Regression Methods and Performance Criteria for Small-Area Population Estimation

P. McCullagh
James V. Zidek

Technical Report No. 06

February 1985

REGRESSION METHODS AND PERFORMANCE CRITERIA FOR SMALL-AREA POPULATION ESTIMATION

by

P. McCullagh
Imperial College

J. V. Zidek
University of
British Columbia

January 1983

ABSTRACT

In this paper a regression model is developed for the post-censal estimation of the population sizes of small-areas. The approach is nonstochastic. It is assumed that current population sizes have been determined by a function of those obtained at the last census, together with the associated values of certain symptomatic variables. As well, the current values of these variables are assumed to be in-hand. Natural properties of such a rule are shown to imply a specific, log linear form for this function. Existing models are shown to be approximations to the result.

An objective function for evaluating arbitrary estimation procedures or fitting regression models, is derived on the assumption that revenue allocation is the objective of the estimation program. By appealing to an equilibrium theory for group decision processes, it is shown that under certain conditions the appropriate criterion is given by the Kullback-Leibler discrimination function.

KEYWORDS: Ratio-correlation; difference-correlation; populations of small-areas; local-areas; log-linear models; Kullback-Leibler information; Nash solutions.

ACKNOWLEDGEMENTS

Interest in the problem considered in this paper was stimulated by a problem posed to the second author by the City of Vancouver and by discussions with Mr. Phil Mondor of its Planning Office and Professor H. Hightower, University of British Columbia.

The City of Vancouver supported related research, carried out by the second author and reported elsewhere.

The University of British Columbia and Imperial College, where the work reported herein commenced and finished, respectively, both very generously provided facilities during the period of the leaves of absence of the first and second authors, respectively.

Support for the work was provided by the Natural Sciences and Engineering and the Social Sciences and Humanities Research Councils of Canada, the latter through a Leave Fellowship to the second author.

1. INTRODUCTION

Government and private agencies depend on estimates of the populations of small-areas for a variety of purposes, such as revenue allocation and planning. In the United States, for example, annual, post-censal estimates are prepared under the Federal-State Co-operative Program (FSCP) for something like 39,000 municipalities, counties and so on (Kitagawa et al. 1980).

The great scale on which this activity is carried out makes the need for accurate, simply applied procedures acute. A variety of methods exist (Purcell and Kish 1979, Kitagawa et al. 1980, Zidek 1982). None are simpler and more adaptable than those which use regression models for making these estimates (Schmitt and Crosetti 1954, Morrison and Relles 1975, O'Hare 1976). These models turn easy-to-measure quantities (symptomatic variables) into estimates of population sizes which are expensive to measure. The coefficients obtained by fitting these models using two successive censuses and the measured values of all variables, one set per sub-area, (implicitly) account for migration, demographic trends and so on, so that the latter are not needed in the preparation of the estimates as they would be in, say, administrative records methods. Different symptomatic variables may be used in different sub-areas, depending on what data is readily available in each. And comparative empirical studies in the works cited above, show that these methods can be very effective, surprisingly so since the regression methodology is misappropriated in the sense that there is no conceivable experiment which generates these variable-values at random as in the conventional case. One such method, that using the ratio-correlation model, is among the three, regression and other, which constitute the overall strategy used in the FSCP (Kitagawa et al. 1980).

In spite of the relatively high precision of regression-based methods, their simplicitly and adaptability, there have been surprisingly few models proposed and these have not been very much refined. Rosenberg (1968) points out the need to stratify local-areas by the dichotomies, first of urban versus rural and second of rapid versus slow growth. His proposal seems to have been ignored. Namboodiri (1972) argues pursuasively against the needed temporal stationarity of regression models and gives an analysis which may well show the ill-effects of multicollinearity among the symptomatic variables. A systematic residual analysis has not been published and little seems to be known about the possibly serious negative impact of

influential observations in this context (see Belsey, Kuh and Welsch (1980) for a general discussion of this problem.

Existing regression models have been proposed on an *ad hoc* basis. The major result of Section 3 is a new model which is obtained by what might be called an axiomatic approach. Stochastic and approximation errors aside, it is supposed that population sizes of small-areas are assigned, hypothetically, by some unspecified function of the symptomatic variables and previous counts. The most reasonable requirements of such a rule are specified and then shown to be equivalent to a log-linear model similar but not identical to that of Morrison and Relles (1976). The model has yet to be empirically assessed.

