Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

SINGLE FRAME ESTIMATION IN MULTIPLE FRAME SURVEY

Fulvia Mecatti[1]

ABSTRACT

Multiple Frame Survey has been originally proposed according to an *optimality* approach in order to persecute survey cost savings, especially in the case of a complete list available but expensive to sample. In the modern sampling practice it is frequent the case where one complete and up-to-date list of units, to be used as sampling frame, is not available. Instead, a set of two or more lists singularly partial, usually overlapping, with union offering adequate coverage of the target population, can be available. Thus the collection of the partial lists can be used as Multiple Frame. Literature about Multiple Frame estimation theory mainly concentrates over the Dual Frame case and it is only rarely concerned with the important practical issue of the variance estimation. By using a *multiplicity* approach a fixed weights Single Frame estimator for Multiple Frame Survey is proposed. The new estimator naturally applies to any number of frames and requires no information about unit domain membership. Furthermore it is analytically simple so that its variance is given exactly and easily estimated. A simulation study comparing the new estimator with the major Single Frame competitors is also presented.

KEY WORDS: Confidentiality; Difficult-to-Sample Population; Multiplicity; Simulation; Variance Estimation.

1. INTRODUCTION

In the modern sampling practice it is not rare the case where a unique, complete and up-to-date list of units, to be used as sampling frame, is not available and it may not be built unless expensive or unfeasible screenings. For instance when dealing with a *rare* population such as persons with a rare disease, with *elusive* and/or *hidden* populations such as homeless, illegal immigrants or drug consumers, and in general when treating with *special* or *difficult-to-sample* populations (Sudman and Kalton, 1986, Lepkowski, 1991). Instead, a set of $Q \geq 2$ unit lists can be available. In the general case, lists are singularly partial and overlapping each other, though their union might offer an adequate coverage of the target population *U*. That is the context known as Multiple Frame Surveys (MF). Although literature on MF dates back to early sixties, in a recent paper Lohr and Rao stated: << *As the U.S., Canada and other nations grow in diversity, different sampling frames may better capture subgroups of the populations.* […] *We anticipate that modular sampling designs using multiple frames will be widely used in the future* >> (Lohr and Rao, 2005).

Literature mostly concentrates on the case of $Q = 2$ available unit lists, called Dual Frame Survey (DF), where two frames *A* and *B* are given with $A \cup B = U$ and sizes N_A and N_B, usually known, such as $N_A + N_B \geq N$, where *N* denotes the population size, usually unknown. The two overlapping frames are virtually divided into three disjoint *domains*: $a = A - B = A \cap B^C$, $b = B - A = B \cap A^C$ (where ^{C} denotes complementation) and the so-called *overlapping domain* $ab = A \cap B$. Hence the population size equals the sum of the domain sizes $N_a + N_b + N_{ab} = N$ and the total $Y = \sum_{i \in A \cup B} y_i$ of the survey variable, assumed as the parameter to be estimated, equals the sum of the domain totals $Y_a + Y_b + Y_{ab} = Y$ where for instance $Y_a = \sum_{i \in a} y_i$. Two random sample s_A and s_B are selected independently from the two frames under a given sampling design possibly different into each frame. Sample data from the two frames are used to produce estimates for the domain totals. Estimated domain totals are finally combined to provide estimation for the population total *Y*.

[1]Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8 – Ed. U7, 20126 Milano, Italy. (fulvia.mecatti@unimib.it)

