it is frequently difficult to say, a priori, which model is better among a multitude of alternatives. A researcher may even be able to decide which class of models is more appropriate from a behavioural point of view, but one still has to choose between many possible permutations of a given model. An empiricist may not share this problem of choosing the "best" model since a decision can be made on a straightforward and objective basis--the best model is the one which most closely fits observed data. This choice-making tactic has a major weakness for planning purposes--the model chosen on an empirical basis is not necessarily generalizable to other situations of recreation supply-demand systems. This difficulty of deciding on a most appropriate model is also discussed for related topics in TN 3, 12, 19, 20, 27, 30 and 35.

The supply factor computation described in this paper has the desirable quality of avoiding the difficulty of having to select among several formulations. Further, once a supply factor is obtained it may be possible to explain this empirically derived factor by using conventional supply measure formulae. For those regions in which there is evidence of supply effects on participation it may be possible to examine the levels of facility and resource development logically or intuitively associated with those activities exhibiting possible supply effects. Ideally this analysis would lead to the further identification of what natural and man-made resources (in terms of quantity, quality, location, costs, advertising, management policies and so on) define supply for a given activity.

HISTORY OF THE PRESENT WORKOF DEVELOPING A MEASURE OF SUPPLY ON PARTICIPATION:INTRODUCTION TO THE METHODOLOGY

Work on a version of this paper began in 1973. This work began using 1972 National Survey information broken down so that observations were associated with 75 different geographical areas in Canada (on these data and geographic codings on these data see Volume III). As work on various Canadian Outdoor Recreation Demand Study projects proceeded, it became clear that considerations of the way that people reacted to alternative facilities made it important to derive a measure of how people perceived supply. It was recognized that this should not be some ad hoc formulation that a researcher dreamed up such as the Equation 1 supply measure for an origin area for some activity

(1) Supply Measure = Σ [(attractiveness of a site)**A/ (distance from the origin to the site)**B}
WHERE the sum is over all of an origin's sites for a given activity.

The general kind of supply measure just defined is called an alternative factor elsewhere (TN 1, TN 3) and used in other ways in other research (TN 5, TN 11). Regardless it suggests that a person perceives the amount of supply based on how far away it is and how attractive it is. The issue of concern here is: "Is the function that was chosen correct?" An approach that may ultimately lead to an answer is not to start with a function but to start with assumptions and proceed to derive values which reflect peoples, response to supply. Having such values or supply measures for each of a number of geographic locations allows one to work backwards to find the function that explains the measures in terms of what is on the ground (see TN 4 for Cesario's approach to explaining attractiveness values).

The conceptual strategy adopted was to measure the effects of supply in several steps:

(1) Assume that supply was homogeneously distributed in all regions in Canada;
(2) Use an existing model (such as the one reported on in TN 12) to make estimates of the amount of participation in some activity in each of these regions;
(3) Obtain actual participation levels and compare these to the predictions;
(4) Examine the differences between actual and predicted participation figures to see if this difference shows that there is a statistically significant difference between observation and predictions that may reasonably be accounted for by the level of supply in the different geographical areas considered.

The kind of supply factor being described can be defined as follows:

SF(i) = Supply factor area i
= Σ (observation - predictions)/N(i)

WHERE the sum is over all persons from i for whom there are data in the survey used and N(i) is the total number of such people in i.

Given the assumptions just cited a negative sum of residuals and thus a negative average value of the residuals could be expected for a region when observed use levels are less than the predicted. Because this average is negative even taking into account the socio-economic factors, it is reasonable to refer to this average value of residuals as a supply factor for the given region for the activity considered subject to some qualifications introduced subsequently. If an area has abundant facilities for the activity considered, if it exceeds a national average and these are being used, it may be expected that observed participation will exceed the national average and in this case the average value of a residual will be positive. So, the supply factor defined by the average value of residuals for the geographic area will be positive.

There are some points of criticism that can be raised about the rather naive formulation just introduced, but these are only raised in the discussion section of the paper. Now further background on the actual estimation attempts that were carried out in an effort to obtain supply measures and to verify that these were statistically significant is introduced. This is because there has been a history of problems in developing what may already sound to the reader as a very simple measure of how people respond to supply.

Initially, when predictions were made about peoples, behaviour in various outdoor activities both using the model indicated in TN 12 and using a model generated by the Michigan AID program (see TN 20) the supply effects determined were highly variable. This was true even though they were based on interview information from 4,000 people, and it was true both in trying to explain participation and non-participation and in trying to explain the amount of participation. The reason that the amount of participation was finally disregarded in the analysis is that it is known that in Canadian Outdoor Recreation Demand Study National Survey data people's reported volume of participation (1) may reflect enthusiasm for an activity which destroys the relationship between the total volume of supply in an area and participation; (2) the amount of participation reported is known to be very unreliable and (3) it seemed more plausible when this research was being carried out to assess the amount of supply perceived to be available by assessing general perceptions of this supply than to let the analysis be dominated by a fairly high level of participation by some people in some geographic areas whereas in other geographic areas there may be a quite general but lower level of participation by the people.

Regardless, the initial analysis attempt involved developing both frequency of participation based supply measures and participation or non-participation based supply measures. The results produced were disappointing. After much computer programming and sorting out of residuals for a large number of activities, it was concluded that significant supply factors had been estimated. However, these supply factors had been estimated on the assumption that the CORDS National Survey information resulted from interviews in 4,000 different households to get the information on 4,000 different individuals. Subsequent checking of this matter showed that the commercial company that carried out the interviews, interviewed more than one person in a large number of residences. This introduced a correlation between participation figures that resulted in a significant relation between supply and participation becoming an insignificant relationship. Many hours of work carrying out rather sophisticated statistical tests to show a result significant at the 5% level was simply lost!

The only option that offered any promise for measuring the effect of supply on participation was to use a data set in which the kind of intercorrelation just cited did not occur. Ideally a data set which was much larger so that there would be a much better chance of observing a supply effect. One problem with 4,000 observations and 75 geographic areas was that there is not much information about each area and consequently the variation of each supply factor tends to be large in comparison to anything that is to be explained (see the Appendix).

Fortunately in 1972, data were collected by Secretary of State, Citizenship Branch, on participation in leisure activities. These data included information on participation in some outdoor activities and thus these data offered good potential for analysis. Unfortunately, the 52,000 interviews carried out in this study were not carried out in different residences but it was possible to select about 25,000 interviews from the Secretary of State data set on the basis of a random selection of single interviews per household. This gave a reduced tape on which all of the interviews could be considered to be independent. So after supply factor research work had continued for over two years a new data set was adopted.

These data from the Secretary of State tape which were actually used in the analysis are indicated in Figures 1 and 2. The sixty-two geographic areas used are relatively homogeneous economic areas of the different Canadian provinces as defined by Statistics Canada for the 1972 Labour Force Survey.