Both the fitting of regression models and the assessment of estimation methods requires a criterion function which accumulates in some sensible way the errors made over all sub-areas. No particular choice has yet been indicated and various alternatives such as average relative absolute error are used, sometimes several in the same study. None of these criteria seem directly related to a primary objective of the program, namely, the allocation of revenue. In Section 2 is derived from first principles, a criterion function which takes account of the need to choose an allocation scheme which would be a jointly acceptable compromise to all members of the community, provided their desires obey certain weak constraints. The derivation is based on the theory of Nash (1950).

2. CRITERIA

Evaluating the performance of an estimation methodology requires answers to two fundamental questions. Against what are the answers it produces to be compared and by what criterion is the comparison to be made? This section is addressed to the second of these questions. We have no alternative to propose to the practical answer which is commonly given to the first question, namely, the corresponding census counts. The latter typically underestimate the true counts by something like 2% (Hauser 1981) and therefore seem somewhat unsatisfactory.

Kitagawa et al.(1980) describe various estimation performance criteria in terms of P_i, the "actual" and $\hat{P}_i$, the estimated population sizes for

subregion $i = 1, \ldots, n$. The former would be the last available census counts corrected, possibly, as Spencer suggests (Appendix I, Kitagawa et al. 1980) for undercoverage. These criteria include average error, average relative error, number of extremely large relative errors and bias. The first two of these are obtained by putting $a = 1$, $b = 0$ and $a = 1$, $b = 1$, respectively, in the general index $\sum |P_i - \hat{P}_i|^a / P_i^b$.

While Kitagawa et al.(1980) refer to the need to take account of the purposes of the estimation program, no criterion has been given which does so, other than that which is derived below.

Its derivation relies on Nash's theory of bargaining as it might be applied in the present situation (Nash 1950). Let P and A denote, respectively, the population size and amount (of revenue, say) to be allocated. Let $\underset{\sim}{A} = (A_1, \ldots, A_n)$ be a feasible allocation of A among the n subregions and $u_{ij}(\underset{\sim}{A})$, $j = 1, \ldots, P_i$, $i = 1, \ldots, n$ represent the gain-in-value (utility) to individual j in subregion i which would result from this allocation scheme. Certain weak assumptions imply that any equilibrium solution must maximize an objective function which will now be described.

Before doing so, it should be pointed out that Nash's theory admits as potential solutions, not only the allocations, $\underset{\sim}{A}$, themselves but as well, all randomized mixtures of the $\underset{\sim}{A}$'s. Deadlocks can therefore be broken by tossing a coin, as it were. The domains of the u_{ij}'s are extended to this more general solution set by invoking the expected utility hypothesis. However, the feasible solutions, randomized and non-randomized alike, are required to satisfy $u_{ij} \geq 0$. In this way, the Nash theory ensures that no individual can be made to suffer a net expected loss of utility as a result of the proposed allocation.

In agreement with practice, we will restrict ourselves to nonrandomized allocations and define as optimal any that are feasible and maximize the so-called Nash product given by

$$NP(\underset{\sim}{A}) = \prod_i \prod_j u_{ij}^{I/P}(\underset{\sim}{A}) \tag{2.1}$$

If the u_{ij}'s are unknown, as would be the case in the situation under consideration, an approximation to (2.1) becomes necessary. If the A_i's are moderate, it is reasonable that $u_{ij}(\underset{\sim}{A})$ would equal, approximately,

A_i/P_i for all i and j. This approximation is supported by the assumptions that the i-th subregion's allocation is evenly distributed among its constituents, that they receive no benefit from the allocations made to other subregions and that the utilities are linear. This approximation yields

$$MP(\underset{\sim}{A}) = \prod (A_i/P_i)^{P_i} \tag{2.2}$$

where $p_i = P_i/P$ is the proportion of the population in subregion i for all i.

Equation (2.2) may be reduced further by letting $A_i = a_i A$, $i = 1, \ldots, n$. Then

$$MP(\underset{\sim}{A}) = (A/P) \exp\left[-I(\underset{\sim}{p},\underset{\sim}{a})\right] \tag{2.3}$$

where $I(\underset{\sim}{p},\underset{\sim}{a}) = \sum p_i \log(p_i/a_i) \geq 0$ is the Kullback-Leibler "distance" between $\underset{\sim}{a}$ and $\underset{\sim}{p}$.

