ECOLOGICAL PANEL INFERENCE IN REPEATED CROSS SECTIONS

ECONOMETRIC INSTITUTE REPORT EI 2002-22

BEN PELZER[1], ROB EISINGA[1], AND PHILIP HANS FRANSES[2]

ABSTRACT. This paper presents a Markov chain model for the estimation of individual-level binary transitions from a time series of independent repeated cross-sectional (RCS) samples. Although RCS samples lack direct information on individual turnover, it is demonstrated here that it is possible with these data to draw meaningful conclusions on individual state-to-state transitions. We discuss estimation and inference using maximum likelihood, parametric bootstrap and Markov chain Monte Carlo (MCMC) approaches. The model is illustrated by an application to the rise in ownership of computers in Dutch households since 1986, using a 13-wave annual panel data set. These data encompass more information than we need to estimate the model. This additional information allows us to assess the validity of the parameter estimates. Software implementing the model is available.

Paper prepared for the Ecological Inference Conference, Harvard University, Cambridge, MA, June 17-18, 2002.

1. INTRODUCTION

It has sometimes been argued that King's ecological inference model can be adapted and fruitfully applied to independent repeated cross-sectional (RCS) samples (e.g., Penubarti and Schuessler 1998, King, Rosen and Tanner 1999). To date, however, surprisingly little research has been devoted to the development of cross-level inference models that draw panel conclusions from non-panel data.[1] The objective of this paper is to address this issue.

[1]Department of Social Science Research Methods, University of Nijmegen, The Netherlands
[2]Econometric Institute, Erasmus University Rotterdam, The Netherlands

Authors' note. The data for *Socio-Economic Panel* used in this paper were collected by STATISTICS NETHERLANDS and were made available by the SCIENTIFIC STATISTICAL AGENCY of the NETHERLANDS ORGANIZATION FOR SCIENTIFIC RESEARCH. Our program *CrossMark* implements all the simulations and estimations reported here. It is programmed in Delphi but distributed as a stand-alone program running under Windows. The program (including documentation) is free software and available upon request.

[1] The only studies we were able to locate that are directly related to this issue include Franklin (1989), Moffitt (1990 1993), Sigelman (1991), Mebane and Wand (1997), and Penubarti and Schuessler (1998). The framework presented here has originally been proposed by Moffitt (1990 1993). Pelzer et al. (2002) discuss the dissimilarities between this model and the ecological panel inference (EPI) method of Penubarti and Schuessler (1998) and the two-stage auxiliary instrumental variables (2SAIV) approach of Franklin (1989).

One reason for this is the lack of genuine panel data. Panel designs are, rightfully, highly regarded for the opportunity they offer to measure transitions of state or value from repeated observations on the same sample units. For many research issues, however, adequate panel data are hard to collect or simply unavailable. Another major reason is that panel data are potentially subject to non-sampling biases. An important one is sample attrition that results from the progressive loss of (often selective groups of) respondents willing to participate in the data collection. While non-response is also a limitation for cross-sectional surveys, it is a more serious problem for panel data because non-response often accumulates over time. A related limitation is that it is often difficult to ensure that changes in the target population are reflected in the panel. While panels are typically designed to be representative of the population at the beginning of the study, the panel ages over time and few panels are, in addition to providing longitudinal data, also designed to premanently provide fully representative information of the population by continuous renewal of the sample.

A large number of cross-sectional surveys conducted by public and private organizations are repeated at regular time intervals. These repeated cross-sectional surveys do not suffer from panel mortality and reflect changes in the universe that cannot be accounted for by a panel study. Estimating individual transitions from such data has the connotation of performing an impossible task, of obtaining information from nowhere. Indeed, it is often argued that panel data are necessary to study individual-level change (e.g., Kish 1987, p. 167). While individual change is obviously only *visible* in panel data, this paper will show that this argument is not correct and that data from successive, separately drawn samples can be used to validly estimate transitions using a model that is no more magical than the use of 'plug-in' estimates and bridging assumptions in other areas of statistical modelling.

The outline of the paper is as follows. Section 2 presents a Markov transition model for repeated cross sections designed to deal specifically with binary responses. This model has its origins in the work of Moffitt (1990 1993). We briefly review its main features and discuss maximum likelihood (ML), parametric bootstrap and Markov chain Monte Carlo (MCMC) approaches to estimation and inference. Sections 3 considers an application of the model to the rise in computer penetration rates in Dutch households from 1986 to 1998, using annual panel data from the *Socio-Economic Panel* (*SEP*) survey of STATISTICS NETHERLANDS. We examine the determinants of the transitions from 'have-not' to 'have' (and back again) using well-known socio-economic and demographic covariates of the digital divide. Parametric bootstrap and Bayesian simulation are used to evaluate the accuracy and the precision of the RCS Markov ML estimates and the results are also compared with those of a first-order dynamic panel model. To mimic genuine RCS data, we additionally analyze samples of independent observations randomly drawn from the panel. The summary in Section 4 concludes the paper.

2. Estimating binary transitions from rcs data

2.1. **Binary transition model.** Obviously, the estimation of dynamic models with repeated cross-sectional data is hampered by the lack of information about lagged variables. Let y_{it} denote the observed response for the binary random variable y of unit i at time period t. The crucial characteristic of RCS data is that y_{it} is observed, but y_{it-1} is not. Consequently, no estimate of the serial covariance of successive y_{it} necessary to estimate dynamic models is available in RCS data. This does not imply that dynamic models cannot be estimated with repeated cross sections. However, it does imply that estimation of the unobserved transitions is possible only by putting certain constraints on the transitions for unit i and/or time period t.