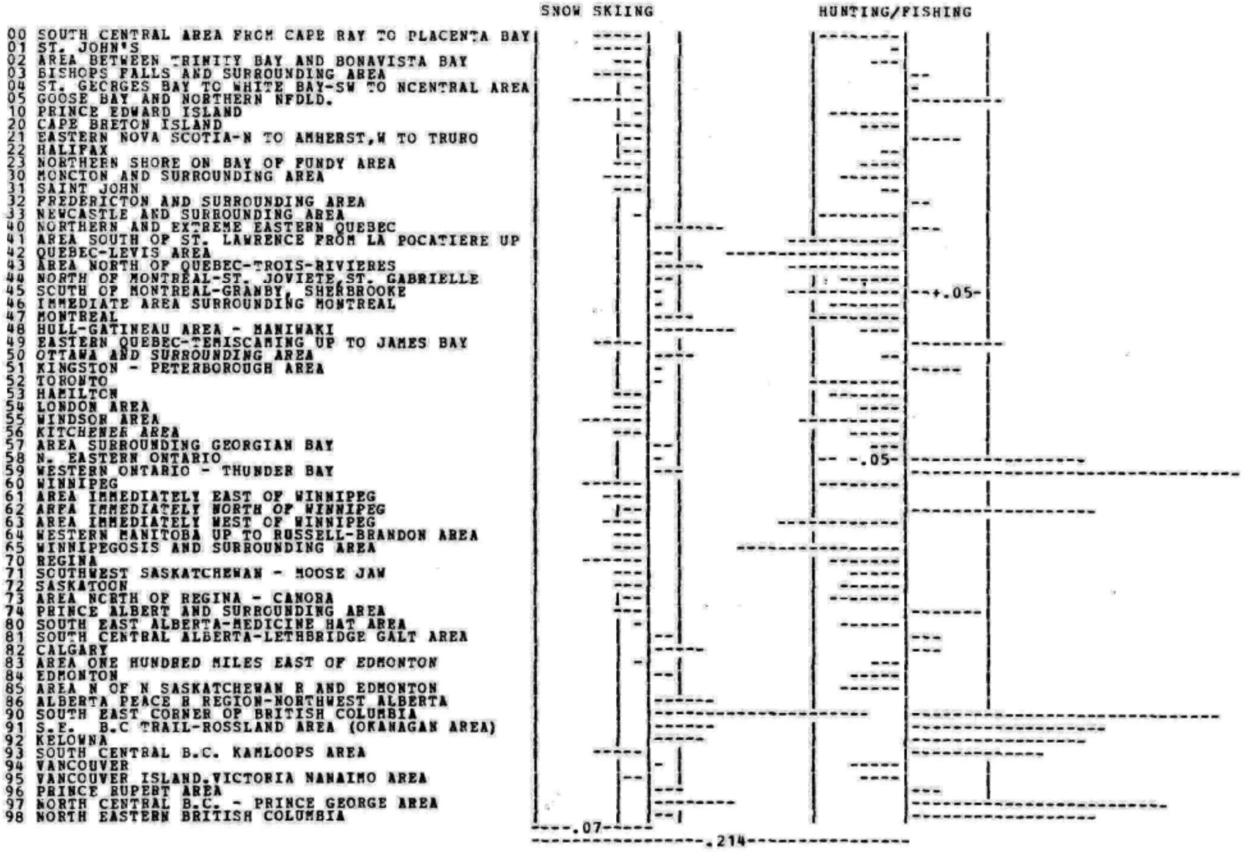
Now, there is a final point before proceeding to presenting results. As research work on this project went ahead, it was recognized that estimation need not always be carried out by producing residuals and using a special program to process these in a number of steps. The reader who is familiar with the theory behind least squares estimation will know that if one sets up a model such as the one indicated by the equation following then one can estimate the coefficients in the model by simply using a regression analysis program:

$$\begin{array}{c}\text{Probability of person in region r}\\ \text{with given socio-economic}\\ \text{characteristics participating}\end{array} = M+ \begin{array}{c}\text{(sum of socio-}\\ \text{economic effects)}\end{array} + \text{(region r suppy effect)} + \text{error}$$

WHERE (M + . . .) is as defined in TN 12.

The regional supply effect is the regression coefficient that "explains" how, on the average, the region differs from an average region after correcting for socio-economic effects. In actually carrying out computations to determine if the supply factors indicated in the equation shown above were significant, first a regression was carried out in which no supply factors were included. This allows one to determine the variance explained when supply is considered not to be important. It also allow one to know how much variance remains to be explained. Then, when 62 additional parameters were introduced, the 62 geographic areas (indicated in Figure 1) were considered and supply factors computed for them. To assess statistical significance of these supply factors, one need only ask whether the introduction of these parameters has resulted in more variance being explained than would be expected to be explained by chance (for a similar test see TN 27). In this particular case the introduction of the supply factors as indicated subsequently should have explained less than 5 thousandths of the variance that remained to be explained after socio-economic factors were considered. Yet far more variance than this was explained.

FIGURE 1: Supply Factors Plotted for the Origins to Which They apply



* The scales on which the snow skiing and hunting and/or fishing supply factors are plotted differ slightly. For snow skiing, there is about .0017 units per mm.; for hunting and/or fishing, there is about .0015 units per mm. This difference occurs because of the way that the computer program used scale results. Given the slight discrepancy between the scales and the fact that the supply factors should not be compared, no common scale was established even though both plots were related to a common baseline.

TABLE 1:Socio-Economic Effects B(i,j), by ANOVA and Sums of Squares Explained by AID and ANOVA for Snow Skiing and Hunting and Fishing

Part 1 – Snow skiing		ANOVA Model with socio-economic effects		ANOVA Model with socio-economic and origin, effects		**AI**D socio-economic analysis	
Variable	Level	Beta	St Dev	Beta	St Dev		
VAR005		-.01783	.00289	-.01604	.00330	See Table 2	
VAR005	2	.00092	.00519	.00008	.00587	for AID	
VAR005	3	-.00487	.00581	-.00130	.00634	results	
VAR005	4	.02178	.00300	.01726	.00395		
VAR006	I	-.00766	.00227	-.00759	.00226		
VAR006	2	.00766	.00227	.00759	.00226		
VAR007	1	.02881	.00559	.02884	.00557		
VAR007	2	.01986	.00312	.01951	.00311		
VAR007	3	.00080	.00310	-.00224	.00309		
VAR 007	4	-.04768	.00324	-.04610	.00323		
VAR059	1	.03727	.00379	.03573	.00377		
VAR059	2	-.01860	.00374	-.01762	.00373		
VAR059	3	-.01867	.00384	-.01811	.00382		
VAR060	4	.01409	.00502	.01264	.00500		
V AR060	2	.00846	.00559	--00735	.00557		
V AR060	3	-.00786	.00531	.00632	.00529		
VAR060	4	.00224	.00416	.00103	.00404		
HSHLD	1	.00783	.00443	.00889	.00443		
FISH1.D	2	.00001	.00263	.00076	.00263		
HSHLD	3	.00372	.00282	.00367	.00281		
HSHLD	4	-.01156	.00450	-.01181	.00452		
General	Mean	.06996	(Y.05344)	.07145	(Y .5344)	Y .05344	$\ddot{Y}^{=}$.05344
Total Sum of Squares			1279			1279	
Explained Sum of Squares			51.74	69.83			
R^2			.0404			.0546	
Residual Sum of Squares (RSS)			1228.0			1209.91	
RSS Degrees of Freedom			25280			25219	
Regression Degrees of Freedom			15			76	
F Statistics			71.01			19.15	
Standard error of residuals			.220			.219	

