

Tunali, Insan; Ekinci, Emre; Yavuzoglu, Berk

**Working Paper**

## Rescaled Additively Non-ignorable (RAN) Model of Attrition and Substitution

Working Paper, No. 1220

**Provided in Cooperation with:**

Koç University - TÜSİAD Economic Research Forum, Istanbul

*Suggested Citation:* Tunali, Insan; Ekinci, Emre; Yavuzoglu, Berk (2012) : Rescaled Additively Non-ignorable (RAN) Model of Attrition and Substitution, Working Paper, No. 1220, Koç University-TÜSİAD Economic Research Forum (ERF), Istanbul

This Version is available at:

<https://hdl.handle.net/10419/108604>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

KOÇ UNIVERSITY-TÜSİAD ECONOMIC RESEARCH FORUM  
WORKING PAPER SERIES

**RESCALED ADDITIVELY NON-IGNORABLE (RAN)  
MODEL OF ATTRITION AND SUBSTITUTION**

Insan Tunali  
Emre Ekinci  
Berk Yavuzoglu

Working Paper 1220  
July 2012

---

KOÇ UNIVERSITY-TÜSİAD ECONOMIC RESEARCH FORUM  
Rumelifeneri Yolu 34450 Sarıyer/Istanbul

# RESCALED ADDITIVELY NON-IGNORABLE (RAN) MODEL OF ATTRITION AND SUBSTITUTION\*

Insan Tunali<sup>†</sup>, Emre Ekinici<sup>‡</sup>, Berk Yavuzoglu<sup>§</sup>

July 2012

---

\*Tunali would like to acknowledge discussions with Geert Ridder that prompted this line of research. Funding was provided by grant no. 109K504 by TUBITAK, The Scientific and Technological Research Council of Turkey. We are grateful to Huseyin Ikizler, Bengi Ilhan Yanik and Hayriye Ozgul Ozkan for research assistance, and to Elvan Ceyhan for consultations on our bootstrap approach. A version was presented as a Keynote Lecture by Tunali during the 12th International Symposium on Econometrics, Operations Research and Statistics held at Pamukkale University, Denizli, 26-29 May 2011. Comments from seminar and workshop participants at Bilkent, Koç and LSE, and detailed feedback from John Kennan are gratefully acknowledged.

<sup>†</sup>Corresponding author, Department of Economics, Koç University, Rumelifeneri You, Sariyer, 34450 Istanbul; phone: +90-212-3381425; fax: +90-212-338-1653; e-mail: itunali@ku.edu.tr

<sup>‡</sup>Department of Business Administration, Universidad Carlos III de Madrid: e-mail: eekinci2610@hotmail.com

<sup>§</sup>Department of Economics, University of Wisconsin-Madison: e-mail: yavuzoglu@wisc.edu

## Abstract

We modify the Additively Non-ignorable (AN) model of Hirano et. al. (2001) so that it is suitable for data collection efforts that have a short panel component. Our modification yields a convenient semi-parametric bias correction framework for handling endogenous attrition and substitution behavior that can emerge when multiple visits to the same unit are planned. We apply our methodology to data from the Household Labor Force Survey (HLFS) in Turkey, which shares a key design feature (namely a rotating sample frame) of popular surveys such as the Current Population Survey and the European Union Labor Force Survey. The correction amounts to adjusting the observed joint distribution over the state space using deflation factors expressed as parametric functions of the states occupied in subsequent rounds. Unlike standard weighting schemes, our method produces a unique set of corrected joint probabilities that are consistent with the margins used for computing the published cross-section statistics. Inference about the nature of the bias is implemented via Bootstrap methods. Our empirical results show that attrition/substitution in HLFS is a statistically and substantially important concern.

*Keywords:* attrition; substitution; selectivity; short panel; rotating sample frame; labor force survey.

# 1 Introduction

Attrition has been a major concern in applied research based on panel data. The study by Hausman and Wise (1979) constitutes an early attempt to model attrition as the outcome of rational economic behavior that can systematically bias the findings based on the balanced panel (subsample of non-attriters). As such the attrition problem is intimately related to the class of problems collected under the title of selectivity (Heckman, 1987). The subject has also drawn the attention of survey researchers (Madow et al., 1983). Formalizations by Rubin (1976), and Little (1982) (collected in Little and Rubin, 1987) have paved the way for establishing common terminology such as missing completely at random (which describes situations where non-attriters constitute a random subsample of the full sample) and ignorable attrition (when attrition does not impart bias on the outcome under study).