Consider a 2×2 transition table in which the internal cell values sum to unity across rows. If we define $p_{it} = P(y_{it} = 1)$, $\mu_{it} = P(y_{it} = 1|y_{it-1} = 0)$, and $\lambda_{it} = P(y_{it} = 0|y_{it-1} = 1)$, then we have the well-known accounting equation

$$E(y_{it}) = p_{it} = \mu_{it}(1 - p_{it-1}) + (1 - \lambda_{it})p_{it-1} \tag{1}$$

This identity is the critical equation that needs to be solved in estimating dynamic models with repeated cross sections, as it relates the marginal probabilities (p_{it} and p_{it-1}) to the entry (μ_{it}) and exit (λ_{it}) transition probabilities.[2] A more concise form for the same equation is $p_{it} = \mu_{it} + \eta_{it}p_{it-1}$, so that $\eta_{it} = 1 - \lambda_{it} - \mu_{it}$. It is also sometimes convenient to define $\kappa_{it} = 1 - \lambda_{it} = P(y_{it} = 1|y_{it-1} = 1)$. If we recursively substitute in for p_{it} in (1), and derive its reduced form in terms of past μ_{it} and λ_{it}, then we get

$$p_{it} = \mu_{it} + \sum_{\tau=1}^{t-1}\left[\mu_{i\tau}\prod_{s=\tau+1}^{t}\eta_{is}\right] + p_{i0}\prod_{\tau=1}^{t}\eta_{it} \tag{2}$$

This is the model equation that will be used in this paper. It is obviously not uniquely solvable with RCS data without identifying constraints. Two types of restrictions may be used in this context. One is to impose some direct restraint on the patterns of μ_{it} and λ_{it}. For example, the parameters in (2) are clearly identifiable with RCS data if we assume that the transition probabilities are homogeneous with respect to both units i and time periods t. With the assumption that $\mu_{it} = \mu$ and $\lambda_{it} = \lambda$, for all i and t, the long-run value of p_{it} in (2) reduces to $p_{it} = \mu/(\mu + \lambda)$ (see, e.g., Ross 1993, pp. 152-153). Another type of restriction may be imposed if the cross-sectional data include covariates $\mathbf{x}_{it}$ that are measurable in the past (by 'backcasting'), and if the current and lagged $\mathbf{x}_{it}$ affect μ_{it} and λ_{it}. In that case, the covariates $\mathbf{x}_{it}, \mathbf{x}_{it-1}, \ldots, \mathbf{x}_{i1}$ can be employed to obtain current and backward predictions of the entry ($\mu_{it}, \mu_{it-1}, \ldots, \mu_{i1}$) and exit ($\lambda_{it}, \lambda_{it-1}, \ldots, \lambda_{i2}$) transition probabilities, by specifying

$$\mu_{it} = F(\mathbf{x}_{it}\beta) \text{ and } \lambda_{it} = 1 - F(\mathbf{x}_{it}\beta^*) \tag{3}$$

where β and β^* are two potentially different sets of parameters associated with two potentially different sets of (time-invariant or time-varying) covariates $\mathbf{x}_{it}$, and F is the - in this paper logistic - link function. Estimates of the model parameters are obtained by substituting (3) into (2). The critical identifying restriction used here is that the regression parameters are taken to be constant over time, but this

[2] The non-panel nature of cross-sectional data could be made explicit by indexing individual units in cross-section t by $i(t)$ and the marginal probabilities, for example, by $p_{i(t)t}$. However, to ease the notation we shall simply write i for $i(t)$ and p_{it} for $p_{i(t)t}$, as the nature of the data is obvious.

assumption may easily be relaxed if we have a sufficient number of repeated cross sections. We can model the parameters as a function of time using polynomials or splines, for example, or we may use a semi-parametric approach that assumes the parameters to be constant within but different across discrete time periods. Note that the underlying Markov chain is not assumed to be homogeneous here, implying that the entry and exit transition probabilities may vary across both units i and time periods t. Also note that to obtain p_{it}, we actually integrate (sum) over all possible unobserved state-to-state transition paths for each individual unit i, starting at $t = 1$ and ending at the cross-sectional observation period t. This implies that the probabilities are estimated as a function of all the available cross-sectional samples, rather than simply the observations from the current time period.

Other perhaps more implicit assumptions underlying the application of model are that $p_{i0} = 0$, that all the covariates $\mathbf{x}_{it}$ included in the model should have known values in the past, and that the estimation of the entry and exit transitions depend exclusively on variations in the covariates observed. With respect to the first assumption, it should be noted that p_{i0} is not the first observed outcome (which is p_{i1}), but rather the value of the state prior to the start of the Markov chain. It is generally difficult to incorporate the initial state into the model, since this requires assess to the Markov process from the beginning for each individual unit i. We therefore invoke the restriction that $p_{i1} = 0$, the consequence of which is that $p_{i1} = \mu_{i1}$. Because in many applications the latter assumption is untenable, we define a separate logistic function for the cross section at $t = 1$, i.e., $P(y_{i1} = 1) = F(\mathbf{x}_{it}\delta)$. The δ-parameters are estimated simultaneously with the entry and exit parameters of interest at $t = 2, \ldots, T$, and they are estimated as a function of all cross-sectional data, rather than simply the observations at $t = 1$.

If some of the covariates are 'non-backcastable' (i.e., their past history is unknown), the model may be modified by estimating two different sets of parameters for both μ_{it} and λ_{it}: one for the current transition probability estimates and a separate one for the preceding estimates. If we denote the time-dependent covariate with unknown past history by $\mathbf{v}_{it}$ and the associated parameter vector representing the effect on μ_{it} by ζ, then we have $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta^{**} + \mathbf{v}_{it}\zeta$ for cross section t, and $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta$ for the cross sections $1, \ldots, t-1$. This specification allows one to express the current transition probability estimates as a logistic function of both backcastable and non-backcastable variables. A similar model may be specified for λ_{it}. It should be noted here that in our application below we assume that $\beta^{**} = \beta$.

If the assumption that all relevant variables are included in the model is not a realistic one, it may be useful to include an individual-specific random error term ε_i in the linear predictor of the transition probabilities to account for omitted variables, at least in so far as these variables are time-invariant for each individual. In this logistic-normal mixture model we have $\text{logit}(\mu_{it}) = \mathbf{x}_{it}\beta + \gamma_0\varepsilon_i$ and $\text{logit}(1 - \lambda_{it}) = \mathbf{x}_{it}\beta^* + \gamma_1\varepsilon_i$, where γ_0 and γ_1 are coefficients of the random variable ε_i having zero mean and unit variance. To estimate the parameters, the marginal likelihood of this model is integrated with respect to the distribution of ε_i using, for example, the Gauss-Hermite quadrature approximation. As unobserved heterogeneity is not examined in the empirical application below, we will not elaborate on this topic here. Pelzer et al. (2002) provide further details.