To be very specific the regression analyses were carried out as described because it was more convenient to carry out regressions, with only socio-economic variables and with both socio-economic variables and supply effects than to follow the more complicated analysis procedure described subsequently for the AID analysis. This is because with one run of the regression program made results were generated.

Part 2- Hunting & Fishing		Unweighted - only SE effects		Unweighted with suppy effects		Weighted ANOVA - only SE effects		Weighted ANOVA with suppy effects		AID
Variable	Level	Beta	St Dev	Beta	St Dev	Beta	St Dev	Beta	St Dev	
VAR005	1	.02493	.00504	.01669	.00436	.01074	.00574	.00965	.00435	See Table 2
VAR005	2	-.00165	.00906	-.00857	.00723	-.01677	.01020	-.01225	.00731	for AID
VAR005	3	.00973	.01013	.00179	.00861	.01189	.01102	-.00029	.00797	results
VAR005	4	-.03302	.00523	-.00991	.00406	-.00586	.00687	.00290	.00492	
VAR006	1	-.12065	00395	-.11499	.00374	-.12007	.00393	-.10929	.00375	
VA R006	2	.12065	.00395	.11499	.00374	,12007	.00393	.10929	.00375	
VAR007	1	.06577	.00975	.04255	.00930	.06939	.00967	.03363	.00865	
VAR007	2	.02608	.00544	.01865	.00548	.02559	.00540	.01439	.00501	
VAR007	3	-.00364	.00540	-.00127	.00519	.00082	.00537	.00260	.00489	
VAR007	4	-.08821	.00564	-.05992	.00487	-.08580	.00561	-.05008	.00473	
VAR059	1	-.01380	.00660	.01362	.00435	-.01193	.00655	-.00941	.00434	
VAR059	2	.01648	.00653	.02260	.00547	.01543	.00648	.02022	.00545	
VAR059	3	-.00267	.00670	-.00898	.00404	-.00350	.00664	-.01080	00406	
VAR060	1	.02386	.00875	.01110	.00678	.02039	.00869	.00919	.00663	
VAR060	2	-.03110	.00975	-.01763	.00741	.02655	.00967	-.01665	.00700	
VAR060	3	-.00430	.00926	.00642	.00559	-.00462	.00919	.00414	.00559	
VAR060	4	.01154	.00708	.00010	.00505	.01078	.00703	.00333	00500	
HSHLD	1	.00549	.00773	.00619	.00510	.00058	.00770	.00303	.00511	
HSHLD	2	.00953	.00459	.00791	.00334	.00905	.00456	.00580	.00421	
HSHLD	3	-.00203	.00493	-.00615	.00561	-.00148	.00489	-.00276	.00342	
HSHLD	4	- .01299	.00785	-.00796	.00558	-.00814	.00785	-.00607	.00535	
General Mean		.06996	(Y .2079)	.2135	(Y .2079)	.24206	(Y .2079)	.22848	(Y .2079)	Y = .2079
Total Sum of Squares			4265.07		33428.4		4165.07		36018.8	
Explained Sum of Squares			428.72		3778.90		510.40		4259.16	
R Squared			.1005		.113		.1225		.1182	
Residual Sum of Squares			3736.36		296493		3654.68		31759.7	
Residual Degrees of Freedom			25280		25280		25219		25219	
Regression Degrees of Freedom			15		15		76		76	
F Statistics			193.378		214.799		46.342		44.5	
Standard Error of Residual			.384		1.083		.381		1.122	

Note: These results are based on 25,296 cases selected from data reported on in 'A Leisure Study, Canada 1972' by Carol Kirsh, Brian Dixon and Michael Bond. The report was produced for the Department of the Secretary of State, Canada.

However, one caution is appropriate at this point. The procedure just outlined is perfectly adequate for determining if there are significant supply effects. But, there is the remote possibility that the supply factors computed may have variance in common with the socio-economic effects which are included in the model. If one wanted to get supply factors that were absolutely not "polluted" by socio-economic effects, it would be necessary to make a first regression run and produce the residuals from this regression run. Then the supply factor would be computed directly from the residuals (by a regression or some less costly procedure). The point is that when a regression is carried out and then the residuals are subjected to further analysis, the variance that is found in the residuals is variance which is not in common with the variance that has previously been explained by the socio-economic variables used in the original regression. This occurs because as is well known for least-squares estimation, the residuals are uncorrelated with, or in other words, orthogonal to the parameter space. Regardless, examination of Table 2 shown clearly that whether or not supply factor effects were computed the socio-economic effects had roughly the same value, e.g. in row 1, one sees that the effects for snow skiing of being at level 1 of VAR005 are -.0178 and -.0160 which differ by less than .003, which is the standard duration for both coefficients. The standard deviation of the differences is $(0.003^2+0.003^2)^{1/2} \approx 0.004$. Therefore, the difference is under one standard deviation.

FIGURE 2: Secretary of State Data Variables and Their Levels As Used in Deriving Supply Factors

Variables		Levels of Variables	
Identification	Description	Identification	Description
VAR005	Population of Sampling Area	1	Under 20,000
		2	20,000-49,999
		3	50,000-99,999
		4	100,000 and Over
VAR006	Gender of Respondent	1	Female
		2	Male
VAR007	Age of Respondent	1	Under 16
		2	17-24
		3	25-44
		4	45 and Over
VAR059	Marital Status	1	Single
		2	Married
		3	Other
VAR060	Relation to the Head of Household	1	Spouse
		2	Son of Daughter
		3	Relative, Boarder, Employee, Other
		4	Head
Household*	Number of Individuals In Household Interviewed	1	1
		2	2
		3	3 or 4
		4	5 and Over
Origin	Origin of Respondent	1-62	For description see Figure 2
Fishing/ Hunting	Participation in Fishing and or Hunting	0	Non participant
		1	Participant
Snowskiing	Participation in Snowskiing	0	Non participant
		1	Participant

*Note: This variable was not on the original tape. It was generated by a program that selected 1 individual from each household and at the same time counted the number of individuals interviewed in the household.