Our paper builds on an important contribution by Hirano, Imbens, Ridder and Rubin (Hirano et al., 2001). Their paper approaches the attrition issue as an identification problem that amounts to recovering the joint distribution of interest for the full population, when all we have is a subsample subjected to potentially non-ignorable attrition. They work with a discrete joint distribution that characterizes the finite outcomes of interest, express the attrition probability as a function of the set of outcomes before and after attrition, and establish that identification can be achieved when unbiased estimates of the marginal distributions are available. While the typical panel data collection effort yields an unbiased estimator of the first round marginal distribution, attrition renders subsequent round marginals suspect. Hirano et al. (2001) exploit an independently conducted cross-section survey (so-called refreshment sample) to provide an unbiased estimator for the second round marginal distribution. As usual, adjustment of the balanced sample proceeds by using the inverted attrition (selection) probabilities as weights. Equating the row and column sums of the reweighted balanced panel cell counts (fractions) to the respective marginals, a just-identified system of equations that yields the parameter estimates of the weighting function is obtained. Since the weighting function only allows for main effects and rules out interactions, Hirano et al. (2001) name this model Additively Non-ignorable (AN) model of attrition. They show that both the popular formulations of Little-Rubin and the Hausman-Wise model are nested within the AN model. Thus the AN model not only offers a theoretically appealing correction for attrition, but it also affords

tests of widely used models.

In this paper we establish that the key ideas embedded in the AN model can be used for addressing a broader class of non-response problems. In particular we modify the AN model so that it is suitable for data collection efforts that have a short panel component. Household surveys that have this feature – such as the Household Labor Force Surveys (HLFS) in Turkey we use below, as well as popular data sets such as the Current Population Survey (CPS) in the U.S., and most country surveys included in the European Union Labor Force Survey (EULFS) – call for repeat visits to the same household according to a pre-determined schedule but limit the maximum number of visits. The schedule is supported by a rotational design that ensures nationwide representation as well as updating. Towards that end, after each round households that exit the sample frame are replaced by a group of new households. If the data collection agency provides the weights needed for rendering the subset of new household nationally representative, this amounts to having a refreshment sample, as in Hirano et al. (2001).

Even if weights were available, the idea of treating the subsample that is rotated in as a refreshment (i.e. independent) sample has an important drawback. The properly weighted marginals based on the subsample will typically not be the same as the weighted marginals calculated on the full cross-section and published as period-specific official statistics. Since our approach does not require a refreshment sample, it is not subject to this criticism.

Surveys that rely on a rotational design use typically have an address- or dwelling-based sample frame. In some cases a longitudinal view is adopted, so that households (or individuals) that enter the sample frame are followed even when they leave the original address (such as the CPS, see BLS, 2002). In other cases the data collection agency prefers to treat each round of the data as an independent cross-section (such as some country components of the EULFS, see EUROSTAT, 2007). The data set we work with, HLFS-Turkey, is a typical example of the latter (TURKSTAT, 2001). Residential addresses are kept in the sample frame for a certain time and visited according to the rotation schedule whether or not any respondents are found. Standard non-response adjustments (based on demographics) are used to obtain marginal distributions, which in turn serve as the source of published official statistics. Since a subset of the households are surveyed in two adjoining periods, such surveys also lend themselves for dynamic analyses. However finding suitable weights is a challenge.

The problem is attributable to the fact that such data not only suffer from attrition (response followed by non-response) but also from substitution (non-response followed by response). As we show below, in these cases a key parameter of the AN model is not identified. However, a correction scheme which renders the dynamic estimates consistent with the official cross-sectional statistics can still be found. Since this amounts to treating the unidentified probability as a nuisance parameter, we term the new model Rescaled Additively Non-ignorable (RAN) model of attrition. We show that the model can be estimated with semi-parametric methods which are computationally simpler than the EM-algorithm based imputation methods used in Hirano et al. (2001). In fact the data requirements for implementation of RAN methodology are extremely minimal: namely, the joint distribution obtained from the balanced panel, the marginal distributions obtained from another data source that does not have the representation problems, and the sample sizes that yielded the respective distributions.

The idea of reconciling observed flow data between states with cross sectional stocks via probabilistic adjustments expressed as a function of the states predates Hirano et al. (2001). Abowd and Zellner (1985) and Stasny (1986, 1988) work with counts obtained from short panels, and focus on adjusting the flow data so that they are consistent with the properly weighted margins that represent the target population. The contrasts with their approaches and ours will be taken up below. In fact all these approaches can be situated within a broader statistical framework directed at reconciling key statistical features of survey data with what is known about the population (Little, 1993). Although it is not directed to panel data, the model based adjustment of Little and Wu (1991) in particular echoes the fundamental ideas exploited in AN and RAN models.

We begin our formal treatment in Section 2 by introducing our model and establish its links with the AN Model. In Section 3 we discuss our estimation and inference methodology. We then relate our approach to others developed in the statistics literature. Section 4 contains examples that illustrate the utility and potential limitations of the proposed approach. Section 5 offers a short compilation of the lessons learned from a broader investigation. We conclude the paper with a brief summary of the key aspects of our model and potential uses.