Finally, equation (1) may be rearranged into $\mu_{it} = p_{it}/(1 - p_{it-1}) - p_{it-1}/(1 - p_{it-1})\kappa_{it}$ where $\kappa_{it} = 1 - \lambda_{it}$. This expression resembles the equation that King

(1997) termed 'tomography line'. Since the estimated marginal probabilities p_{it} and p_{it-1} are guaranteed to lie in the (0,1) range, bounds are enforced on the maximum likelihood estimators of μ_{it} and κ_{it}. These bounds are not informative as in the 'methods of bounds' (Duncan and Davis 1953), however, but rather logical limits implied by the model. The methods of bounds obtains ranges of feasible entries that are consistent with the observed margins in the 2 by 2 table. Together the two margins provide (at least some) information on the internal cells. In the analysis of individual data from repeated cross sections, however, only one of the margins (y_{it}) is observed, and the other one (y_{it-1}) is not. This implies that the repeated cross sections cannot provide any deterministic, informative restrictions on the entries (unless one aggregates the micro data into profiles defined by covariates, as in Penubarti and Schuessler 1998). Consequently, the ecological inference problem in RCS data is greater (in the sense of a larger number of unknowns) than in the more common application where the margins are known. The approach proposed here is to completely express the marginal probabilities p_{it} in terms of μ_{it} and κ_{it}, recursively, so that estimating the latter automatically renders the former.

2.2. **Estimation and simulation.**

2.2.1. *Maximum likelihood estimation.* The method of maximum likelihood may be used to compute the estimates of the parameters in (3) - plugged into (2) - and their variances. For a sample of n statistically independent observations - were each observation is treated as a single draw from a Bernoulli distribution - with success probability p_{it}, model (2) has the log likelihood function

$$LL = \sum_{t=1}^{T}\sum_{i=1}^{n_t} \ell\ell_{it} = \sum_{t=1}^{T}\sum_{i=1}^{n_t} [y_{it}\log(p_{it}) + (1-y_{it})\log(1-p_{it})]. \tag{4}$$

were T is the number of cross sections and n_t the number of units of the cross-sectional sample at time period t. Maximization of this function has to be performed iteratively and requires the derivatives of the log likelihood with respect to the parameters, θ, say. If we suppress subscript i to ease notation, the first order derivatives with respect to θ are

$$\frac{\partial \ell\ell_t}{\partial\theta} = \frac{y_t - p_t}{p_t(1-p_t)} \cdot \frac{\partial p_t}{\partial\theta},$$

were

$$\frac{\partial p_t}{\partial\theta} = \frac{\partial\mu_t}{\partial\theta} + \frac{\partial p_{t-1}}{\partial\theta}\eta_t + p_{t-1}\frac{\partial\eta_t}{\partial\theta}.$$

If θ is used to estimate μ_t, then $\partial\mu_t/\partial\theta = \mathbf{x}_t\mu_t(1-\mu_t)$ and $\partial\eta_t/\partial\theta = -\partial\mu_t/\partial\theta$. If it is used for λ_t, then $\partial\mu_t/\partial\theta = 0$ and $\partial\eta_t/\partial\theta = \mathbf{x}_t\lambda_t(1-\lambda_t)$. The values for $\partial p_t/\partial\theta$ can be obtained by recursive substitution, setting $\partial p_0/\partial\theta = p_0 = 0$, and starting from $\partial p_1/\partial\theta = \partial\mu_1/\partial\theta = \mathbf{x}_1\mu_1(1-\mu_1)$. The second order derivatives are

$$\frac{\partial^2\ell\ell_t}{\partial\theta\partial\theta'} = -\frac{(y_t-p_t)^2}{p_t^2(1-p_t)^2}\cdot\frac{\partial p_t}{\partial\theta}\cdot\frac{\partial p_t}{\partial\theta'} + \frac{y_t-p_t}{p_t(1-p_t)}\cdot\frac{\partial^2 p_t}{\partial\theta\partial\theta'},$$

were

$$\frac{\partial^2 p_t}{\partial\theta\partial\theta'} = \frac{\partial^2 p_{t-1}}{\partial\theta\partial\theta'}.\eta_t + \frac{\partial p_{t-1}}{\partial\theta'}.\frac{\partial\eta_t}{\partial\theta} + \frac{\partial^2 \mu_t}{\partial\theta\partial\theta'}.(1 - p_{t-1}) - \frac{\partial\mu_t}{\partial\theta'}.\frac{\partial p_{t-1}}{\partial\theta}.$$

If θ belongs to $\mathbf{x}_t$ and θ' to $\mathbf{x}_t^*$, then $\partial^2\mu_t/\partial\theta\partial\theta' = \mathbf{x}_t\mathbf{x}_t^*\mu_t(1-\mu_t)(1-2\mu_t)$. Again, if we set $\partial^2 p_0/\partial\theta\partial\theta' = \partial p_0/\partial\theta = \partial p_0/\partial\theta' = p_0 = 0$, the values for $\partial^2 p_t/\partial\theta\partial\theta'$ can be obtained recursively, starting from $\partial^2 p_1/\partial\theta\partial\theta' = \partial^2\mu_1/\partial\theta\partial\theta'$.

The parameter estimates may be obtained by Newton's method, which uses the Hessian matrix of the actual second derivatives. To speed up computation, we may avoid calculating the exact Hessian by approximating it instead by the expected second derivatives, and use Fisher's method-of-scoring. This method is used here. In addition to providing parameter estimates, the Fisher optimization algorithm produces as a by-product an estimate of the asymptotic variance-covariance matrix of the model parameters, given by the inverse of the estimated information matrix evaluated at the converged values of the estimates. If the estimates are MLE, each element of the inverse of the information matrix is a minimum variance bound for the corresponding parameter and the positive square root of the diagonal elements of this matrix (i.e., the standard errors of the estimated coefficients) may be used for significance tests and to construct confidence intervals.

Finally, according to asymptotic theory, ML estimators become progressively more unbiased, more normally distributed and achieve a minimum possible variance more closely as the sample size increases. However, these asymptotic assumptions may be violated by the nature of our relatively complex Markov chain model. Moreover, the estimator in our model has essentially unknown properties for small to moderate sample sizes and we cannot present any guidelines as to when a sample is sufficiently large for the asymptotic properties to be closely approximated. It is therefore important to investigate the behavior of the estimators of the parameters in (2) by examining their finite-sampling distribution. The bootstrap and MCMC simulations provide useful tools in this situation.