A DF estimator has been originally proposed by Hartley (1974) focusing on the case where a complete frame *A* is available but expensive to sample and a second frame *B* is available but partial. With the purpose of persecuting cost savings by achieving the same or greater efficiency, expensive data from the complete frame *A* are combined with cheaper information from frame *B* under an *optimality approach*. Particularly, the estimator for the population total Y is produced by combining estimators of the domain totals with *optimum weights* $\hat{Y}_H = \sum_{i\in s_a^{(A)}} y_i + \sum_{i\in s_b^{(B)}} y_i + \vartheta\sum_{i\in s_{ab}^{(A)}} y_i + (1-\vartheta)\sum_{i\in s_{ab}^{(B)}} y_i$ namely with ϑ minimizing the estimator variance, where for instance $s_{ab}^{(A)}$ denotes the subsample of s_A of units included into the overlap domain *ab*. The optimal Hartley's estimator has been successively improved. Particularly Fuller and Burmeister (1972) introduced a optimal estimator for the DF which has been interpreted as a maximum likelihood estimator (Skinner, 1991) and proved to be asymptotically efficient (Lohr and Rao, 2000). Some practical and methodological problems can be listed when the optimality approach is adopted and the optimal estimators applied: *i)* the optimum weights turn out to depend on unknown variances and covariances so that they have to be estimated from sample data, which could be complicated and affecting optimality itself; *ii)* estimated weights depend on the survey variable values so that they differ for different survey variables, which is unpractical for multipurpose surveys; *iii)* the generalization to MF setup is not straightforward or even impossible (Skinner, 1991); *iv)* the important practical issue of variance estimation is scarcely considered; *v)* the *k*nowledge of the domain membership for every sampled units is required. *i.e.* the correct classification into domains of units sampled into each frames has to be performed in order to apply optimal estimators. This is a strong assumption as stated for instance in Lohr and Rao (2005) since optimal estimators result sensitive to misclassification of sampled units into domains.

Point *ii)* has been addressed by Skinner and Rao (1996), along with the application to complex sampling designs, by introducing a pseudo-maximum likelihood estimator (PML) for DF; in Lohr and Rao (2005) point *iii)* is concerned for both optimal and PML estimators; in the present paper we focus on points *iv)* and *v)*. Particularly, in Section 2 the single frame approach is recalled. In Section 3 a multiplicity estimator for MF is proposed. Some simulation results are presented in Section 4.

2. SINGLE FRAME ESTIMATION

In alternative to the optimality approach, a *single frame* (SF) *approach* can be used. In a SF estimator data from the two frames are combined by using fixed weights depending on the inclusion probabilities under the design induced by the two frame designs over the *total* sample, *i.e.* the union of the two frame samples (Bamkier, 1986; Kalton and Anderson, 1986; Skinner, 1991): $\hat{Y} = \sum_{i\in s_A} w_i y_i + \sum_{i\in s_B} w_i y_i$ where $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$ with $\pi_{Ai} = 0$ if $i \in b$ and $\pi_{Bi} = 0$ if $i \in a$. For simple random sampling (SRS) of each frame, the SF estimator for DF is given by: $f_A^{-1}\sum_{i\in s_a^{(A)}} y_i + (f_A + f_B)^{-1}\left(\sum_{i\in s_{ab}^{(A)}} y_i + \sum_{i\in s_{ab}^{(B)}} y_i\right) + f_B^{-1}\sum_{i\in s_b^{(B)}} y_i$ where $f_A = n_A/N_A$ and $f_B = n_B/N_B$ *i.e.* the frame sampling fractions. Since fixed weights usually differ from optimum weights, the SF estimator is in general (asymptotically) less efficient than the optimal estimators (Lohr and Rao, 2000). Furthermore the correct classification of sampled units into domains is still required. On the other hand, the SF estimator does not require to identify duplicate units since sampled units from the overlap domain are weighted by the same fixed coefficient; moreover it naturally extends to the MF setting. With this purpose the resourceful MF notation by Lohr and Rao (2005) is here extensively applied. A collection of $Q \geq 2$ overlapping frames $A_1 \cdots A_q \cdots A_Q$ is given, assuming that $\bigcup_q A_q = U$. Define the index sets *K* as subsets of the range of the frame index $q = 1 \cdots Q$. For every index set $K \subseteq \{1 \cdots q \cdots Q\}$ a domain is defined as $D_K = \left(\bigcap_{q\in K} A_q\right) \cap \left(\bigcap_{q\notin K} A_q^C\right)$ with $2^Q - 1$ different domains. For example, with $Q = 3$ there are 7 domains D_K denoted by the index sets $K = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. Define *Domain Membership Indicator* the random variable $\delta_i^{(D_K)}$ taking value 1 if $i \in D_K$ and 0 otherwise; then the population total, to be estimated, is expressed as a sum over the set of $2^Q - 1$ domains through the unit domain membership, *i.e.* $Y = \sum_K \sum_{i\in\bigcup_q A_q} \delta_i^{(D_K)} y_i$. Let s_q be the sample selected from frame A_q under a given design, independently for $q = 1 \cdots Q$. The (simple) SF estimator in the general case of MF is then given by