## 2 RAN Model

Consider data collection efforts directed to households which utilize a rotational design, whereby each household remains in the sample frame for a predetermined number of periods. Several advantages are apparent: Firstly, by limiting the number of revisits, the cost of the data collection effort is balanced against the response burden imposed on the households. Secondly, by including a fresh subsample every period, the sample is kept up to date. Thirdly, the rotational design yields a short panel. However use of the panel component ushers in new challenges when drawing inferences about the population. In fact it is often not fully exploited for want of weighting schemes consistent with those used in obtaining the cross-sectional estimates.

Without loss of generality we refer to the equally spaced rounds of data collection as the first period and the second period. We distinguish between the complete panel (CP), which includes all subjects intended for repeat visits, and the balanced panel (BP), which only includes subjects who have been successfully interviewed in both periods. We also keep track of households which are rotated out of the sample after period 1, and households which are rotated in during period 2. We introduce three random variables and associated parameters:

$$D = \begin{cases} 1 & \text{if designated for the Complete Panel (w/prob.} = \delta) \\ 0 & \text{if not (w/prob.} = 1 - \delta) \end{cases}, \quad (1)$$

$$C = \begin{cases} 1 & \text{if observed in the 1}^{st} \text{ period only (w/prob.} = \gamma_1) \\ 2 & \text{if observed in the 2}^{nd} \text{ period only (w/prob.} = \gamma_2) \\ 3 & \text{if observed in both periods (w/prob.} = \gamma_3 = 1 - \gamma_1 - \gamma_2) \end{cases}, \text{ given } D = 1; \quad (2)$$

$$R = \begin{cases} 1 & \text{if observed in the 1}^{st} \text{ period for the last time (w/prob} = \phi) \\ 2 & \text{if observed in the 2}^{nd} \text{ period for the first time (w/prob} = 1 - \phi) \end{cases}, \text{ given } D = 0. \quad (3)$$

Although  $D$  and  $R$  are usually predetermined as part of the sampling frame, it may be useful to



treat them as random variables because of practical issues such as encountering an establishment rather than a household at the address, and non-response by households. We shift the focus to individuals, so that  $D = 1$  indicates that the individual is designated for the CP. For such individuals, there are 3 possibilities:  $C = 1$  denotes attritors,  $C = 2$  denotes substitutes, and  $C = 3$  denotes individuals observed in both periods (i.e., those in the balanced panel). If an individual is not designated for the CP ( $D = 0$ ), then she either rotates out ( $R = 1$ ) or rotates in ( $R = 2$ ).

Let  $y$  and  $x$  denote random variables which are the main objects of the data collection effort. We distinguish between endogenous outcomes ( $y$ ) and exogenous covariates ( $x$ ). Some of the exogenous covariates may serve as objects of stratification. Others may identify subpopulations of interest. The primary objective of the statistical agency is to produce period-specific statistical indicators based on  $y$ , conditional on  $x$ . In what follows we use subscripts to denote period-specific values of  $y$ , and for notational convenience (and without loss of generality) treat  $x$  as time invariant. The joint distribution of interest is  $f(y_1, y_2|x)$ . In the typical application this is a discrete distribution which classifies individuals of a given type according to a pair of outcomes  $(y_1, y_2)$ . We suppress the conditioning on  $x$  for brevity, and use equation (1) to express the joint distribution as:

$$f(y_1, y_2) = f(y_1, y_2, D = 0) + f(y_1, y_2, D = 1). \quad (4)$$

We then use equations (2)-(3) and break down the components further as

$$\begin{aligned} f(y_1, y_2) &= f(y_1, y_2, D = 0, R = 1) + f(y_1, y_2, D = 0, R = 2) + f(y_1, y_2, D = 1, C = 1) \quad (5) \\ &\quad + f(y_1, y_2, D = 1, C = 2) + f(y_1, y_2, D = 1, C = 3). \end{aligned}$$

Let's examine each of the five components in turn. We begin with the terms for individuals who are not designated for the CP. Repeated use of Bayes' Theorem yields

$$\begin{aligned} f(y_1, y_2, D = 0, R = 1) &= \Pr(R = 1|y_1, y_2, D = 0)f(y_1, y_2, D = 0) \\ &= \Pr(R = 1|y_1, y_2, D = 0) \Pr(D = 0|y_1, y_2)f(y_1, y_2) \\ &= \Pr(R = 1|D = 0) \Pr(D = 0)f(y_1, y_2) \\ &= \phi(1 - \delta)f(y_1, y_2), \end{aligned} \quad (6)$$

where we used the fact that designation of an individual for rotation, or for the CP is done independently of  $(y_1, y_2)$ . Likewise,