2.2.2. *Parametric bootstrap simulation.* Bootstrap uses Monte Carlo simulation to empirically approximate the probability distribution of the parameter estimates and other statistics rather than relying on assumptions about its shape that may only be asymptotically correct.[3] The technique used here is model-based parametric bootstrap (Davison and Hinkley 1997). For the parametric bootstrap, re-samples are taken from the original data via a fitted parametric model to create replicate data sets, from which the variability of the quantities of interest can be assessed. In the repeated simulations, it is assumed that both the form of the deterministic component of the model and the nature of the stochastic component are known. Bootstrap samples are generated using the same fixed covariates as in the original sample and a set of predetermined values for the parameters, allowing only the stochastic component to change randomly from sample to sample. By this means, many bootstrap samples are generated, each of which provides a set of estimates of the parameters that may then be examined for their bias, variance, and other

[3] The examination of the sampling distribution of the estimators will be restricted to studying the marginal distribution of each parameter separately. Studying the multivariate behavior of the estimators is a more complicated problem we shall not undertake here.

distributional properties and used for bootstrap confidence intervals and hypothesis testing. The parametric bootstrap re-sampling procedure is implemented here according to the following algorithm.

(1) Estimate the unknown parameter, θ, according to model (2) using the original sample, $\{x_{it}, y_{it}\}, i = 1, \ldots, n_t, t = 1, \ldots, T$, with the estimate denoted as $\hat{\theta}$, and obtain the fitted values $\hat{p}_{it}$ of the probability that the binary dependent variable $y_{it} = 1$.
(2) For each x_{it} in the original sample, $\{x_{it}, y_{it}\}$, generate a value of the bootstrap dependent variable y^*_{it} by random sampling from a Bernoulli distribution with success probability given by $\hat{p}_{it}$.
(3) Use the bootstrap sample, $\{x_{it}, y^*_{it}\}$, to fit the parameter estimate θ^*.
(4) Repeat Steps 2 and 3 R times, yielding the bootstrap replications denoted as $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_R$. The empirical distribution of these replications is used to approximate the finite sample distribution of $\hat{\theta}$.

In this study we look at the density of the values of $\hat{\theta}^*$ under re-sampling of the fitted model to examine bias and variance and to see if it is multi-modal, skewed, or otherwise differs from normality. To obtain an accurate empirical approximation, we use $R = 5,000$ replications of the original data set. While the bootstrap estimates of bias and variance under the fitted model are important in their own right, parametric re-sampling may also be useful in testing problems when standard approximations do not apply or where the accuracy of the approximation is suspect. The key to applying the bootstrap for hypothesis testing is to transform the data so that the null hypothesis is true in the bootstrap population. That is, we simulate data under the null hypothesis so that bootstrap re-sampling resembles sampling from a population for which the null hypothesis holds (Hall and Wilson 1991). To be specific, the bootstrap hypothesis test compares the observed value in the original sample to the R values $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_R$, which are obtained from samples independently generated under the null model that satisfies H_0. The bootstrap P-value may then be obtained by $p^*(\hat{\theta}) = P(\hat{\theta}^* \geq \hat{\theta}|H_0) = R^{-1} \sum_{i=1}^{R} I(\theta^* \geq \hat{\theta})$, were the indicator $I(.)$ equals one if the inequality is satisfied and zero if not (Davison and Hinkley 1997). We reject the null hypothesis if the selected significance level exceeds $p^*(\hat{\theta})$.

2.2.3. *Markov chain Monte Carlo simulation.* Another powerful tool next to MLE and parametric bootstrap is Bayesian simulation, which is easily implemented using Markov chain Monte Carlo (MCMC) methods. Bayesian data analysis is not concerned with finding the parameter values for which the likelihood reaches the global maximum. It is primarily concerned with generating samples from the posterior distribution of the parameters given the data and a prior density and this distribution may be asymmetric and/or multi-modal. Other advantages of the Bayesian approach include the possible incorporation of any available prior information and the ability to make inferences on arbitrary functions of the parameters or predictions concerning specific individual units in the sample. A popular method for MCMC simulation is Metropolis sampling (Tanner 1996). The Metropolis sampler obtains a chain of draws from the posterior multivariate distribution, $\pi(\theta|y)$, of the parameter θ. In sampling from the unknown target distribution, the algorithm uses a known auxiliary density A – e.g., a (multivariate) uniform or normal distribution - to select candidate parameters θ^c. The Metropolis algorithm proceeds as follows

(1) Choose a start value for the parameter (e.g., the MLE).
(2) Randomly draw parameter θ^c from A, were A is a symmetric proposal distribution with mean equal to the previous draw, θ, and an arbitrary variance.
(3) If $\pi(\theta^c|y) \geq \pi(\theta|y)$, add candidate θ^c to the chain of draws. If $\pi(\theta^c|y) < \pi(\theta|y)$, calculate the ratio $r = \pi(\theta^c|y)/\pi(\theta|y)$ and add candidate θ^c with probability r to the chain of draws.
(4) If candidate θ^c is not added to the accepted draws in Step 3, add θ so that two successive elements of the chain have the same parameter value θ. Else, proceed with the next step.
(5) Repeat Steps 2-4 K times, yielding a sample from the posterior distribution of θ.

In the Markov chain sampling used here, we assumed a priori that we are ignorant about the values of the parameters. That is, no value of the parameter is any more probable than any other value (i.e., a vague prior belief). This implies that $\pi(\theta|y)$ equals the likelihood of parameter θ. Once stationarity has been achieved, a value from a chain of draws from the Metropolis algorithm is supposed to have the same distribution as the target density. We run the Metropolis algorithm $K = 100,000$ times, excluding an initial burn-in of $10,000$ samples, and subsequently obtained the mean, standard deviation, and limits of the 95% credibility interval of θ.

3. Application

3.1. PC penetration in Dutch households.

The major concern of this section is how the RCS Markov model performs in practice. The empirical application is concerned with modelling the rise in computer penetration rates in Dutch households in the 1986-98 period using data from the *Socio-Economic Panel (SEP)* collected by STATISTICS NETHERLANDS. The reason for using this 13-wave annual household panel study is that it offers the opportunity to check the estimation results against the panel findings. However, it is important to note that in the RCS Markov analysis below the panel data are treated as if they were observations of a temporal sequence of 13 independent cross-sectional samples. That is, no use is made of information about lagged values of y_{it}.

The binary dependent variable y_{it} is defined to equal one if the household owns a personal computer and zero if not. Table 1 reports the proportions of Dutch households with a PC in the 1986-98 period along with the observed entry and exit transition rates. As can be seen, there is a marked upward time trend in PC ownership, from 12% in 1986 to 57% in 1998. While the entry rates, $(\bar{y}_t|y_{t-1} = 0)$, also show an increase over time, the exit rates, $((1 - \bar{y}_t)|y_{t-1} = 1)$, reveal erratic change.