If one carries out an analysis of the relation between socio-economic variables and participation using the Michigan AID program, when the AID program has been used and the residuals from the AID analysis have been stored, one need only carry out an analysis of the analysis of the residuals based on the following equation to see if significant supply factors can be estimated:

Residual for a person in a region r = (regional supply effect for r) + error
$= SF(r) + \varepsilon$

In actual fact when AID was used in estimating supply effects and the residual differences between observations and predictions were stored so that these could be analysed using the regression program it was realized that the regression program need not be used. Because only 1 supply effect applied to each individual and the residuals already had a mean of zero overall a special purpose "regression program" was written which calculated the supply factors, presented in Table 3 (the original Table 3 is only partially reproduced because only some values are needed in this paper). This program took advantage of the fact that each supply factor was only a weighted sum of residuals. Once the "residual tape" which had region as an identifier was sorted by region observations were read one after the other with supply factors being produced as sub totals and total sums of squares explained being accumulated throughout the run. The supply factors program was also used to compute sums of squares explained and sums squares that could not be explained by supply factors (see Tables 1, 2, and 3). It then became possible as described before to compare the sums of squares explained with those that could be expected to be explained by chance.

TABLE 2: Summary Results on Aid Analyses

	SNOW SKIING			HUNTING AND FISHING	
Group	No. of People in the group	Mean for the group	Group	No. of people in the group	Mean for the group
20	411	0.241	16	27	0.556
21	1019	0.184	28	11	0.545
6	178	0.174	20	335	0.540
2	513	0.140	21	2023	0.453
16	355	0.130	15	2008	0.388
15	1077	0.103	18	2167	0.315
26	544	0.096	29	2455	0.285
23	1976	0.088	19	826	0.276
27	2253	0.073	9	1565	0.235
24	68o	0.063	23	400	0.213
9	558	0.045	7	367	0.166
28	4121	0.043	26	3740	0.149
19	1259	0.021	24	5995	0.088
11	10190	0.014	27	1088	0.070
29	162	0.012	25	2289	0.033

But, what is being alluded to becomes more clear when the actual results of analysis are presented, for now it should suffice to indicate that the actual regressions performed produce a weighted sum of residuals. If, for example, socio-economic effects are estimated and eliminated then supply effects are computed or if the socio-economic variables are included supply factors are determined by a formula which indicates that a supply effect is a weighted sum of residuals:
Supply effect for r= (Σ W(i) residual(i))/(Σ W(i)) region r for an activity

The weighting of residuals to reflect the variance in observations is mentioned in TN 6 and 20 and discussed in detail in the Cicchetti and Smith article which appears as an appendix to this volume. The weight is based n estimating a probability *p* so, as for flipping a coin or rolling a dice, standard deviation when given *p* is *p*(1-*p*).

RESULTS

The results of this analysis are presented in Tables 1 through 5 and Figure 1. Tables 3 and 4 permit the reader to see in certain respects how different ANOVA and AID analyses compare. In particular Table 2 presents regression coefficients which are laid out in such a way that one can see that, whether or not supply factors are incorporated into the snow skiing model or the hunting and/or fishing participation model, the effects of socio-economic variables remain the same. In particular if one looks at the first and third columns of Table 1 one finds the effect for level 1 of VAROO6 (being female) indicated by a coefficient of -.00766 with a relatively small standard of deviation of .002. This value occurred when only socio-economic effects were considered in developing a model. This coefficient only shifted in value to a minus .00759 when origin effects were introduced. This shift is well within one standard deviation. A similarity of coefficients is even more striking when one looks at effects for different levels of VAR007. These shows the effects of age on participation. For example the coefficients for the youngest age group is .0288 whether or not origin effects have been incorporated into the snow skiing model.

TABLE 3: ESTIMATED SUPPLY FACTORS (SF) SNOW SKIING, SF's & Their SD's

Origin	From ANOVA		Based on AID Unweighted		Residuals Weighted		Unweighted From		ANOVA Weighted		Based on Unweighted AID		Residuals Weighted	
	SF	SD	SF	SD	SF	SD	SF	SD	SF	SD	SF	SD	SF	SD
1	-.034	.016	-.036	.016	-.023	.012	-.014	.028	-.046	.028	-.006	.028	-.029	.025
2	-.036	.010	-.034	.011	-.012	.007	.013	.017	-.003	.018	.037	.016	.015	.013
3	-.019	.018	-.012	.015	-.008	.012	-.029	.031	-.021	.028	-.018	.033	.000	.028
4	-.037	.014	-.038	.014	-.024	.011	.028	.025	.013 •	.030	.038	.023	.013	.021
5	-.011	.014	-.008	.013	-.005	.010	.028	.025	.004	.028	.030	.022	.000	.021
6	-.055	.048	-.048	.051	-.039	.045	.063	.084	.059	.113	.059	.084	.023	.081
7	-.007	.012	-.007	.011	-.007	.008	-.027	.022	-.043	.019	-.023	.023	-.017	.019
8	-.024	.012	-.023	.011	-.010	.008	-.003	.021	-.022	.019	-.013	.022	-.016	.018
9	-.013	.010	-.012	.008	-.009	.006	.052	.017	.037	.026	.070	.014	.052	.012
10	-.015	.007	-.016	.006	-.009	.005	.004	.012	-.014	.013	.015	.011	.006	.010
11	-.016	.012	-.016	.010	-.011	.008	-.009	.021	-.026	.021	.002	.021	-.004	.018
12	-.026	.012	-.023	.011	-.017	.008	-.046	.021	-.044	.015	-.030	.022	-.015	.018
13	-.017	.010	-.017	.010	-.012	.007	.007	.017	-.017	.018	.037	.015	.010	.013
14	-.002	.010	-.003	.009	-.004	.007	.003	.018	.014	.021	.037	.017	.015	.014
15	-.009	.014	-.006	.013	-.001	.010	-.067	.024	-.050	.018	-.059	.029	-.029	.024
16	.047	.015	.047	.014	.008	.001	.024	.026	.019	.029	.023	.026	.021	.022
17	-.002	.014	-.003	.013	-.005	.010	-.087	.024	-.072	.014	-.079	.028	-.063	.024
18	.012	.013	.012	.013	-.005	.010	-.018	.026	-.102	.012	-.170	.036	-.102	.029
19	.034	.009	.034	.009	.014	.006	-.009	.016	-.068	.009	-.062	.017	-.045	.014
20	.015	.012	.014	.011	.001	.009	-.069	.021	-.038	.015	-.065	.025	-.043	.021
21	.010	.011	.008	.010	.012	.007	-.107	.018	-.066	.106	-.107	.023	-.067	.018
22	.001	.010	.010	.009	-.003	.007	-.109	.018	-.043	.001	-.101	.022	-.059	.018
23	.029	.006	.026	.005	.019	.004	-.102	.011	-.058	.008	-.056	.009	-.030	.007
24	.057	.017	.058	.016	.015	.012	-.004	.029	-.022	.030	.024	.029	.027	.025
25	-.037	.033	-.037	.022	-.021	.017	.055	.038	.060	.049	.064	.038	.067	.032
26	.031	.008	.029	.008	.025	.005	-.009	.014	-.011	.014	.023	.013	.018	.011
27	.007	.011	.007	.009	.001	.007	.034	.019	.032	.021	.026	.017	.035	.014