$$f(y_1, y_2, D = 0, R = 2) = (1 - \phi)(1 - \delta)f(y_1, y_2). \quad (7)$$

Equations (6) and (7) show that the contribution to the joint distribution of interest by individuals who are not designated for the CP is a rescaled version of that distribution. Put differently, there is no additional information in the  $D = 0$  subsample that can be exploited for the purposes of recovering the joint distribution. Next, we turn to the terms in equation (5) for individuals designated for the CP ( $D = 1$  subsample) and establish that this is no longer the case. For attritors we have

$$\begin{aligned} f(y_1, y_2, D = 1, C = 1) &= \Pr(C = 1|y_1, y_2, D = 1)f(y_1, y_2, D = 1) \\ &= \Pr(C = 1|y_1, y_2, D = 1)\Pr(D = 1|y_1, y_2)f(y_1, y_2) \\ &= \Pr(C = 1|y_1, y_2, D = 1)\Pr(D = 1)f(y_1, y_2) \\ &= \Pr(C = 1|y_1, y_2, D = 1)\delta f(y_1, y_2), \end{aligned} \quad (8)$$

where we used the fact that designation of an individual for the CP is done independently of  $(y_1, y_2)$ . Note that in equation (8) the probability of attrition is expressed as a function of  $(y_1, y_2)$ . Likewise for substitutes we have

$$f(y_1, y_2, D = 1, C = 2) = \Pr(C = 2|y_1, y_2, D = 1)\delta f(y_1, y_2). \quad (9)$$

As seen in equation (9), the probability of substitution is also a function of  $(y_1, y_2)$ . The remaining term in (5) may be obtained as:

$$\begin{aligned} f(y_1, y_2, D = 1, C = 3) &= f(y_1, y_2|D = 1, C = 3)\Pr(D = 1, C = 3) \\ &= f(y_1, y_2|D = 1, C = 3)\Pr(C = 3|D = 1)\Pr(D = 1) \\ &= f(y_1, y_2|D = 1, C = 3)\gamma_3\delta. \end{aligned} \quad (10)$$

It is straightforward to see that  $f(y_1, y_2|D = 1, C = 3)$  can be identified non-parametrically from

the balanced panel. However, since the balanced panel consists of individuals who have not been subjected to attrition or substitution, in general  $f(y_1, y_2|D = 1, C = 3) \neq f(y_1, y_2)$ .

Substitution of the terms on the right hand sides of equations (6)-(10) in equation (5) yields:

$$\begin{aligned} f(y_1, y_2) &= \phi(1 - \delta)f(y_1, y_2) + (1 - \phi)(1 - \delta)f(y_1, y_2) + \Pr(C = 1|y_1, y_2, D = 1)\delta f(y_1, y_2) \\ &\quad + \Pr(C = 2|y_1, y_2, D = 1)\delta f(y_1, y_2) + f(y_1, y_2|D = 1, BP = 3)\gamma_3\delta. \end{aligned} \quad (11)$$

Upon collecting terms, simplifying and rearranging we get

$$f(y_1, y_2) = \frac{f(y_1, y_2|D = 1, C = 3)\gamma_3}{[1 - \Pr(C = 1|y_1, y_2, D = 1) - \Pr(C = 2|y_1, y_2, D = 1)]}. \quad (12)$$

Finally, using the fact that  $\sum_{m=1}^3 \Pr(C = m|y_1, y_2, D = 1) = 1$ , we get

$$f(y_1, y_2) = \frac{f(y_1, y_2|D = 1, C = 3)\gamma_3}{\Pr(C = 3|y_1, y_2, D = 1)}. \quad (13)$$

The last equation is equivalent to the key equation of the AN Model of Hirano et al. (2001: 1647). Recall that the case they study involves a two period panel, and the only concern is non-ignorable non-response (attrition) in the second period.<sup>1</sup> Thus  $\gamma_3 = \Pr(C = 3|D = 1)$  is non-parametrically identified. They specify the probability in the denominator of (13) as a parametric function of  $(y_1, y_2)$ , and investigate the conditions under which it can be identified. In our case the sampling design involves rotation, whereby non-ignorable non-response may occur either in period 1 (substitution) or period 2 (attrition). This poses additional challenges for the identification of  $\gamma_3$ .<sup>2</sup> As we proceed to show, it can be treated as a nuisance parameter. Thus our version of (13) is:

$$f(y_1, y_2) = w(y_1, y_2)f(y_1, y_2|D = 1, C = 3), \quad (14)$$

where  $w(y_1, y_2) = \gamma_3/\Pr(C = 3|y_1, y_2, D = 1) > 0$  by construction. Additional restrictions on  $w(y_1, y_2)$  are needed for identification.

As in Hirano et al. (2001), we use the restrictions on joint probabilities imposed by the marginals.

---

<sup>1</sup>Non-response in the initial period can be handled by reweighting via  $x$ .