To maximize MP is to minimize I, i.e. to choose $\underset{\sim}{a} = \underset{\sim}{p}$. However, in practice $\underset{\sim}{p}$ would not be known. It would seem natural then to choose for $\underset{\sim}{a}$ the best available estimate, say $\hat{\underset{\sim}{p}}$ of $\underset{\sim}{p}$.

The approximate MP-criterion given in equation (2.3) can be obtained by entirely different reasoning. Suppose a random sample of the region's current population is drawn with replacement and each individual so-obtained is classified by subregion. Let p_i be the observed sample fraction of individuals from subregion i. Then the function of $\underset{\sim}{a}$ given in equation (2.3) is the likelihood function for these data. It would be maximized to find the optimal estimate among allowable choices of $\underset{\sim}{a}$ to find the maximum likelihood estimate of the true subregion population proportions. This would be $\underset{\sim}{a} = \underset{\sim}{p}$ unless $\underset{\sim}{a}$ were constrained.

This sampling-theoretical point of view suggests a natural alternative to the criterion given in equation (2.3). If the hypothetical sample were large and the a_i's were the "true" regional proportions, then the consistency of the (unconstrained) maximum likelihood estimator, $\underset{\sim}{p}$, would yield the approximation $\log(a_i/p_i) = (a_i/p_i - 1) - (a_i/p_i - 1)^2/2$. This, in turn would yield

$$MP(\underset{\sim}{A}) = (A/P)\exp\left[-\tfrac{1}{2}\sum(a_i-p_i)^2/p_i\right] \qquad (2.4)$$

Equation (2.4) suggests the minimum chi-squared criterion for choosing the $\{a_i\}$, namely minimize

$$\chi^2 = \sum(a_i-p_i)^2/p_i\,. \qquad (2.5)$$

Again, if $\underset{\sim}{a}$ were unconstrained, $\underset{\sim}{a} = \underset{\sim}{p}$ would be optimal.

This last criterion among others receives special attention in Kitagawa et al.(1980). It represents according to these authors, a compromise between $\sum(a_i-p_i)^2$ and $\sum|a_i-p_i|/p_i$. The first would be unduly sensitive to large individual, subregional misallocations and the second, to misallocations in small areas.

3. REGRESSION METHODS

Such methods yield post-censal estimates, $\hat{P}_{2i}$ of current (time $t = 2$, say) population sizes, P_{2i}, for subregions. They rely on models which involve coefficients which must be fitted, the observed current values of symptomatic variables, $\underset{\sim}{S}_{2i} = (S_{2i1}, \ldots S_{2ip})$, and their corresponding values at time $t = 1$, $\underset{\sim}{S}_{1i}$, when population sizes, P_{1i}, are available for all regions, $i = 1, \ldots, n$.

To fit the coefficients, a criterion is chosen, usually least squares but possibly that given by I or χ^2 in equations (2.3) or (2.5) respectively. The times, $t = 1$ and 2, are taken to be successive census years so that the $\{P_{ji}\}$, $\{\underset{\sim}{S}_{ji}\}$ $j = 1,2$, $i = 1, \ldots, n$ are observable. Then p_i is taken to be the observed value of the "dependent" variable, $P_{2i}/\sum P_{2i}$, $i = 1, \ldots, n$ while a_i is supplied by the regression model, albeit with as yet unspecified coefficients. The coefficients are then chosen to give the criterion-functional its least possible value.

Erickson (1973, 1974) proposed an interesting variation of this scheme which might be called "sampling-regression". The time $t = 2$ is the present and the model is fitted to the results of a census carried out on a subsample of subregions. This approach takes account of the inevitable temporal nonstationarity of any regression model by "tuning" its coefficients to the present. The sampling scheme which was introduced in Section 2 to

provide an interpretation of the MP^2 criterion could provide a (no doubt inferior) alternative to Erickson's plan. A bonus of the sampling-regression approach which is pointed out by Purcell and Kish (1979) is that, unlike the conventional approach, it carries with it a basis for inference.

It remains to choose a suitable model. As will now be shown, simple intuitive requirements lead quite easily to particular models.