In Tables 1 and 5, it will be noticed that the F - test for the regressions are given. All of the models developed have F - test values which are very highly significant. At the same time all of the models have R^2 value which are low enough that many people would consider the models quite unacceptable (on this matter see particularly TN 35 and comments in TN 6). In reality the low R^2 values only reflect the problem involved in predicting the actions of an individual. As shown in TN 6 when predictions are made for many people the accuracy of results on what percent will participate in an activity can be plus or minus a few percent using CORDS data.

TABLE 4: CORRELATION COEFFICIENTS BETWEEN AID AND ANOVA BASED SUPPLY FACTOR ESTIMATES AND AVERAGE (RMS) SIZE OF SUPPLY FACTORS[1]

SNOW SKIING RESULTS[3] CORRELATIONS

	Average size of Supply Factor(SF)[2]	ANOVA SF's	AID SF's unweighted	AID SF's weighted
ANOVA SF's	.0393 (.018)	1.0	.99 (.98)	.93 (.99)
AID SF's unweighted	.0390 (.016)	.99 (.98)	1.0	.93 (.99)
AID SF's weighted	.0205 (.012)	.93 (.99)	.93 (.99)	1.0

HUNTING AND FISHING RESULTS[3] CORRELATIONS

	Average Size of Supply Factor		ANOVA SF's unweighted		ANOVA SF's weighted		AID SF's unweighted		AID SF's weighted	
ANOVA SF's unweighted	.0660	(.030)		1.0	.92	(.72)	.92	(.97)	.88	(.98)
ANOVA SF's weighted	.0666	(.037)	.92	(.72)	1.0		.89	(.71)	.91	(.72)
AID SF's unweighted	.0708	(.030)	.92	(.97)	.89	(.71)	1.0		.94	(.99)
AID SF's	.0542	(.026)	.88	(.98)	.91	(.72)	.94	(.99)	1.0	

1. For actual values of the various supply factors see Table 3.
2. The average size is the root mean square average size. It gives one some idea of how much on the average a supply factor correction to a prediction will be. The number in brackets gives one an idea of the average size of the standard deviation in a supply factor though as one can see from Table 3, standard deviations tend to be somewhat proportional in size to the supply factor to which they relate.
3. The numbers in brackets following the supply factor correlation coefficients are weighted correlation for the standard deviations in supply factors. The generally high value reflects in part that standard deviation depends on sample.size. The correlation of about .7 for the weighted hunting-fishing supply factors standard deviations simply shows the influence of introducing the variants corrections for observations rather than weighting each observation by one as is done in AID and ANOVA analyses. The .9 correlation of the supply factors from the weighted model with other supply factors shows that weighting only has a "drastic" effect on coefficient standard deviations, not on coefficients. (This holds when the model is appropriate to the data.)

Those readers who are familiar with the AID program will understand why the coefficients from the AID analysis that explain behavior in terms of socio-economic characteristics are not included in Table 2. Those who are not familiar with the AID program may wish to refer to CORDS TN 4 or 27 or to a reference on the AID program (see Reference 20). When the AID Program is used people are grouped into relatively homogeneous clusters based on complicated collections of socio-economic characteristics which serve to identify people for whom an appropriate estimate of their probability of participating in an activity is the group mean for the terminal group they are assigned to based on the socio-economic characteristics they have. The important information on AID analysis which can be displayed in a summary form is the size of the group which are identified by the AID Program and the various means for these groups. Actually these means are the prediction that would be made if one wished to estimate the probability participation for the people in a particular group.

From Table 2 it can be seen that there was a mean participation for the whole universe of .05344 (the average value of the unweighted dependent variable) for snow skiing.

TABLE 5: STATISTICAL TESTS FOR THE SIGNIFICANCE OF THE SUPPLY FACTORS

	Snow Skiing			Hunting and Fishing			
	ANOVA	AID Results		Unweighted ANOVA	Weighted ANOVA	AID Results	
	Results with/ without 0rigins	Unweighted	Weighted	Results with/ without Origins	Results with/ without Origins	Unweighted	Weighted
TSS	1228	1211.29	25257.4	3736.36	35964.101	3699.24	25244.2
ESS#1	18	17.5663	231.698	82.68	4204.461	78.5557	369.496
R(2)(%) ESS/TSS	1.467	1.450	0.917	2.212	11.69	2.124	1.464
RSS#2	1210	1193.72	25025.70	3653.68	31759.64	3620.68	24874.7
F-Statistics#3	6.04	5.98	3.77	9.18	5.37	8.82	6.02

1. D.F. for ESS = 62.
2. D.F. for RSS = 25,296 - 1 - (15 + 62) = 25,218.
3. An F(62; 25,218) greater than 1.90 has a probability less than 0.001 of occurring.

Note: Total sum of squares (TSS) and explained sum of squares (ESS) for the weighted ANOVA results for Hunting and Fishing were based on TSS and ESS in Table I using the following formulae:

TSS (weighted) TSS (weighted ANOVA results of Socio-Economic effects and origin effects) - k ESS (unweighted ANOVA results of Socio-Economic effects only)

ESS (weighted) = ESS (weighted ANOVA results of Socio-Economic and origin effects) - k ESS (unweighted ANOVA results of Socio-Economic effects only)

where k = TSS (unweighted ANOVA results of Socio-Economic effects only) TSS (unweighted Socio-Economic and origin effects)

From Table 3, one can see that there was one group of 411 people identified for which a most appropriate prediction of their probability of participation was .24. Also, around 2,500 people out of the 25,000 for whom the model was developed are in groups 20, 21, 6, 2 and 16 for which the mean for the group is more than twice the mean of the dependent variable. One notes a similar pattern for the AID results on hunting and fishing. However for hunting, it is much more striking because based on the mean of the dependent variable of about .2 one suspects that a person has one chance in five of being a hunter yet in the AID Analysis certain socio-economic characteristics are sufficient to identify people to the point that one can reasonably assign these people one chance in two of participating in hunting. At the other extreme with respect to hunting and snow skiing one can see that there are large numbers of people in groups with group means that are less than half of the average participation rate for the population, below .10 and below .027 respectively.