Table 1 goes here

The most important structural determinants of the presence of a PC in homes - in the Netherlands as elsewhere - include educational attainment, household income, the size of the household, and age (see, e.g., OECD 2001). These variables are included in the *SEP* household study, but they would generally also be available in a repeated cross-sectional survey. The time-varying variable age of head of household (hereafter: age) is categorized into three different age categories (18-34, 35-54, and 55+ years of age). The time-varying variable number of household members is constructed from cross-sectional information about the number and the ages of the children in the household and the presence of a spouse. It is assumed that a family with children has two adults. The variable highest completed education of head of household (hereafter: education) is taken to be fixed over time. Next to these backcastable variables, the analysis also includes the temporary, nonbackcastable covariate household income. The variable used here is the standardized (i.e., corrected for size and type of household) disposable household income, categorized into quintiles.

3.2. **RCS Markov model.**

3.2.1. *Maximum likelihood.* The first model fitted was a time-stationary Markov chain with constant terms only. This model produces the parameters $\beta(\mu_t) = -2.543$ and $\beta^*(\lambda_t) = -3.310$ and a log-likelihood value of $LL = -15895.214$. These estimates imply constant transition probabilities $\mu = .073$ and $\lambda = .035$, hence predicted rates that underestimate the observed sample frequencies reported in Table 1. The model was subsequently modified to a non-stationary, heterogeneous Markov model by adding the covariates reported above. In analyzing the data with this model, it became apparent that the covariates have a substantial effect on the transition from have-not to have, but that they contribute little to the explanation of the reverse transition. We therefore decided to model the exit transitions using a single constant term only. Further, it turned out that the inclusion of a linear time trend in the prediction of obtaining a computer appreciably improves the fit. We therefore included the variable time in the model. The results are reported in the second column of Table 2.

Table 2 goes here

The top part of the table gives the estimated effects on the marginal probabilities p_{i1}. The table indicates that both education and the number of household members positively affect the presence of a PC in homes. While there is no significant difference in PC ownership between the 18-34 year age group and others aged 35-54, ownership is significantly more widespread among the younger age group than among those aged 55 and over. The middle part of Table 2 presents the effects on the transition from have-not to have with respect to PC ownership. The results show that educational attainment of head of household, household size, household income, and time have a positive effect on obtaining a computer. This finding confirms the conclusion of cross-sectional studies that computer ownership has spread most rapidly among the affluent, well educated families with children (OECD 2001). The coefficients of the age terms again imply similar entry rates among younger and middle age groups. The older age group has considerably lower access rates. The parameter estimate of the constant term for λ_{it} is shown in the bottom part of the table. An intercept of -2.292 implies a time-constant exit transition probability of $\lambda = .092$ (i.e., $\kappa = .908$), which perfectly matches the observed mean frequency of .092.

3.2.2. *Parametric bootstrap.* As indicated, the benefit of parametric simulation is that the bootstrap estimates give empirical evidence that likelihood theory can be trusted, while providing alternative methods for calculating measures of uncertainty if this theory is unreliable. To examine the sampling distribution of the parameter estimates, we generated $R = 5,000$ bootstrap samples according to the algorithm given in Section 2.2.2. Table 2 provides for each parameter the mean and the sample standard deviation of the bootstrap estimates. It is typical of likelihood methods that the variability of likelihood quantities is underestimated. As the table shows, however, the effect is small enough to be unimportant here. The bootstrap mean values are close to the ML estimates and the sample standard deviations are similar to the likelihood-based standard errors. The bootstrap estimates of bias and other distributional properties are given in Table 3.

Table 3 goes here

The ML estimates of the model parameters appear to be only slightly biased, with the largest absolute bias being .0086. When the estimated bias is expressed as a percentage of the parameter estimate (not reported in Table 3), the largest differences between standard theory and the bootstrap results are found for the $\delta(p_{i1})$ parameter of the age 35-54 dummy, for which the percentage bias is 1.85%. All other parameters have percentage biases of less than 1%. The parameters also tend to have a small bias compared to the magnitude of their standard deviation. A frequently applied rule of thumb is that a good estimator should be biased by less than 25% of its standard deviation (Efron and Tibshirani, 1993). As can be seen in Table 3, the ratios of estimated bias to standard deviation are all less than .25. Small values are also found for the root mean square error, which takes into account both standard deviation and bias. The bootstrap sample variance may be compared to the estimated ML variance using a chi-square test to examine whether the sample variance from the bootstrap is significantly larger than the variance from ML (Ratkowsky 1983). For none of the parameters is the bootstrap variance significantly in excess of the ML variance. The largest value was again found for the $\delta(p_{i1})$ parameter of the age 35-54 dummy. The statistic $\chi^2 = (N-1)(\hat{\sigma}^2_{bootstrap}/\hat{\sigma}^2_{ML})$ is distributed as chi-square with 4,999 degrees of freedom (df), a transform of which may be closely approximated by the standard normal distribution yielding, for this dummy variable, $z = \sqrt{2\chi^2} - \sqrt{2(df)-1} = 1.857$. Table 3 also reports skewness, excess kurtosis and the Jarque-Bera (1980) statistic, that may be used to test whether the estimators are normally distributed. The null hypothesis of normality is only rejected for the constant and the age 55+ parameter of $\delta(p_{i1})$, and for the constant term parameter of $\beta^*(\lambda)$. The distribution of the latter is somewhat peaked and all three estimates have an extended tail to the left. The normal approximation is least accurate for the $\beta^*(\lambda)$ constant. However, even for this estimate the deviation from normality is negligible. The same goes for the distribution of κ $[= 1/(1 + \exp(\beta^*(\lambda)))]$, shown in the left panel of Figure 1.

Figure 1 goes here

No obvious visual departure is apparent in the histogram of the κ estimates from that expected for a normally distributed random variable.

3.2.3. *Markov chain Monte Carlo.* The Metropolis sampler posterior estimates for each parameter are reported in Table 2. The findings are based on $K = 100,000$ samples, excluding 10,000 samples for initial settling. Inspection of the posterior means reveals that there are no gross discrepancies in magnitude compared to the ML estimates. The MCMC standard deviations and the ML standard errors are also similar to one another. The same goes for the .95 percentile intervals of the parametric bootstrap estimates and the Bayesian credibility intervals. Thus Bayesian and frequentist methods for obtaining estimates produce roughly similar results.