For expository simplicity it will be assumed that $p = 1$ so that the symptomatic variables at times $t = 1$ and $t = 2$ become S_{i1} and S_{i2} respectively. The "dot" will as usual represent summation over any subscript it replaces. Thus $P_{2\cdot}$ would denote the population of the entire region at time $t = 2$.

Let $\hat{\underset{\sim}{P}}_2 = \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$ represent an unspecified post-censal regression estimation model. Here $\underset{\sim}{T} = (T_1, \ldots, T_n)^T$ is the vector of subregional estimation models while $\underset{\sim}{S}_{jt} = (S_{j1}, \ldots, S_{jn})^T$, $j = 1,2$, and $\underset{\sim}{P}_1 = (P_{11}, \ldots, P_{1n})^T$ are the corresponding vectors of symptomatic variables and population sizes at time $t = 1$." Observe that, in general, T_i is a function of $\underset{\sim}{S}_2$, $\underset{\sim}{S}_1$ and $\underset{\sim}{P}_1$ for all i.

Suppose, hypothetically, that subregional population sizes were to be assigned without error by means of the model, $\underset{\sim}{T}$ introduced above. Various reasonable requirements of such a rule suggest themselves and lead, as is shown below, to explicit forms for $\underset{\sim}{T}$. These may be taken as first approximations to the actual population sizes and so used to obtain estimates. The precision of such estimates and hence the value of the models suggested by the approach outlined above, would need to be ascertained by empirical study.

Possible requirements for $\underset{\sim}{T}$ are presented and discussed. Their implications are given below.

Regional population sizes may be estimated with a relatively high precision compared to those of its subregions. So they may be regarded as known. The following requirement is therefore considered as fundamental:

CTT (Controlled-to-Total). $\hat{P}_{2\cdot} = P_{2\cdot}$

It implies that we may regard $\underset{\sim}{P}_2$ and $\underset{\sim}{P}_1$ as vectors of proportions, each set summing to 1, rather than as vectors of population counts whenever it is expedient to do so.

Another important condition is:

PI (Permutation invariance). For all $n \times n$ permutation matrices Φ

$$\underset{\sim}{T}(\Phi \underset{\sim}{S}_2, \Phi \underset{\sim}{S}_1, \Phi \underset{\sim}{P}_1) = \Phi \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1).$$

This condition simply ensures that the order in which the sub-regions are listed is irrelevant. It is not the same as spatial homogeneity because the symptomatic variables, $\underset{\sim}{S}_1$ and $\underset{\sim}{S}_2$, may contain information pertaining to the geography of the sub-regions.

The following invariance requirements ensure that the model has good robustness properties. They make use of the fact that the scale on which the symptomatic variables are measured is usually arbitrary. Furthermore, if the symptomatic variables are counts computed locally (e.g. births, marriages, deaths) there may be a degree of under-reporting that varies from one region to another. To some extent these effects are eliminated if the following invariance conditions are satisfied:

RI (Regional invariance). For all positive diagonal matrices D

$$\underset{\sim}{T}(D\underset{\sim}{S}_2, D\underset{\sim}{S}_1, \underset{\sim}{P}_1) = \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$$

TI (Temporal invariance). For all positive scalars $a_1, a_2 > 0$

$$\underset{\sim}{T}(a_2 \underset{\sim}{S}_2, a_1 \underset{\sim}{S}_1, \underset{\sim}{P}_1) = \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1).$$

The next condition embraces a natural equivariance requirement. It derives from the recognition that the estimated growth rate in each region, i, $\hat{P}_{2i}/P_{1i}$, would not change even if $\underset{\sim}{P}_1$'s co-ordinates were replaced by their corresponding densities, say per hectare or per square kilometre, for example. This suggests the condition $T_i(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)/(P_{1i}) = T_i(\underset{\sim}{S}_2, \underset{\sim}{S}_1, D\underset{\sim}{P}_1)/(d_i P_{1i})$ $i = 1, \ldots, n$, for any positive $D = \mathrm{diag}(d_1, \ldots, d_n)$. This is equivalent to