$\hat{Y}_{SF} = \sum_K \sum_{q \in K} \sum_{i \in s_q} w_i \delta_i^{(D_K)} y_i$ with fixed weights w_i ensuring design-unbiasedness. Under SRS of every frame it is given by:

$$\hat{Y}_{SF} = \sum_K \sum_{q \in K} \left(\sum_{q \in K} f_q \right)^{-1} \sum_{i \in s_q} \delta_i^{(D_K)} y_i \qquad (1)$$

As mentioned above, fixed weights are in general non-optimal. Hence in order to improve efficiency it has been proposed to correct the simple SF estimator via raking ratio by using the known frame sizes N_q. In the DF, Skinner (1991) derived the limiting (closed) form of the SF raking ratio estimator as the number *r* of the raking runs increases to infinity. In the MF framework the raked SF estimator $\hat{Y}_{SFrak}$ becomes computationally more complex: estimator (1) has to be corrected every $r \bmod(Q)$ raking run with respect to the size N_q of frames involved into every domain D_K, *i.e.* $\forall K \ni q$, iteratively until convergence. Note that since SF estimators simple or raked involve the domain membership indicator $\delta_i^{(D_K)}$, as appears for instance in equation (1), their application requires a known and correct classification of sampled units into domains. Besides the risk of misclassification, this implies to collect additional information since sampled units have to be asked, alongside about the survey variable, also about *to which* frames they belong (besides the one in which they have been sampled). This assumption can be removed by adopting a *multiplicity approach* to correct the SF estimator.

3. A MULTIPLICITY ESTIMATOR

The notion of multiplicity has been first introduced in connection with Network Sampling (Sirken, 2004; Casady and Sirken, 1980). It is also a tool of the Generalized Weight Share Methods (Lavallée, 2002) as well as of the Center Sampling estimation theory (Mecatti, 2004) since Center Sampling and MF frameworks are equivalent under certain conditions. By using a multiplicity approach, a design-unbiased MF estimator is proposed. Unlike the optimal and SF estimators, the multiplicity estimator does not depend on the domain membership of sampled units for requiring knowledge of *how many* frames they belong to instead of *to which* frames. As already mentioned this is a practical advantage for reducing the risk of misclassification and for reducing the amount of information asked to sampled units who might be somehow sensitive to the frame membership, for instance when sampling ex-prisoners, drug addicts, patients, illegal immigrants and so on.

In Lohr and Rao (2005), the multiplicity of domain D_K is defined as the cardinality of the index set *K*. Since domains are mutually exclusive, *i.e.* every unit *i* belongs to one and only one domain, the multiplicity is also a characteristic of every unit $m_i = \sum_q \delta_i^{(A_q)}$ where $\delta_i^{(A_q)}$ denotes the *frame membership indicator*, *i.e.* the random variable taking value 1 if $i \in A_q$ and 0 otherwise. Unit multiplicity m_i equals the number of frames in which unit *i* is included. Hence it can be observed by simply asking units *how many frames they belong to.* By using multiplicity the population total to be estimated is expressed as a sum over frames instead of a sum over domains:

$$\begin{aligned} Y &= \sum_K \sum_{i \in \bigcup_q A_q} \delta_i^{(D_K)} y_i \; m_i^{-1} \sum_{q=1}^{Q} \delta_i^{(A_q)} = \sum_{q=1}^{Q} \sum_{i \in \bigcup_q A_q} \delta_i^{(A_q)} y_i m_i^{-1} \sum_K \delta_i^{(D_K)} \\ &= \sum_{q=1}^{Q} \sum_{i \in A_q} y_i m_i^{-1} \end{aligned} \qquad (2)$$