In sum, according to both parametric bootstrap and MCMC simulations, the maximum likelihood estimators in this application are almost unbiased, with a variance close to the minimum variance bound, and a distribution close to that of a normal distribution. This implies that the ML point estimates of the parameters are accurate and that the inverse of the Fisher information matrix may be used as a good estimate of the covariance matrix of the parameter estimates.

3.3. **Dynamic panel model.** It is compelling to compare the RCS Markov ML estimates with the corresponding parameter estimates of a dynamic panel model that allows for first-order dependence. Most directly related to the RCS Markov model is a panel model that specifies a separate logistic regression for $P(y_{it} = 1|y_{it-1} = 0, 1)$, and includes y_{it-1} as an additional predictor. This model can conveniently be written in a single equation as $\operatorname{logit} P(y_{it} = 1|\, y_{it-1} = 0, 1) = \mathbf{x}_{it}\beta + y_{it-1}\mathbf{x}_{it}\alpha$, where $\alpha = \beta^* - \beta$ (see Amemiya 1985, Diggle, Liang and Zeger 1994, and Beck et al., 2001). The results of applying this logistic model to the binary panel data are shown in the right most columns of Table 2. A comparison of the RCS Markov and panel estimates indicates that most of the findings are insensitive to choice of model. The point estimates of all parameters, except perhaps the coefficients for age 35-54 and those for income, are rather similar and the standard errors also correspond. This implies that inferences about the parameters do not change considerably with the choice of model. Moreover, the two models predict equivalent transition probabilities μ_{it} and λ_{it} for all individual cases (not reported), and the accuracy of the two models as judged by a ROC curve analysis is almost identical (curve not shown either).[4] The RCS Markov model is only clearly inferior to the panel model with respect to the likelihood. It should be noted, however, that the two models differ in the computation of p_{it} and thus the likelihood. In binary panel data, the marginal probability p_{it} is either μ_{it} or $(1 - \lambda_{it})$, conditional on y_{it-1}, and the likelihood contribution can be written as $\ell_{it} = \mu_{it}^{y_{it}(1-y_{it-1})}(1 - \lambda_{it})^{y_{it}y_{it-1}}(1 - \mu_{it})^{(1-y_{it})(1-y_{it-1})}\lambda_{it}^{(1-y_{it})y_{it-1}}$. In the RCS Markov model, however, the marginal probability p_{it} is always a weighted sum of two probabilities - μ_{it} and λ_{it} - weighted by p_{it}, and the likelihood is given by $\ell_{it} = [\mu_{it}(1 - p_{it-1}) + (1 - \lambda_{it})p_{it-1}]^{y_{it}}\, [(1 - \mu_{it})(1 - p_{it-1}) + \lambda_{it}p_{it-1}]^{1-y_{it}}$. This implies that even if panel and RCS data produce identical transition probabilities μ_{it} and λ_{it}, the two likelihood functions may differ because of p_{it-1}. The likelihood

[4] The area under the ROC curve for $y_t|y_{t-1} = 0$ observations is .763 for the RCS Markov model and .768 for the panel model.

values are only identical if p_{it-1} is equal to y_{it-1}; that is, if the lagged covariates perfectly predict the previous response.

3.4. **Samples of independent observations.** As indicated, in the RCS Markov model the panel data are treated as independent cross sections implying that there is no information on the autocov (y_{it}, y_{it-1}) available in the data file used for analysis. Nevertheless, the best way to make sure that the results are not artificial findings is to analyse independent observations. To do so, we randomly draw (without replacement) samples of 2,028 different households from the (2,028 x 13=) 26,364 panel observations, were each sample consists of 13 separate sets - one for each time period - of 156 households. Hence each household is selected only once in the 'cross-sectional' sample. The total number of possible 'cross-sectional' samples in the our application is approximately $10^{2,242}$.[5] We randomly draw 5,000 samples and analyzed each data set separately using maximum likelihood estimation.

Table 4 goes here

Table 4 reports the average value of the parameters across the samples along with the standard deviation divided by $\sqrt{13}$. A comparison of the Tables 2 and 4 suggests that for almost all parameters the mean values are close to the MLE obtained for the original full sample size. The only noticeable difference is in the constant term parameter of $\beta^*(\lambda)$. This mismatch can be explained by referring to the distribution for κ, shown in the right panel of Figure 1. For several 'extreme' small sized samples the true maximum of the likelihood function is attained when κ is on the boundary value of $\kappa = 1$. This implies that the true MLE of $\beta^*(\lambda)$ is minus infinity and the Fisher optimization algorithm thus fails to converge.

Since the re-sample size is much smaller than the original sample size, it is not surprising that there is a large drop in efficiency relative to the estimates from the original full sample. However, dividing the standard deviations by $\sqrt{26,364/2,208} = \sqrt{13}$ scales them back to the standard errors of the parameters in the original sample. As can be seen, the standard deviations in Table 4 agree well with the ML standard errors reported in Table 2, the exception again being the constant parameter of $\beta^*(\lambda)$.

3.5. **Parametric bootstrap test.** Under parametric bootstrap, hypothesis testing is remarkably easy. We simply need to fit the hypothesized null model, generate bootstrap replications under the assumptions of this model, and calculate the measure we wish to test, both for the real data and for the R sets of bootstrap data. If the value from the real data is in the 5% of the most extreme values in the combined set of $R + 1$ values, the hypothesis is rejected at the .05 level of significance. For illustrative purposes, we selected a single sample from the 'cross sections' of size 2,028, with ML estimates close to those reported in Table 2. The estimated value for κ in this sample was .916. Now consider testing the hypothesis $H_0 : \kappa \geq .999$, against the one-sided alternative $H_1 : \kappa < .999$ ($H_0 : \kappa = 1$ would be a theoretically implausible hypothesis to test for all cases). In $R = 4,999$ bootstrap re-samples from H_0, we found 51 values less then or equal to .916, so the p^* value is 51/5,000 = .0102. This finding leads us to reject the null hypothesis for this single sample.

[5] The total number of different samples is obtained as $\prod_{s=0}^{12} \frac{(2,028-s\times156)!}{156!(2,028-156-s\times156)!} \approx 10^{2,242}$.