<sup>2</sup>Typically the subsets of CP are identified in the data set. However, since we do not know which weights to attach to a given individual, this information is not sufficient for identifying  $\gamma_3 = \Pr(C = 3|D = 1)$ .

In the AN model, the original sample yields the unbiased marginal distribution for the first period, and the refreshment sample provides the unbiased marginal distribution for the second period.<sup>3</sup> In our case identifying information comes from the marginal distributions which are the (properly weighted) cross-sectional statistics published by the data collection agency. Restoring the conditioning on covariates  $x$ , the equations of interest are:

$$\sum_{y_2} f(y_1, y_2|x) = \sum_{y_2} w(y_1, y_2|x) f(y_1, y_2|D = 1, C = 3, x) = f_1(y_1|x), \quad (15)$$

$$\sum_{y_1} f(y_1, y_2|x) = \sum_{y_1} w(y_1, y_2|x) f(y_1, y_2|D = 1, C = 3, x) = f_2(y_2|x). \quad (16)$$

Equation (14) has a form which is familiar to survey data users. Once the function  $w(y_1, y_2)$  is estimated, it can be used to inflate/deflate (i.e. reflate) the cells of the balanced panel so that the object of interest  $f(y_1, y_2|x)$  can be recovered. Suppose  $y$  has  $k$  distinct values so that  $f(y_1, y_2|x)$  can be viewed as a  $k \times k$  table. Equations (15)-(16) provide the restrictions that must be satisfied by the reflated balanced panel fractions where  $w(y_1, y_2)$  serve as the reflation factors. Since  $\sum_{y_1} \sum_{y_2} f(y_1, y_2|x) = 1$ , for  $k \geq 2$  the marginals provide  $2k - 1$  pieces of independent information. Thus the  $k^2$  reflation factors viewed as functions of  $(y_1, y_2)$  can have at most  $2k - 1$  unknown parameters. We mimic the approach in Hirano et al. (2001) and impose additivity. To assess the role of parametric assumptions, we follow Chen (2001) and entertain three different specifications for this function, respectively linear, convex and concave. Details will emerge in the next section.

It is straightforward to establish that the RAN model has all the features that render the AN model attractive. Firstly, since the RAN model preserves the additivity restriction of the AN model, identification proof in Hirano et al. (2001) still applies.<sup>4</sup> Secondly, it nests the popular models of attrition. If non-response is ignorable,  $w(y_1, y_2) = 1$  for all  $(y_1, y_2)$  combinations. This is the case dubbed as Missing Completely at Random (MCAR) by Rubin (1976). If non-response is a function of the first period outcomes only,  $w(y_1, y_2) = w(y_1)$ . Little and Rubin (1987), and others – for example Fitzgerald et al. (1998), Hirano et al. (2001) – call this case Missing at Random because in a regular panel it is straightforward to adjust the balanced panel fractions using probability weights

---

<sup>3</sup>To study the consequences of attrition in the standard panel context, Fitzgerald et al. (1998) and MaCurdy et al. (1998) rely on comparisons of later wave distributions with independent samples but do not propose a formal model of correction for attrition.

<sup>4</sup>For a simpler proof see Bhattacharya (2008).

expressed as a function of observables in the first period. Note that in the present case we are dealing with substitution as well as attrition. Since substitution implies that first period outcomes are unobserved, it is not possible to carry out the adjustment based on period 1 information alone. We use the naming convention anyway, to convey the fact that the deflation factors are expressed as a function of first period outcomes only (even though they may be unobserved for a subset of the sample). Finally, if non-response is a function of second period outcomes only,  $w(y_1, y_2) = w(y_2)$ . Hirano et al. (2001) call this the Hausman and Wise (HW) model because the case was first studied by Hausman and Wise (1979).<sup>5</sup>

### 3 Estimation and Inference in RAN Model

For a given set of observed fractions  $f(y_1, y_2 | D = 1, C = 3, x)$  obtained from the balanced panel, and cross-section estimates  $f_1(y_1 | x)$  and  $f_2(y_2 | x)$  obtained from official statistics, estimation problem boils down to solving a system of  $2k - 1$  equations in at most  $2k - 1$  unknowns.<sup>6</sup> We impose a functional form for  $w(\cdot)$ , and estimate it parametrically, as  $\hat{w}(y_1, y_2 | x) = w(\hat{\theta} | y_1, y_2, x)$  where  $\theta$  has at most  $2k - 1$  elements. We then compute the joint probabilities of interest (reflated panel fractions) as a product of the estimated deflation factors and the observed fractions:

$$f(y_1, y_2 | x) = \hat{w}(y_1, y_2 | x) f(y_1, y_2 | D = 1, C = 3, x) \quad (17)$$