The results on the AID analysis that have been presented are obviously unsatisfactory to one who is interested in why a certain group has a certain group mean. However, to present the detailed results on how different groups were defined in terms of socio-economic variables and to comment on this takes one into matters which are quite irrelevant to the general theme of this paper. For that reason further AID results are not included here. The data used in this analysis are available (for AID or ANOVA) should someone wish to pursue the matter of why certain AID clusters occur (see contact notes for University of Waterloo at the start of CORDS web pages).

Table 4 is where one for the first time sees the actual supply factors which were calculated in this analysis. One can manually compare the supply factors to see that the supply factors computed when AID was used to explain participation in terms of socio-economic variables agrees with supply factors computed when ANOVA was used. However, Table 5 provides weighted correlation coefficients which will probably give one a better idea, at least a quantitative idea, of the degree of agreement between the different supply factors. Standard deviations in the supply factors were used as weights in computing the correlation coefficients. The correlation coefficients between the 3 sets of snowing skiing supply factors and the four sets of factors for hunting and fishing being in the general order of magnitude of .9 indicated that for each activity any one of the supply factors explains around 80% of the variance in the other supply factors. And, given the standard deviations in the supply factors this level of explanation is all that can be expected. So, the results of the analysis appear very good in one respect. That is that the supply factors computed from the AID and the ANOVA analyses agree very well. One could say that this agreement shows something about the reliability of the supply factors when they are determined in a similar way.

But one may be concerned about the validity of these supply factors. Does the magnitude of the different supply factors show something that would be expected in terms of what is known about the distribution of supply for snow skiing and for hunting and fishing in Canada? Now, rather than meticulously examining the numbers in Table 4 in trying to answer this question one can look at Figure 1 in which origin information and supply factors are displayed together. There one sees that people hunt and fish and also snow ski in the mountains of Western Canada. This is no surprise. One sees that skiing is below the national average in the Canadian prairies where supply conditions are poor and it is well above the national average in the areas around Quebec City and Montreal where there is an abundance of ski slopes. Again one sees large negative supply factors in Southern Ontario and the Maritimes which are areas which offer few opportunities for snow skiing because of the weather. It is left for the reader to examine the effects in detail to decide if certain details correspond with what experience suggests. However in this examination the reader should keep in mind that some of the supply factors do have rather large variances. For example, the difference in the supply factors for Regina and Saskatoon, Saskatchewan could be taken as reflecting the existence of the Blackstrap "Mountain" ski development near Saskatoon while Regina has no comparable supply but this is a tenuous conclusion because it is based on rather weak evidence. Still Thunder Bay and other special skiing areas do regularly appear to have the supply factor which one might expect.

Given the statistical significance of many of the supply factors and given the way that the relationship between the supply factors and the areas with which they are associated makes sense in terms of what most Canadians know about supply for the activities considered, there may appear to be little reason to give statistical tests for the significance of supply factors. However, such significance tests are presented in Table 6. There one sees that all of the supply factors computed are highly significant in terms of the sums of squares which they explain above and beyond sum of the squares that are explained by socio-economic variables. However, the reader may find Table 6 particularly distressing in terms of the R^2 values presented there. The feeling may be that it was bad enough to see R^2 values of .05 or .10 for the socio-economic effect models but the values for supply factors are absolutely too low to be meaningful. Actually, they are not too low to be meaningful. The fact that supply factors explain another 1.4% of the variance in participation after socio-economic variables have explained 5% really shows that supply factors are too important to ignore if models are to be developed that are to be at all accurate in explaining people's participation in activities in particular areas of Canada.

One can see the importance of the supply factors, even though they only increase R^2 by 20 to 30%, in terms of the following argument. Consider that R^2 is increased from Δ by an amount $r\Delta$ where r is say .25 (as it was when supply factors were introduced into the unweighted model to explain participation in hunting and/or fishing). It may be noted that for the relation defined above to hold the contributions of the supply factor to R^2 are not defined by r but by the square root of r. This is the case because the sums of the squares of these effects, which are orthogonal to the socio-economic effects must add up to $r\Delta$ and if the socio-economic factors have an amplitude X and the supply factor an amplitude SQRT(r)X the sum of square is X^2 $(rX)^2$. So, in practical terms if r is .2 or .3 this implies that supply factors are on the average 40 to 60% of the magnitude of socio-economic factors. If effects of this size are ignored in making predictions for particular geographic areas one can see that substantial errors could result. However, one should recognize that the mean participation level is not considered when R^2 is computed. considering them to be unimportant compared to socio-economic variables, other matters should be considered in determining their overall importance in making correct predictions. Consequently, this matter is returned to in the discussion section of this paper.

DISCUSSION

The results presented in the last section are quite definitive. They leave no doubt that there are significant supply effects that can be calculated for different regions of Canada that help explain the amount of participation in various activities. However, as is made clear in the appendix of this note, extremely large sample sizes are required to measure these supply effects accurately enough so that they can be subjected to secondary analysis focusing on how actual supply "on the ground" is related to perceived supply. Relating the regional supply effects to actual "on the ground" supplies or measures of perceived supply can, and should be, the goal of future work.

This future work, however, should proceed with caution. As was pointed out earlier, original supply factor research involved information on the incidence of participation (yes or no). Some measure of supply based on the frequency of participation might also be developed. But if it is, it is plausible that two different analyses, one based on incidence and one based on frequency, will result in different supply factors. If so, the implication is that different models should be used by planners according to the measures they are trying to predict. This is what is implied in the Cicchitti, Senecca and Davidson (1969; see also Cicchetti 1973, ch. 3 and TN 34).

Another matter of importance regarding the models is that many of the socio-economic variables considered as causing or shaping participation are not truly causal. More precisely, some are antecedents of other causal variables. It is only later in this volume in the review of Chapter VII and in Chapter IX of this volume that any comment is made about the need for a better and more precise understanding of causal variables in modelling. For example, age is a partial antecedent to education, which is in turn a partial antecedent to income. The result is that the socio-economic variables included in a regression model are not truly independent of each other. Participation is a dynamic phenomenon and the type of analysis presented in this paper is static or cross-sectional. At best, the solutions presented in this paper are approximations of an equilibrium condition and approximations are where research must start. For the purposes of this paper and from an empirical viewpoint, the measurement of supply effects does not depend on the internal structural validity of the relation between socio-economic variables implied by the models used. The critical point for the work here is to simply include certain variables. In other words, there is a need for more work on identifying and measuring the causal variables of participation, but this is not a crucial issue in this paper.

A related issue concerning the development and refinement of a model is that the supply factors developed here were achieved under the assumption that all people have knowledge of the supply of facilities for given activities. Future computations could be based on interviews with only those people who really do know something about the supply of facilities. Or, more generally, a model could be formulated in such a way that the response of a given person about his participation is weighted according to the amount of knowledge he has about the supply of facilities for the activity under consideration. A model based on this type of formulation might from an aggregate view adequately combine considerations both of the incidence of and the frequency of participation.