Note that in equation (2) the domain membership indicator is not involved (for being $\sum_K \delta_i^{(D_K)} = 1, \; \forall i \in \bigcup_q A_q$). Equation (2) represents a practical advantage since domains are virtual objects while the sampling is actually performed into the *Q* frames. As a consequence a SF multiplicity estimator is given by:

$$\hat{Y}_{SFmulti} = \sum_{q=1}^{Q} \sum_{i \in s_q} w_i y_i m_i^{-1} \tag{3}$$

with fixed weights w_i ensuring, for instance, unbiasedness under the chosen frame designs, *i.e.* the inverse of the inclusion probability. For SRS we have $w_i = f_q^{-1}, \forall i \in s_q$. Owing to its Horvitz-Thompson structure, the exact variance of estimator (3) is given in closed form and hence easily estimated. For SRS of every frame the estimator variance is given by

$$Var\left(\hat{Y}_{SFmulti}\right) = \sum_{q=1}^{Q} \frac{N_q - n_q}{n_q\left(N_q - 1\right)} \left[N_q \sum_{i \in A_q} y_i^2 m_i^{-2} - \left(\sum_{i \in A_q} y_i m_i^{-1} \right)^2 \right]. \tag{4}$$

A customary unbiased variance estimator for SRS is then

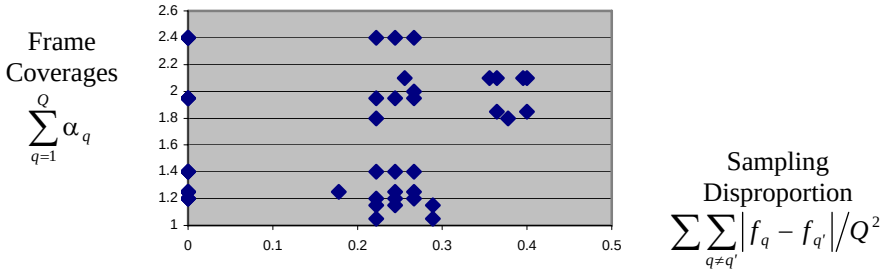
$$\hat{v}\left(\hat{Y}_{SFmulti}\right) = \sum_{q=1}^{Q} \frac{N_q\left(N_q - n_q\right)}{n_q^2\left(N_q - 1\right)} \left[N_q \sum_{i \in s_q} y_i^2 m_i^{-2} - f_q^{-1} \left(\sum_{i \in s_q} y_i m_m^{-1} \right)^2 \right]. \tag{5}$$

Estimator (3) has been first compared in the MF framework with direct SF competitors, namely the simple and raked SF estimators, under SRS of every frame. To start with, notice that while in the simple SF estimator as given by (1) unit values are weighted by $\left(\sum_{q \in K} f_q\right)^{-1} \delta_i^{(D_K)} = w_i(K)$ *i.e.* an *average* coefficient over the frames involved into each domain, in the multiplicity estimator unit values are weighted by $\left(f_q m_i\right)^{-1} = w_i(q)$ *i.e.* a *specific* frame coefficient. Thus they coincide for proportionate sampling of every frame, *i.e.* for constant sample fractions $f_q = f \ (q = 1 \cdots Q)$, while they differ for disproportionate sampling.

4. A SIMULATION STUDY

A simulation study has been performed in order to compare the proposed multiplicity estimator with the direct SF competitors and to investigate inferential properties for finite sample sizes.

Figure 1: *Simulated Populations*



Simulation concerns artificial populations of size N=120 randomly divided into Q=3 overlapping frames; the survey variable is provided with Discrete Uniform distribution between 1 and 100 and the population total Y is assumed as the parameter to be estimated. Different scenarios are produced by varying the frame coverage $\alpha_q = N_q / N$ under the constraint $\sum_q \alpha_q > 1$ in order to control the overlapping among frames, and by varying the sampling fractions $f_q = N_q / N$. In Figure 1 the simulated populations are represented as points in the plane formed by the two main simulation parameters: the sum of frame coverages on the vertical axis and a measure of dispersion among the

sampling fractions on the horizontal axis. As the dispersion of sampling fractions increases, increasing sampling disproportion is obtained. Hence by simulating along the vertical axis the case of proportionate sampling is investigated while all over the plane disproportionate sampling is simulated. In this study 45 populations/points are considered. For every population, a monte carlo distribution of the three SF estimators $\hat{Y}_{SF}$,$\hat{Y}_{SFrak}$ and $\hat{Y}_{SFmulti}$ is produced under SRS of each frame and their empirical mean and mean square error (*EMSE*) are calculated. The main objective of the simulation study is the evaluation of the proposed multiplicity estimator with respect to its relative efficiency so that the empirical efficiency ratio $EEff = Var\left(\hat{Y}_{SFmulti}\right) / EMSE\left(\hat{Y}\right)$ is provided. The ratio *EEff*