For inference, we rely on standard Bootstrap methodology (Efron, 1979). Each of the random components  $f(y_1, y_2 | D = 1, C = 3, x)$ ,  $f_1(y_1 | x)$  and  $f_2(y_2 | x)$  need to be bootstrapped. Technically speaking the joint distribution for the balanced panel is extracted from the same data set that yields the marginals. That is, all three distributions are functions of the survey data that have been collected during the two periods under study. These functions involve predetermined features, such as censoring due to the rotation design. They also involve the unknown attrition/substitution process. In addition, the function that maps the cross-section data into official statistics includes

---

<sup>5</sup>Fitzgerald et al. (1998) also study this model and contrast it with MAR using popular selection terminology. They point out that while selection in the MAR model is on (first period) observables, selection in the HW model is on unobservables (unobserved second period outcomes). The observable/unobservable distinction is not useful for characterizing the attrition/substitution encountered in a short panel obtained from a rotating sample frame.

<sup>6</sup>In our empirical work we relied on MATLAB's predefined function  $f_{solve}(\cdot)$  to find the solution to this system. In fact EXCEL's predefined function 'solver' is also capable of handling the computations.

the weights used by the statistical agency, which may be known, or unknown (as in our case). Thus, a joint bootstrap scheme is elusive.

In the case of HLFS-Turkey, the rotation schedule of the sample frame ensures that about half of the addresses overlap ( $D = 1$  in the set-up of section 2). Two distinct groups of individuals who are not designated for the complete panel ( $R = 1$  and  $R = 2$ ) also contribute to the raw marginals. Furthermore, TURKSTAT manipulates the raw marginals using period specific weights based on demographic characteristics of individuals (namely age, sex and geographic location). This adjustment aims to bring the distribution of demographic attributes in line with those obtained from independent population projections. The corrected marginals are reported as official statistics.<sup>7</sup> With these features in mind, we propose drawing three independent bootstrap samples that have the same sample size as in the raw data. We resample from the actual balanced panel that yields  $f(y_1, y_2 | D = 1, C = 3, x)$  and two artificially created marginal samples which yield the fractions  $f_t(y_t | x)$  published by TURKSTAT. Using these three independent bootstrap samples, we can solve (15)-(16) to calculate a new  $\hat{\theta}$ . After conducting a suitable number of replications (we used 100), we can obtain bootstrap means, standard errors and estimated variance-covariance matrix for  $\hat{\theta}$ . These statistics can be used along with standard asymptotic theory (Efron, 1979) to test the statistical significance of the parameters of the RAN model, and hypotheses concerning the nature of the attrition process.

Apart from choice of the functional form for  $w(\cdot)$ , our procedure is fully non-parametric. We propose treating each distinct  $x$  as a separate stratum, and repeating the estimation/inference exercise. Clearly there are some practical limits to this fully non-parametric procedure; we will return to this issue below, when we discuss the lessons learned from a broader empirical investigation.

At this point it is appropriate to provide a brief account of how our adjustment procedure differs from existing methods. As mentioned earlier, Abowd and Zellner (1985) and Stasny (1986, 1988) deal with the same substantive issues, but work with counts. The goal is to adjust the gross flow data so that it can be reconciled with the marginals. Abowd and Zellner (1985) use a multiplicative model to distribute those who are not observed in both periods to the appropriate margins (original set of

---

<sup>7</sup>By analogy to the widely used concept of “stratification,” Little (1993) refers to the procedures for reweighting survey data using information from aggregate data on the population obtained from other sources as “post-stratification.” Clearly, the weights do some correction for attrition and substitution, but whether this is adequate can be debated in light of the evidence in Tunali (2009). Since this methodology is sanctioned by Eurostat, we do not question it here.

states plus two others, respectively attritor/substitute and rotated in/out). Like us (see Section 4) they study three states (nine cells in the flow matrix), but estimate 18 unknown parameters subject to six restrictions coming from the margins. Thus, they not only allow interaction effects, but they also distinguish between attrition and substitution parameters. Clearly this overparametrized model cannot be used to implement separate adjustments for each period pair. They assume stationarity and use multiple rounds of CPS data to estimate “average” parameters by minimizing the weighted squared deviation of the adjusted gross flow margins from the observed population margins.

Stasny (1986, 1988) uses an additive model that resembles ours. In her model an individual designated for the panel can lose either its row or column designation, with different probabilities. In the richest models (A and D in Stasny, 1988) she expresses one of these probabilities as a function of states occupied in the first period, the other as a function of states occupied in the second period. Thus, the probability that someone designated for the complete panel ends up in one of the margins, is a function of the states in both periods. Clearly this treatment is exactly the same as ours. In her examples there are three states and six free parameters for each period pair, which can be estimated subject to the six restrictions on the margins. She is able to identify an extra parameter because she uses count data, while we work with shares. She uses maximum likelihood estimation on multiple rounds of data from CPS and Canadian LFS.