$\underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, D\underset{\sim}{P}_1) = D\underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$, i.e. $T_i(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1) = P_{1i} H_i(\underset{\sim}{S}_2, \underset{\sim}{S}_1)$, $i = 1, \ldots, n$ for some function H. However, this condition is inconsistent with CTT which is considered more fundamental. And it is unnecessarily strong for it is possible, as our analysis will show, to derive a sufficiently small class of potential models under the weaker requirement that $g_{ij}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1) \triangleq (\hat{P}_{2i}/P_{1i}) \div (\hat{P}_{2j}/P_{1j})$, the relative growth rates of regions i and j, $i,j = 1, \ldots, n$, be invariant in the sense described above. This condition is $g_{ij}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, D\underset{\sim}{P}_1) = g_{ij}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$ $i,j = 1, \ldots, n$, for all positive diagonal matrices, D which is equivalent to

RGRI (Relative Growth Rate Invariance). For $i,j = 1, \ldots, n$ and all positive $D = \mathrm{Diag}(d_1, \ldots, d_n)$,

$$G_{ij}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, D\underset{\sim}{P}_1) = G_{ij}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$$

where $G_{ij}(\underset{\sim}{x}, \underset{\sim}{y}, D\underset{\sim}{z}) = \left[T_i(\underset{\sim}{x}, \underset{\sim}{y}, D\underset{\sim}{z})/d_i\right]\left[T_j(\underset{\sim}{x}, \underset{\sim}{y}, D\underset{\sim}{z})/d_j\right]^{-1}$

for all $\underset{\sim}{x}$, $\underset{\sim}{y}$ and $\underset{\sim}{z}$ in T's domain.

Instead of regarding, as we may because of CTT, $\underset{\sim}{P}_1$ and $\underset{\sim}{P}_2$ as vectors of proportions, it is more convenient to deal with the $(n-1)$-vector of ratios, $\underset{\sim}{P}_1^*$, with elements P_{1i}/P_{1n}, $i = 1, \ldots, n-1$. Similar changes in $\underset{\sim}{T}$ yields $\underset{\sim}{P}_2^* = \underset{\sim}{T}^*(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1^*)$. The advantage of $\underset{\sim}{P}_2^*$ over $\underset{\sim}{P}_2$, the vector of proportions, is that the elements of $\underset{\sim}{P}_2^*$ are unrestricted even when CTT is imposed.

Conditions RI and TI are equivalent to $\underset{\sim}{T}^*(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1^*) = \underset{\sim}{T}^*\left[(\underset{\sim}{S}_2/\underset{\sim}{S}_1), \underset{\sim}{1}, \underset{\sim}{P}_1^*\right]$ and RGRI to $T_i^*(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1^*) = P_i^* T_i^*(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{1})$ $i = 1, \ldots, n$. So RI, TI and RGRI combined are equivalent to

$$T_i^*(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1^*) = P_i^* H_i\left[(\underset{\sim}{S}_2/\underset{\sim}{S}_1)^*\right] \tag{3.1}$$

for some function H_i, $i = 1, \ldots, n$.

We regard CTT, PI, RI, TI and RGRI as important: the first two would seem to be essential but it is possible to think of conditions under

which the invariance requirements are not so compelling. The following conditions, although plausible, seem less important. They, or conditions like them, are needed to reduce the general model given in equation (3.1) to a more explicit, applicable form:

TC (Temporal Coherence). Given an additional time, $t = 0$

$$\underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_0, \underset{\sim}{P}_0) = \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \hat{\underset{\sim}{P}}_1) \quad \text{where}$$

$$\hat{\underset{\sim}{P}}_1 = \underset{\sim}{T}(\underset{\sim}{S}_1, \underset{\sim}{S}_0, \underset{\sim}{P}_0).$$

TR (Time Reversibility). For all S_1, S_2, P_1

$$\underset{\sim}{P}_1 = \underset{\sim}{T}(\underset{\sim}{S}_1, \underset{\sim}{S}_2, \hat{\underset{\sim}{P}}_2) \quad \text{where}$$

$$\hat{\underset{\sim}{P}}_2 = \underset{\sim}{T}(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1).$$

The intuitive basis for TC is clear and it nearly implies TR since the latter becomes the former when the past is reflected into the future provided $\underset{\sim}{T}(\underset{\sim}{S}_0, \underset{\sim}{S}_1, \underset{\sim}{P}_1)$ is given the value $\underset{\sim}{P}_0$, the subregions' known population sizes when the symptomatic variable, $\underset{\sim}{S}_0$, reassumes its original (time $t = 0$) value.