4. Summary

Repeated cross-sectional surveys have increasingly become an important data source for research over the past decades. The accumulation of these surveys presents researchers from various disciplines with a growing opportunity to analyze longitudinal change. Dynamic models for the analysis of repeated cross-sectional data are, however, relatively rare and one may even argue that there is an increasing lag between the availability of surveys and models to analyze them.

The results presented here illustrate the usefulness of exploiting repeated cross-sectional surveys to identify and to estimate entry and exit transition probabilities, which are generally thought to be non-estimable from RCS data. The bootstrap and MCMC findings for the PC ownership example suggest that the ecological Markov model produces accurate estimates in large sample sizes. It also turns out that, in our empirical application at least, the Markov chain model performs almost as well as a first-order dynamic panel model. To rule out artificial results, samples of independent observations from the panel data were also analyzed, with similar results as for the full sample.

Topics not fully covered here are the distributional properties of the estimators in different model specifications and the sensitivity of inference procedures to varying sample sizes, so further Monte Carlo work on this panel ecological inference model is needed. Also, in addition to parametric bootstrap, nonparametric re-sampling could be used to examine the robustness of specification. Nonparametric simulation requires generating artificial data without assuming that the original data have some particular parametric distribution. Finally, although the impetus behind developing the methodology presented here came from the intend to dynamically model RCS data, it would be of interest to apply the model to panel data with missing observations for y_{t-1}. The Markov model could then be used, in conjunction with a first-order panel model for observations with non-missing y_{t-1}, to obtain model-based imputations for the missing data.

References

[1] Amemiya, Takeshi. 1985. *Advanced Econometrics.* Oxford: Basil Blackwell.

[2] Beck, Nathaniel, David Epstein, Simon Jackman, and Sharyn O'Halloran. 2001. *Alternative Models of Dynamics in Binary Time-Series-Cross-Section Models: The Example of State Failure.* Paper presented at the 2001 Annual Meeting of the Society for Political Methodology, Atlanta, GA.

[3] Davison, A.C., and D.V. Hinkley. 1997. *Bootstrap Methods and their Application.* Cambridge: Cambridge University Press.

[4] Diggle, Peter J., Kung-Yee Liang, and Scott L. Zeger. 1994. *Analysis of Longitudinal Data.* Oxford: Clarendon Press.

[5] Duncan, Otis Dudley, and Beverly Davis. 1953. An Alternative to Ecological Correlation. *American Sociological Review* 18:665-666.

[6] Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.* New York: Chapman and Hall.

[7] Franklin, Charles H. 1989. Estimation across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation (2SAIV). *Political Analysis* 1:1-23.

[8] Hall, Peter, and Susan R. Wilson. 1991. Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics* 47:757-762.

[9] Jarque, Carlos M., and Anil K. Bera. 1980. Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters* 6:255-259.

[10] King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Cambridge: Cambridge University Press.

[11] King, Gary, Ori Rosen, and Martin Tanner. 1999. Binomial-beta Hierarchical Models for Ecological Inference. *Sociological Methods and Research* 28:61-90.

[12] Kish, Leslie. 1987. *Statistical Design for Research.* New York: Wiley.

[13] Mebane, Walter R., and Jonathan Wand. 1997. *Markov Chain Models for Rolling Cross-Section Data: How Campaign Events and Political Awareness Affect Vote Intentions and Partisanship in the United States and Canada.* Paper presented at the 1997 Annual Meeting of the Midwest Political Science Association, Chicago Il.

[14] Moffitt, Robert. 1990. The Effect of the U.S. Welfare System on Marital Status. *Journal of Public Economics* 41:101-124.

[15] Moffitt, Robert. 1993. Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-sections. *Journal of Econometrics* 59:99-123.

[16] OECD. 2001. *Understanding the Digital Divide.* Pdf file available at http://www.oecd.org/pdf/M00002000/M00002444.pdf (May 2002).

[17] Pelzer, Ben, Rob Eisinga, and Philip Hans Franses. 2002. Inferring Transition Probabilities from Repeated Cross Sections. *Political Analysis* 10:113-133.

[18] Penubarti, Mohan, and Alexander A. Schuessler. 1998. *Inferring Micro- from Macrolevel Change: Ecological Panel Inference in Surveys.* Los Angeles CA: University of California.

[19] Ratkowsky, David A. 1983. *Nonlinear Regression Modeling: A Unified Practical Approach.* New York: Marcel Dekker.

[20] Ross, Sheldon M. 1993. *Introduction to Probability Models (5th ed.).* San Diego CA: Academic Press.

[21] Sigelman, Lee. 1991. Turning Cross Sections into a Panel: A Simple Procedure for Ecological Inference. *Social Science Research* 20:150-170.

[22] Tanner, Martin. 1996. *Tools for Statistical Inference.* New York: Springer.

Figure legends

Figure 1: *Histogram of ML estimates of κ for 5,000 bootstrap samples from the original full data, normal curve superimposed (left panel), and for 5,000 'cross-sectional' samples of 2,208 observations, one observation per household (right panel)*

Table 1. *Proportions of PC ownership in Dutch households over time, n of cases =2,208*

year	$\bar{y}_t$	$\bar{y}_t \mid y_{t-1} = 0$	$(1 - \bar{y}_t) \mid y_{t-1} = 1$
1986	.12		
87	.15	.05	.10
88	.20	.08	.12
89	.24	.08	.13
90	.28	.08	.08
91	.31	.09	.09
92	.36	.11	.09
93	.38	.10	.13
94	.41	.10	.09
95	.44	.13	.11
96	.48	.13	.07
97	.51	.14	.09
98	.57	.19	.07

Table 2. *ML, parametric bootstrap and MCMC estimates of RCS Markov model and ML estimates of first-order panel model, n of obs = 26,364*