There is a well-established line of research in the statistical literature which is directed at the important distinction between the sampled and the target population, and on methods used in reconciling them (Madow et al., 1983). Little (1993) refers to adjustments of data obtained from surveys (i.e. sampled population) using aggregate data on the (target) population obtained from other sources as “post-stratification.” The bulk of this paper is concerned with the case when the population joint distribution of the post-stratification variables is known. Little briefly discusses a case which is of special interest for us: only the marginal population distributions of the post-stratification variables are known. When non-response is present, the joint distribution of the post-stratification variables in the sample is not adequate for estimation (unless MCAR or MAR is assumed). This case is covered at length in Little and Wu (1991) where a formal model for nonresponse is given. Notably they address the identification issue and show that a model in which the response probability is expressed as a product of row and column effects is just identified. They propose an iterative method (raking) for estimation of this model. This version of the post-

stratification exercise is intimately connected with the AN/RAN approach. Instead of the additive model that drives the correction is AN/RAN models, Little and Wu (1991) have a multiplicative model. Furthermore in estimating the AN model, Hirano et al. (2001) adopt the predictive modeling perspective of Little (1991), whereby imputation of the missing outcomes precedes the estimation of the joint distribution of interest. In the RAN model we proceed with the estimation of the reflation factors and the adjusted cell probabilities without engaging in computationally costly imputation.

## 4 Examples

We illustrate the utility of the RAN model by applying it to a case where  $y$  indicates labor market status and takes one of three values (0 = non-participant, 1 = employed, 2 = unemployed). In this case the equation system (15)-(16) yields five independent equations, so we can estimate up to 5 parameters. We express  $w(y_1, y_2|x)$  as function of a linear index in  $(y_1, y_2)$  and use indicators for distinct labor market states. We take the individuals who are not in the labor force in both periods ( $y_1 = 0, y_2 = 0$ ) as our reference category. The other distinct categories  $y_t$  have their own parameters in each time period. For period  $t(= 1, 2)$ , the indicators may be defined as:

$$\begin{aligned} z_{t1} &= \begin{cases} 1 & \text{if employed } (y_t = 1) \\ 0 & \text{otherwise} \end{cases} ; \\ z_{t2} &= \begin{cases} 1 & \text{if unemployed } (y_t = 2) \\ 0 & \text{otherwise} \end{cases} . \end{aligned} \quad (18)$$

Let  $\underline{z}' = (1 \ z'_1 \ z'_2) = [1 \ z_{11} \ z_{12} \ z_{21} \ z_{22}]$ ,  $\underline{\theta}' = [\theta_{00} \ \theta_{11} \ \theta_{12} \ \theta_{21} \ \theta_{22}]$ , and define the linear index:

$$i(y_1, y_2) = i(\underline{\theta}'\underline{z}|x) = \theta_{00} + \theta_{11}z_{11} + \theta_{12}z_{12} + \theta_{21}z_{21} + \theta_{22}z_{22} \quad (19)$$

This function is additive in the unknown  $\theta$ 's which capture the dependency on the labor market states  $(y_1, y_2)$  via  $\underline{z}_1$  and  $\underline{z}_2$ . As in Hirano et al. (2001), we rule out interactions and focus on the main effects of the labor market states. In obtaining the reflation factors, we use three parametric forms:

(i) linear:  $w_L(y_1, y_2|x) = i(\underline{\theta}'\underline{z}|x)$ , (ii) convex:  $w_X(y_1, y_2|x) = \exp\{i(\underline{\theta}'\underline{z}|x)\}$ , and (iii) concave:



$w_E(y_1, y_2|x) = 2 - \exp \{i(\underline{\theta}'z|x)\}$ . Note that  $w(y_1, y_2) = 1$  iff  $\theta_{00} = 1, \theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} = 0$  in the linear case. In the nonlinear cases,  $w(y_1, y_2) = 1$  iff  $\theta_{00} = \theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} = 0$ .

For the linear case, the restrictions implied by equations (15)-(16) can be represented as in Table 1, where  $p_{y_1 y_2} = f(y_1, y_2|D = 1, C = 3, x)$ . To recapitulate, the task amounts to finding the refutation factors (functions of  $\theta$ 's) which would bring the adjusted cell probabilities in line with the marginals reported by the data collection agency.