These conditions imply an easily derived explicit form for the function H_i of equation (3.1). The solution, $f(y) = ay$ of Euler's functional equation $f(x+y) = f(x) + f(y)$, f continuous, is used. The solution obtains even if the requirement $x+y \leq K$ is imposed, $0 \leq x, y \leq K$. The result is

$$H_i(\underset{\sim}{y}) = \prod_{j=1}^{n} y_j^{\alpha_{ij}} \tag{3.2}$$

for certain constants, $-\infty < \alpha_{ij} < \infty$.

If condition PI is imposed, a straightforward argument which is omitted for brevity shows that $\alpha_{ij} = \alpha$ or 0 according as $i = j$ or $i \neq j$ for some constant α. The model implied by all of the conditions given above is, then, in summary,

$$\hat{P}_{2i} = P_{2\cdot}\ P_{1i}\ R_i^{\alpha}\left[\sum_{k=1}^{n} P_{1k}\ R_k^{\alpha}\right]^{-1}$$

where $R_i = S_{2i}/S_{1i}$. Of course, P_{1i} and R_i can be replaced by P_{2i}^* and R_i^* in equation (3.3) without changing the result. Alternagively, they can be replaced by their corresponding "shares", $P_{1i}/P_{1\cdot}$ and $R_i/R_{\cdot}$, respectively. In the case of more than one symptomatic variable, obvious extensions of the conditions stated above lead to a generalization of the model given in equation (3.3) :

$$\hat{P}_{2i} = P_{2\cdot}\ P_{1i}\ \prod_{j=1}^{p} R_{ij}^{\alpha_j}\left[\sum_{k=1}^{n} P_{1k}\ \prod_{j=1}^{p} R_{kj}^{\alpha_j}\right]^{-1} \qquad (3.4)$$

where R_{ij}, the counterpart of R_i, is the growth rate for symptomatic variable j, $j = 1, \ldots, p$.

An alternative form of the model derived above is obtained by taking logarithms. The result is a variant of that suggested by Morrison and Relles (1976) :

$$\log(\hat{P}_{2i}/P_{1i}) = \sum_{j=1}^{p} \alpha_j \log R_{ij} \quad . \qquad (3.5)$$

The basis for their choice is not given. The resulting estimate, $\hat{P}_{2\cdot}$, would not necessarily satisfy CTT and would therefore need to be "controlled" after the parameters, α_j, $j = 1, \ldots, p$ were fitted. The resulting estimates would, in general, differ from those obtained by applying the model given in (3.4). Neither of the models (3.4) or (3.5) has undergone a comparative empirical study.

Approximations to these models may be derived using the approximation, $\log x \approx x - 1$ which is accurate if $x \approx 1$. So in stable subregions i, that is, those undergoing slow changes in population size, equation (3.5), for example, would yield the approximate model

$$\hat{P}_{2i}/P_{1i} = \alpha_0 + \sum_{j=1}^{p} \alpha_j\ R_{ij} \, . \qquad (3.6)$$

This model, or that obtained by replacing all its components by their shares, is called the ratio-correlation model. It is the most commonly used of the various alternatives and is the regression component of the composite methodology used in the Federal State Co-operative Program of the United States (Kitagawa et al. 1981). Empirical studies presented in this last-cited work show that this model yields very good results in exactly those subregions where the above approximation would be accurate.

If in the model of equation (3.5) the P's are replaced by their shares and logarithms are taken, an approximate model is obtained :

$$\hat{P}_{2i} - P_{1i} = a_0 + \sum_{j=1}^{p} a_j (r_{2ij} - r_{1ij}) \qquad (3.7)$$

where $p_{ki} = P_{ki}/P_{k\cdot}$ and $r_{kij} = S_{kij}/S_{k\cdot j}$, $k = 1,2$, $i = 1, \ldots, n$, and $j = 1, \ldots, p$. This model would be expected to be appropriate for subregions whose population sizes comprise a relatively large fraction of that of the region. This so-called difference-correlation model was proposed by O'Hare (1976) who provides empirical results which show that this model is marginally superior to its ratio counterpart. However, under this model, the data may well be heteorscedastic (Dr. D. Herman, personal communication) so its value remains uncertain.

4. CONCLUDING REMARKS

This paper presents a new regression model (equation (3.4)) for estimating the population sizes of small-areas. Its value remains to be determined by empirical study. However, the argument of Section 3 makes it a natural choice. The ratio-correlation model in current use in the U.S.A.'s Federal-State Co-operative Program would approximate the proposed model in sub-regions of stable population size for reasons given in Section 3.