Table 1: *Proportionate sampling.*
Efficiency of $\hat{Y}_{SFmulti} \equiv \hat{Y}_{SF}$ *relative to* $\hat{Y}_{SFrak}$
Equal frame coverages

α_q	*EEff*
high 0.8	[0.97, 1.03]
low 0.4	[0.97, 1.03]

Table 2: *Proportionate sampling.*
Efficiency of $\hat{Y}_{SFmulti} \equiv \hat{Y}_{SF}$ *relative to* $\hat{Y}_{SFrak}$
Different frame coverages

α_q	***EEff***
0.2 0.4 0.8	[0.99, 1.14]
0.2 0.15 0.9	[1.09, 1.12]
0.8 0.9 0.25	[1.02, 1.08]

can result less, equal or greater than 1 indicating that $\hat{Y}_{SFmulti}$ is more, equally or less efficient than the compared estimator. The monte carlo error has been taken under control by accepting exclusively simulations giving a monte carlo relative bias less than 1% for estimators known to be unbiased,. This has required to keep the number of simulation runs between 32000 and 60000 in the worst cases.

Table 1 and 2 refer to some indicative simulation results in the case of proportionate sampling, when the multiplicity and the simple SF estimators coincide. In this case simulations regard the effects of the raking ratio correction over unbiasedness and efficiency. As a general indication, the raking ratio correction does not affect unbiasedness, since $\hat{Y}_{SFrak}$ results always unbiased or nearly so. We then focus on efficiency. Results in Table 1 show that the raking ratio correction does not improve the simple SF estimator for equal frame coverages. In other words, when the frame sizes are all equal $N_q = N/Q$, they do not give additional information and to rake becomes irrelevant. As the difference among frame coverages increases, some efficiency gains between 2% and 14% are observed. Over all cases explored for proportionate sampling, the raking ratio correction improves the simple SF estimator in the 60% of cases while giving the same efficiency in the remaining 40%. About 62 raking runs are needed, on average, to reach convergence. Tables 3 to 5 report the main simulation results in the case of disproportionate sampling, First the effects of the multiplicity correction are concerned by comparing the multiplicity estimator and the simple SF estimator. Table 3 refers to the case of equal frame coverages: $\hat{Y}_{SFmulti}$ is less efficient than simple SF estimator only in one case for high frame coverages while efficiency gains between 12% and 22% are registered in all the other cases. As a general indication, the multiplicity correction improves the simple SF estimator especially for decreasing frame coverages. For different frame coverages, as reported in Table 4, the major efficiency gains of multiplicity estimator over the simple SF estimator are registered, ranging from 12% to 73% as the sampling disproportion increases. As a general indication, multiplicity correction improves the simple SF estimator especially for increasing sampling disproportion.

Table 3: *Disproportionate sampling*
Efficiency of $\hat{Y}_{SFmulti}$ *relative to* $\hat{Y}_{SF}$
Equal frame coverages

α_q	f_q	*EEff*	%Efficiency gains
0.4	0.1 0.3 0.6	0.78	22%
	0.75 0.75 0.15	0.84	16%
0.6	0.9 0.05 0.5	0.79	21%
	0.2 0.2 0.75	0.88	12%
0.7	0.6 0.05 0.25	0.86	14%
0.8	0.2 0.2 0.75	1.14	-14%

Table 4: *Disproportionate sampling*
Efficiency of $\hat{Y}_{SFmulti}$ *relative to* $\hat{Y}_{SF}$
Different frame coverages

α_q	f_q	*EEff*	%Efficiency gains
0.2 0.15 0.9	0.9 0.01 0.81	0.27	73%
	0.75 0.75 0.15	0.81	19%
0.9 0.7 0.5	0.05 0.25 0.95	0.88	12%
	0.5 0.08 0.9	0.85	15%