	RCS Markov						Panel	
	ML[a]		bootstrap[b]		MCMC[b]		ML[a]	
$\delta(p_{t=1})$								
constant	-3.713	(.202)	-3.718	(.205)	-3.754	(.232)	-3.606	(.276)
			[4.137	-3.318]	[-4.225	-3.327]		
education	.382	(.054)	.381	(.055)	.393	(.056)	.364	(.072)
			[.271	.489]	[.288	.504]		
age 35-54	-.058	(.119)	-.057	(.121)	-.037	(.120)	.092	(.170)
			[-.294	.181]	[-.284	.197]		
age 55 and over	-.852	(.162)	-.859	(.165)	-.842	(.178)	-.782	(.252)
			[-1.201	-.551]	[-1.207	-.513]		
no. of household	.331	(.042)	.332	(.043)	.327	(.038)	.310	(.061)
members			[.248	.417]	[.249	.397]		
$\beta\,(\mu_{t=2,\ldots,13})$								
constant	-6.336	(.121)	-6.344	(.124)	-6.339	(.130)	-5.116	(.138)
			[-6.586	-6.110]	[-6.605	-6.105]		
education	.368	(.023)	.369	(.023)	.365	(.026)	.245	(.029)
			[.323	.413]	[.310	.414]		
age 35-54	.137	(.049)	.137	(.050)	.129	(.049)	-.098	(.067)
			[.042	.238]	[.037	.224]		
age 55 and over	-1.364	(.066)	-1.365	(.065)	-1.362	(.067)	-1.270	(.142)
			[-1.494	-1.240]	[-1.499	-1.226]		
no. of household	.421	(.018)	.422	(.018)	.425	(.020)	.375	(.089)
members			[.387	.457]	[.389	.470]		
income	.438	(.015)	.438	(.015)	.438	(.016)	.230	(.022)
			[.408	.468]	[.403	.467]		
time	.218	(.009)	.218	(.009)	.219	(.010)	.171	(.008)
			[.201	.236]	[.198	.240]		
$\beta^*(\lambda_{t=2,\ldots,13})$								
constant	-2.292	(.132)	-2.300	(.133)	-2.307	(.198)	-2.284	(.039)
			[-2.576	-2.058]	[-2.779	-1.938]		
$\ell\ell$	-12895.106						-7766.304	

[a] Standard error in parenthesis.

[b] The mean is reported as the point estimate, the standard deviation in parenthesis and the .95 percentile interval in brackets. The parametric bootstrap results are based on R=5,000 bootstrap samples from the original data and the MCMC findings on K=100,000 Metropolis sampler posterior estimates.

TABLE 3. *Parametric bootstrap estimates, based on R=5,000 bootstrap samples*

	bias x 10^2	bias/sd	rmse	skewness	excess kurtosis	Jarque-Bera
$\delta(p_{t=1})$						
constant	-.493	-.024	.205	-.098*	.094	9.812*
education	-.089	-.016	.055	-.008	.061	.796
age 35-54	.107	.009	.121	.032	-.026	1.008
age 55 and over	-.729	-.044	.165	-.179*	.104	28.954*
no. of household members	.128	.030	.043	.028	-.078	1.985
$\beta\,(\mu_{t=2,\ldots,13})$						
constant	-.862	-.070	.124	-.033	-.012	.931
education	.066	.029	.023	-.050	-.037	2.405
age 35-54	.040	.008	.050	.070	-.067	5.225
age 55 and over	-.059	-.009	.065	-.052	.000	2.285
no. of household members	.084	.047	.018	.010	-.025	.224
income	.065	.043	.015	-.032	.044	1.260
time	.022	.025	.009	.008	-.104	2.338
$\beta^*(\lambda_{t=2,\ldots,13})$						
constant	-.789	-.059	.133	-.293*	.296*	89.691*

Note. The bootstrap estimate of bias ($= \bar{\theta}_{bootstrap} - \theta_{ML}$) is multiplied by 100, and rmse $= \sqrt{\text{sd}^2 + \text{bias}^2}$. The standard errors of skewness and excess kurtosis are .035 and .069, respectively. The Jarque-Bera (1980) test statistic for normality has an asymptotic χ^2_2 distribution; the 5% critical value is 5.991.

*significant at the .05 level.

Table 4. *Mean and standard deviation* ($\div\sqrt{13}$) *of the RCS Markov ML estimates for 5,000 samples of 2,208 observations, one for each household*

	$\delta(p_{t=1})$		$\beta\,(\mu_{t=2,\ldots,13})$		$\beta^*(\lambda_{t=2,\ldots,13})^a$	
constant	-3.845	(.199)	-6.426	(.120)	-2.389	(.260)
education	.403	(.046)	.366	(.027)		
age 35-54	-.045	(.118)	.147	(.045)		
age 55 and over	-.785	(.160)	-1.423	(.063)		
no. of household members	.343	(.032)	.431	(.018)		
income			.447	(.015)		
time			.223	(.010)		

Note. Each sample is drawn without replacement and consists of 13 sets - one for each time period - of size 156. The standard deviation, divided by $\sqrt{13}$, is reported in parenthesis.

[a]Excluding 440 samples with $\beta^*(\lambda_t) \leq -8$ (i.e., $\kappa > .9996$).

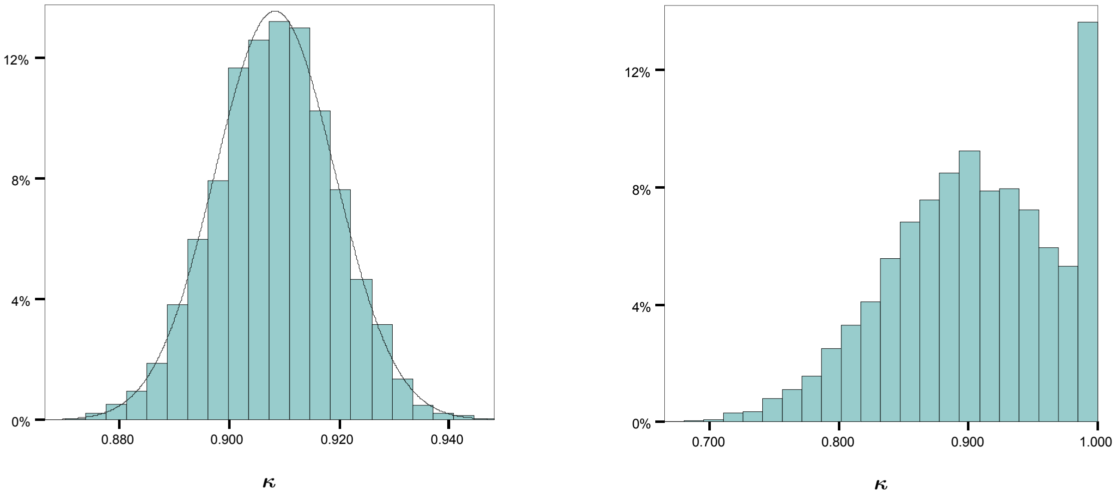

Figure 1 Histogram of ML estimates of κ for 5,000 bootstrap samples from the original full data, normal curve superimposed (*left panel*), and for 5,000 'cross-sectional' samples of 2,208 observations, one observation per household (*right panel*)