The supply-knowledge component should probably be entered into a projection model as a multiplicative component rather than as an additive one. The basis for this is that if the user perception of supply is zero, there will be no participation. This effect is missed if the supply-knowledge component is additive. If the reader is going to get into the kind of considerations just raised, he should also be aware- of an issue discussed by Cicchetti, Seneca and Davidson (1969; Cicchetti 1973, ch. 3). These authors suggested that people who participate at different rates should actually be studied by using separate models. A similar conclusion is implied in the cluster analysis presented at length in CORD TN 10 and commented on in TN s 3, 32 and 37. In these Notes the issue of participation is not simply related to a single activity, but rather the question is raised (Reference 00???????????) whether or not groups of activities should be considered.

As indicated above, the basic computational strategy used here utilizes actual participation data to get a measure of how human behaviour deviates from a model which assumes supply is uniformly distributed. It is necessary that the other assumptions behind the supply factor computation should be made explicit since they concern the interpretation of the supply factor derived here. First it is assumed that there are no regional effects resulting from a unique cultural milieu. Specifically, it is assumed that cultural factors within any of the 62 regions considered in this paper do not result in a change in participation (by modifying the expression of causal socio-economic variables) that might be confused with supply variations. Similarly, it is assumed that any socio-economic variable not included in the analysis either has no effect on recreation participation or it averages out within each of the regions. Also it is assumed that supply factors do not explicitly measure the effect of weather in a given year. For example, a cold summer in the Maritimes and good weather in the Prairie Provinces in the year data were collected could result in a supply factor showing that the supply of opportunities was much larger in the Prairies than it was in the Maritimes because of the effect of weather in modifying average participation patterns. Broadly interpreted, this would be true. The opportunities for a satisfactory recreation experience would be greater in those areas with good weather, but, this interpretation of supply should not be confused with "what is on the ground".This problem of interpretation must be kept in mind continually, and either considerations of the effects of weather should be taken into account or evidence should be presented which suggests that weather had negligible distorting effects with respect to a given activity.

Finally and probably of most practical importance is that though supply factors do not add much to R^2 by their inclusion in a model they are absolutely crucial parameters. Consider for example that for snow skiing there was an average participation rate of about .05. Now, as one can see from Figure 3, for much of the Prairies and Atlantic Canada the supply factor had a value of around -.025. If an ANOVA model were used to make predictions based on national parameters they would imply about 1 person in 20 participated when actually only about 1 in 40 did. The magnitude of error which would occur as just indicated is certainly not acceptable in many cases (possibly most) if planning is to be based on estimates.

The preceding example does not in fact show how poor estimates will be in an exceptional case but rather, since many supply effects are larger in absolute value than .15 and some larger than .35 in absolute value, large errors in predictions due to failure to consider supply factors can be expected to be usual rather than unusual. Certainly the result just cited show that the accurracy estimated using the procedure given in TN 6 is illusory if a supply factor is not included in the model used to make predictions. The supply factor must be included to remove model bias. If supply factors are independently determined (either from residuals of a regression where socio-economic effect were determined or from a separate data set), then the kind of variance estimates obtained in TN 6 need only be modified by adding on variance related to error in the supply factors to determine reliability of estimates:

Added variance = Σ_r (Number of people in area r)^2 (Variance in supply factor for r)

WHERE the sum is over all areas, r, being considered.

Obviously, when supply factors are not accurately known the variance from this source can be larger than error from other coefficients.

CONCLUSIONS

This TN has shown that, under a set of reasonable assumptions, highly significant supply factors can be derived to show how participation varies among regions within Canada. Unfortunately this variation can be the result of variations in physical supply or, possibly, of cultural differences or other variations, The authors feel, however, that supply differences are the most important cause of variation for the activities considered. So a next step is to try to explain the effects measured on the basis of physical supply data modified by variables reflecting such factors as advertising. Such work is quite possible because some of the coefficients measuring the effects of supply are accurate enough that they can be compared to regional inventory data.

In conclusion, this paper has achieved its primary purpose of deriving supply factors. Some other interesting conclusions have been reached about the adequacy of the model, data requirements and so on which should be of help to researchers who may wish to further this work. In particular this paper has given some limited answers about how to proceed in defining a supply parameter for demand analysis, what expected levels of explained variance will be and so on. Possibly of most importance are the guidelines in the Appendix which show the number of observations needed in any region before a researcher can expect to have much chance of deriving adequately accurate supply factors to carry out further analysis. With these guidelines studies which go beyond this one can be designed to succeed rather than to fail.

APPENDIX: THE PROBLEM OF HAVING ENOUGH OBSERVATIONS ON A DESTINATION AREA THAT SUPPLY EFFECTS WILL BE STATISTICALLY SIGNIFICANT AND/OR ACCURATE FOR FURTHER ANALYSIS

Consider that a set of supply effects, r(i), are to be computed for a number, N, of subareas of a country or province. Let an estimate of the participation that would occur if supply were at the same level in all regions (and there were no cultural factors, etc.) be PE(i,x) for an individual x in region Then, one can argue that an appropriate estimate of the real probability of person x participating is:

P(i,x) = PE(i,x)+r(i)

Given the equation above, it is possible to derive some equations that can be used:

(1) to determine if the variances in actual estimated r(i) are what they should be expected to be and

(2) to determine the approximate number of observations necessary in a region i to have a certain probability that a supply effect will be estimated with, say, a S.D. R(i), about δ of the value of r(i) (e.g. S.D. R(i) approx = .1r(i) or less with a certain probability).

Still it must be recognized that one must be interested in having a general accuracy level for a collection of geographic areas because supply effects are calculated around a certain mean and are determined in such a way that they add up to zero. This means that some supply effects are going to be small and others are going to be large simply by virtue of the fact that one area has an "amount of" supply which is close to a national average and another area deviates quite substantially from this. So, it is reasonable to consider that in research design one should have about the same number of interviews in the different geographic areas in which there are roughly the same number of people so no matter which supply factor turns out to be large and which turns out to be small the larger ones will meet some accuracy criteria in terms of S.D. δ=R(i)/r(i).

In expressing error the acceptable amount of error can most conveniently be expressed with respect to some measure of the average deviation of the supply factors from their overall mean value of zero. One convenient measure of this type is R = (1/N) times the sum of the absolute values of the supply factors: in other words a convenient measure against which to assess error is the average absolute value of the supply factors. This number is in some way a measure of “on the average” how large the negative and positive deviations about the mean supply factor values of zero tend to be so let:

R(i) = (1/2N) Σ | r(i) |

Now consider that in most regions the average value of PE(i,x) over all persons in i is near to BAR (es. t 20%) and that by definition PBAR approx = PROl. To be specific, if the average participation rate for hunting for males is .15 and R = .05 then PRO1 = .33 which means that the mean supply factor is of such a size that the effect of its average value is a ±33% change in participation.