Finally estimator $\hat{Y}_{SFmulti}$ has been compared with the raked estimator $\hat{Y}_{SFrak}$ as alternative ways of correcting the simple SF estimator. Since the multiplicity estimator requires less sample information than the raked estimator for using neither the domain membership of sampled units nor the frame sizes, some efficiency loss is expected. The interesting point is *how much* and *when.* With this respect simulation results are highly dispersed: Table 5 is an attempt to summarize them and to grasp any general indication. All the simulation results for disproportionate sampling have been classified on the basis of the empirical efficiency ratio into three classes: 1) *none* or *negligible* efficiency loss (less than 10%), 2) *slight* efficiency loss (between 10% and 20%) and 3) *severe* efficiency loss (more than 20%). Second column in Table 5 shows that none efficiency loss has been registered in 15% of cases with 0% loss on average (third column); in 23% of cases the efficiency loss is 18% on average and a severe efficiency loss (59% on average) appears in 62% of cases. The last three columns of Table 5 report the minimum, the maximum and the 75th quantile of the efficiency losses observed so that negative values indicate efficiency gains of the multiplicity correction with respect to the raking ratio correction. As a general indication, $\hat{Y}_{SFmulti}$ results equally efficient or slightly less efficient than $\hat{Y}_{SFrak}$ especially for decreasing frame coverages.

Table 5: *Disproportionate sampling*
Efficiency of $\hat{Y}_{SFmulti}$ *relative to* $\hat{Y}_{SFrak}$

Efficiency loss		Cases	Average	min	max	75th
none	< 10%	15%	0%	-4%	7	-1%
slight	10% to 20%	23%	18%	15%	20%	20%
severe	> 20%	62%	59%	21%	112%	75%

ACKNOWLEDGEMENTS

The author wishes to thank Jon N.K. Rao for useful discussion and resourceful suggestions.

REFERENCES

Bankier, M. D. (1986), "Estimators based on several stratified samples with applications to multiple frame surveys", *Journal of the American Statistical Association*, 81, pp. 1074-1079.

Casady, R. J., and Sirken, M. G. (1980), "A multiplicity Estimator for Multiple Frame Sampling", *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 601-605.

Fuller, W. A. and Burmeister, L. F. (1972), "Estimates for samples selected from two overlapping frames", *Proceedings of the Social Statistical Section, American Statistical Association*, pp. 245-249.

Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications", *Sankya*, C, 36, pp. 99-118.

Kalton, G. and Andersen, D. W. (1986), "Sampling rare populations", *Journal of Royal Statistical Society*, A, 149, pp. 65-82.

Lavallée, P. (2002), *Le Sondage Indirect ou la Méthode Généralisée du Partage des Poids*, Editions de l'Université de Bruxelles, Editions Ellipses.

Lepkowski, J. M. (1991), "Sampling the Difficult-to-Sample", *Proceedings of the Symposium of the American Institute of Nutrition*, pp. 416-423.

Lohr, S. L. and Rao, J. N. K. (2000), "Inference from dual frame surveys", *Journal of the American Statistical Association*, 95, pp. 271-280.

Lohr, S. L. and Rao, J. N. K. (2005), “Estimation in Multiple Frame Surveys”, *Journal of the American Statistical Association*, to appear.

Mecatti, F. (2004), “Center Sampling: a Strategy for Surveying Difficult-to-Sample Populations”, *Proceedings of the Methodology Symposium 2004*, Statistics Canada.

Sirken, M. G. (2004), “Network Sample Surveys of Rare and Elusive Populations: A Historical Review”, *Proceedings of the Methodology Symposium 2004*, Statistics Canada.

Skinner, C. J. (1991), “On the efficiency of Raking Ratio estimation for multiple frame surveys”, *Journal of the American Statistical Association*, 86, pp. 779-784.

Skinner, C. J. and Rao, J. N. K. (1996), “Estimation in dual frame surveys with complex designs”, *Journal of the American Statistical Association*, 91, pp. 349-356 .

Sudman, S. and Kalton, G. (1986), “New Developments in the Sampling of Special Populations”, *Ann. Rev. Sociol.*, 12, pp. 401-429.