In fitting and empirically evaluating a model such as derived in Section 3, a criterion or "objective function" must be specified. That

obtained in Section 2 (equation (2.3)) is novel in this context. It is, approximately, the Nash criterion (Nash 1950) for determining the joint value to a group (society) of any proposed group action (revenue allocation scheme). It implies that proportional allocation on the basis of population size is optimal. Thus the value of an allocation scheme will depend on how well the best available estimates of sub-regional population proportions approximate the exact values.

The proposed criterion is the product of an attempt to relate the performance of an estimation method to one of the primary objectives of the estimation program, albeit under the oversimplification of linear utility functions. While this criterion may be used in evaluating any estimation procedure through the estimates of proportions, and hence approximate allocation scheme it produces, its form is, fortuitously, particularly appropriate for the regression model developed in Section 3.

Regression models used in current practice are fitted by least squares even though their performance is measured by other criteria, such as average relative absolute error. The justification for this apparent inconsistency is unknown. In any case, we would propose to use the criterion of Section 2 in both fitting and assessment.

The requirements imposed in Section 3 are the most reasonable of the various alternatives. The condition of regional independence, $T_i(\underset{\sim}{S}_2, \underset{\sim}{S}_1, \underset{\sim}{P}_1) = T_i(S_{2i}, S_{1i}, P_{1i})$, $i = 1, \ldots, n$, for example, was ruled out because it forces the model to ignore a vital piece of information, the (assumed) known regional total $P_{2.}$. If it were admitted, then PI would imply that $T_i = T$ for some T and all i. On top of these conditions spatial coherence, $\sum_i \hat{P}_{2i} = T(\sum_i S_{2i}, \sum_i S_{1i}, \sum_i P_{1i})$ for all $\underset{\sim}{S}_2$, $\underset{\sim}{S}_1$, and $\underset{\sim}{P}_1$, would be equivalent to $P_{2i} = \alpha_1 S_{1i} + \alpha_2 S_{2i} + \gamma P_{1i}$ for all i. The result is not inconsistent with RI (regional invariance) and TI (temporal invariance); however, adding this pair of conditions would lead to $\hat{P}_{2i} = \gamma P_{1i}$, a model which takes no account of the symptomatic variables. Replacing the pair in question, successively by one of TC and then TR leads, respectively, to $\hat{P}_{2i} = \alpha(S_{2i} - S_{1i}) + P_{1i}$ and $\alpha(S_{1i} - \beta S_{2i}) + \beta P_{2i}$, $\beta^2 = 1$, according as TC or TR is introduced. The efficacy of such linear models is unknown.

REFERENCES

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*. New York: Wiley.

Erickson, E.P. (1973). A method for combining sample survey data and symptomatic indicators to obtain population estimates for local areas. *Demography*, 10, 137-160.

Erickson, E.P. (1974). A regression method for estimating population changes of local areas. *Jour. Amer. Statist. Assoc.*, 69, 867-875.

Hauser, P.M. (1981). The census of 1980. *Scientific American*, v. 245, # 5, 53-61.

Kitagawa, E.M. et al. (1980). *Estimating population and income of small areas*. Washington: National Academy Press.

Morrison, P.A. and Relles, D.A. (1976). A method of monitoring small-area population changes in cities. *Public Data Use*, v.3, 10-15.

Namboodiri, N.K. (1972). On the ratio-correlation and related methods of subnational population estimation. *Demography*, v.9, 443-453.

Nash, J.F. Jr. (1950). The bargaining problem. *Econometrica*, 18, 155-162.

O'Hare, N. (1976). Report on a multiple regression method for making population estimates. *Demography*, v.13, 369-379.

Purcell, N.J. and Kish, L. (1979). Estimation for small domains. *Biometrics*, 35, 365-384.

Rosenberg, H. (1968). Improving current population estimates through stratification. *Land Economics*, 44, 331-338.

Schmitt, R.C. and Crosetti, A.H. (1954). Accuracy of the ratio correlation method for estimating postcensal population. *Land Economics*, 30, 279-281.

Zidek, J.V. (1982). A review of methods for estimating the populations of local areas. Tech. Rep. 82-4. Instit. of Applied Maths. and Statist., U. of British Columbia.