Further regarding notation, if one is concerned with the reliability of the r(i)'s the concern can be expressed by indicating that the variance in an r(i) with an estimated value of about R should be:

PRO2 = (standard deviation in r(i) with an estimated value of R)/R

and, in the context of making estimates the standard deviations in r(i)'s are influenced by:

(1) The accuracy with which PBAR is estimated (which can be estimated by methods described in TN 6 (which can be approximated as indicated subsequently) and

(2) The variance in the given r(i) about the mean referred to in point (1) immediately preceding.

Now, with N regions and M observations in each, participation PBAR may be considered to be estimated based on MXN observation of a zero-one variable: in other words PBAR = (N of participants)/MXN. It is well known that this type of estimate has a variance approx = PBAR(1-PBAR)/ MXN. As for the variance in r(i) with a value of R. for the region r(i) there are M zero-one observations which by assumption are from a distribution with p = PBAR t R so, again quoting the well known formula, the variance in r(i) given PEAR approx = P(1-P)/M. One must consider PEAR because the results of a particular analysis is being considered and the value of it in this case depends on the statistical deviation to be expected in r(i) measured from the true PEAR. E(PBAR) and on the additional variance introduced into the estimate of r(i) because PEAR is estimated. Adding the two independent variance elements:

V = variance in r(i) = PBAR (1-PBAR)/MXN + P(1-P)/M

If N is 20 or 30 and P(1-P) is not much larger than PBAR(1-PBAR) (which is most unlikely), then the first term on the right can be ignored so:

V ≈ (PBAR ± R) (1—PBAR ± R) /M

And if $V^{*1/2}$ (= SDP, the standard deviation expected in an r(i) of size R) is to be (PRO2)R and R is to be (PBAR)PROl for the reason introduced earlier:

$[(PRO2)(PBAR)PRO1]^2 = (PBAR \pm (PBAR)\ PRO1)\pm(PBAR \pm (PBAR)PRO1^2)/M$

Multiplying through by M and dropping the $(PBAR \pm R)^2$ (i) because this will usually be 1/4 or less and thus will be negligible compared to the other term on the right and also (2) because the results have to do with a statistical approximation:

$M(PRO2)^2(PRO1)^2PBAR^2 \approx PBAR(1 \pm PRO1)$

$M(PBAR) = (1 \pm PRO1)/((PRO1^2(PRO2)^2)$

The formula given above allows one to show why some of the r(i)'s given in the paper have the level of accuracy that they do and why it was a relatively hopeless matter to estimate supply effects for 75 areas based on 4,000 observations (the attempt described at the beginning of the paper). These examples allow one to see how the formula may be used to estimate the size of sample necessary to estimate r(i)'s accurately enough that secondary analysis to determine how r(i)'s are related to what is on the ground can be carried out successfully.

One should notice in what follows that it is critical that one know rough values of PBAR and of the r(i)'s or be willing to make the success of potentially very expensive research dependent on the assumption that PBAR and r(i)'s will have a certain range of values. Furthermore, it should be noted that the appropriate value of the constant PRO1 is related to the choice of PBAR and r(i)'s. The value of PRO2 is independent of the other values in the sense that it is hard to visualize a successful analysis of say 50 or so residuals to relate them to what is on the ground unless PRO2 is .1 and it should probably be more in the range of .05 or .02. Its value must depend on what "accuracy in the r(i)'s" is necessary to achieve the analysis objectives.

To begin with the matter of the accuracy of supply factors, even the possibility of detecting them with 4,000 observations, consider that with 4,000 observations and roughly 70 parameters being estimated, if the r(i)'s are random:

$$\chi^2(4000\text{-}70)= \frac{\Sigma\ r(i)\text{ for the person})^2}{\text{variance expected for the person}}$$

WHERE χ^2 is Chi squared with m degrees of freedom; variance expected in prediction for a person may be taken to be approximately the variance in PBAR.

For χ^2 with this number of degrees of freedom to be significant at the .05 level, the well known approximation for a χ^2 with over 30 degrees of freedom gives:

$1.65 < (2(\chi^2)^{1/2}) - (2(4000\text{-}70)-1)^{1/2}$

and solving the above, recognizing that (4000-70), the degrees of freedom of the χ^2, one obtains:

$1.1 \leq \chi^2$/(degrees of freedom for the χ^2)

So from the above one concludes that on the average r(i)'s must exceed their variance by 10% if they are to produce a χ^2 that can be accepted with 95% certainty as significant. For this to be true on the average, $(1/PRO2)^2$should be greater that 1.2 or in other words PRO2 should be .9 or less. By the formula derived earlier for 4000 observations in 75 areas M = 4000/75 = 53. A typical PBAR for the CORDS activities of which we must chose an activity for males is .15. Finally, the earlier results show that R, as defined earlier, is only about 1/4 PBAR. Thus:

$53*.15)=(1-1/4)/((1/4)^2 pro2^2)$

$PRO2 = 12^{1/2} /8=1.22$

Now even if R/PBAR = 1/3 the chances of results being accepted as significant are questionable because:

$PRO2 = (9*2/3)^{1/2}/8=.86$

However, in either case if the sample size had been three times as large M=150 then there is no question that the results would have been statistically significant.

For another example, by looking at the weighted regression results for activity one, given in Table 1, one sees that for r(i)'s between .1 and .2 the standard deviations are about .02 or .03, which is about 20% of their expected value. Now, for these data there were on the average 400 observations in each of 60 origin areas:

PBAR = .23 and thus PRO1 approx = .1/.23 approx = 1/2 so that:

$PRO2 =(1+PRO1)/(M(PBAR)(PRO1)^2)^{1/2} \approx .14$

The value of .14 is lower than the .2, or so observed, but since results presented in TN 20 show there are structural problems with the model and because the parameter estimates are not efficient weighted estimates (see Reference 00) the difference is not distressing. Also it should be noted that there are not exactly 400 observations in each origin area and this complicates matters.

For future analyses of the r(i)'s to determine how they relate to the supply that is on the ground simple statistical significance of the r(i)'s is not enough. For example the goal might be to study a relation like:

Supply measure=Σ(attractiveness)[a]/distance[b]

As indicated earlier for such analyses PRO2 should be .1 and probably less. So, in the context of the last two examples, one finds that M should be:

$M=(1/.15)(1-1/3)/((1/4)^2(.1)^2)=8000$

So, for 75 areas 600,000 interviews would be needed with larger r(i)'s:

$M==(1/.15)(1-1/3)/((1/3)^2(.1)^2)=6000$

Obviously, research to define the relevant relation between what is on the ground and peoples' behaviour is going to have to be well planned so that data need be collected in only 5 to 10 areas in which there are the necessary distinct supply difference to allow the adequacy of parameters to be estimated and a relation to be tested.