Table 1: A 3x3 Linear RAN Model

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\theta_{00}p_{00}$	$(\theta_{00} + \theta_{21})p_{01}$	$(\theta_{00} + \theta_{22})p_{02}$	$f_1(0)$
$y_1 = 1$	$(\theta_{00} + \theta_{11})p_{10}$	$(\theta_{00} + \theta_{11} + \theta_{21})p_{11}$	$(\theta_{00} + \theta_{11} + \theta_{22})p_{21}$	$f_1(1)$
$y_1 = 2$	$(\theta_{00} + \theta_{12})p_{20}$	$(\theta_{00} + \theta_{12} + \theta_{21})p_{21}$	$(\theta_{00} + \theta_{12} + \theta_{22})p_{22}$	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

For the linear case, the system of equations has the observationally equivalent representation given below:

$$\begin{bmatrix} \sum_{j=0}^2 p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^2 p_{1j} & \sum_{j=0}^2 p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^2 p_{2j} & 0 & \sum_{j=0}^2 p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^2 p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^2 p_{k1} & p_{11} & p_{21} & \sum_{k=0}^2 p_{k1} & 0 \\ \sum_{k=0}^2 p_{k2} & p_{12} & p_{22} & 0 & \sum_{k=0}^2 p_{k2} \end{bmatrix} \begin{bmatrix} \theta_{00} \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{bmatrix} = \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_2(0) \\ f_2(1) \\ f_2(2) \end{bmatrix} \quad (20)$$

Inspection reveals that the system is of the form  $A\theta = \underline{b}$  where  $A$  is of rank = 5. One of the constraints is redundant, in the sense that it will be automatically met once the solution to the reduced system is found. We prove this in the appendix by starting with a particular system of five equations in five unknowns, and showing that any other representation can be transformed to the one we started with by a simple pivoting operation. Consequently, the solution to the reduced system is unique, and does not depend on which constraint is left out.

In Table 2, we compiled a set of parameter estimates from a 3x3 RAN model for annual transitions on data from the Household Labor Force Survey (HLFS) in Turkey, together with bootstrap means and standard errors based on 100 replications. In this case  $x$  denotes the entire working age population, ages 15 and over. The balanced panel contained over 20,000 observations. The first

Table 2: A 3x3 RAN Model - Parameter Estimates  
Annual Transitions Between 2001-Q1 and 2002-Q1  
 $x = \text{age 15 and over}$

Parameter	$\theta_{00}$	$\theta_{11}$	$\theta_{12}$	$\theta_{21}$	$\theta_{22}$
(i) $w(\cdot)$ linear:					
Estimate	0.8987	0.0956	0.2524	0.1315	0.1779
Bootstrap mean	0.8994	0.0999	0.2423	0.1263	0.1755
Bootstrap std. error	0.0063	0.0282	0.0509	0.0290	0.0507
(ii) $w(\cdot)$ convex:					
Estimate	-0.1057	0.0957	0.2306	0.1293	0.1703
Bootstrap mean	-0.1050	0.0999	0.2221	0.1243	0.1672
Bootstrap std. error	0.0070	0.0283	0.0440	0.0288	0.0462
(iii) $w(\cdot)$ concave:					
Estimate	-0.0975	0.0960	0.2848	0.1349	0.1885
Bootstrap mean	-0.0968	0.1007	0.2725	0.1295	0.1875
Bootstrap std. error	0.0060	0.0298	0.0656	0.0308	0.0602
Sample Sizes:					
Balanced panel	21, 731				
First period cross-section	52, 389				
Second period cross-section	53, 810				

Data Source: Household Labor Force Survey, TURKSTAT.

and second period marginals in the raw data contained over 52,000 observations. Thus it is not surprising that all RAN model parameters are estimated extremely precisely.

As we noted earlier, HLFS sample frame ensures that about half of the addresses visited in a given period are also visited the next period. Taking the sample sizes we reported above, we see that the balanced panel sample amounted to about 40 percent of the respective marginals. The fact that this fraction is considerably lower than the expected 0.5 can be taken as a rough statistic that warns us about the potential severity of the attrition/substitution problem.<sup>8</sup> What matters, of course, is whether the process that excludes individuals designated for the complete panel from the balanced panel is ignorable. Given the evidence from the bootstrap exercise, we do not expect this to be the case. In fact, Wald tests provide overwhelming evidence that the attrition and substitution process is non-ignorable. Furthermore, alternatives to RAN model are deemed inadequate for capturing the selectivity (all  $p$ -values are practically zero). The key insight from labor economics, that attrition and substitution behavior is intimately connected with labor market behavior, is vindicated.

In Table 3, we compiled the set of reflation factor estimates we obtained from the RAN model

<sup>8</sup>The realized magnitudes of attrition and substitution in the HLFS over the period 2000-2002 are reported in Tunali (2009).

Table 3: A 3X3 RAN Model - Reflation Factors  
Annual Transitions Between 2001-Q1 and 2002-Q1  
 $x = \text{age 15 and over}$

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\left\{ \begin{array}{c} 0.8986 \\ 0.8997 \\ 0.8976 \end{array} \right\} 0.5052$	$\left\{ \begin{array}{c} 1.0302 \\ 1.0238 \\ 1.0366 \end{array} \right\} 0.0566$	$\left\{ \begin{array}{c} 1.0766 \\ 1.0667 \\ 1.0870 \end{array} \right\} 0.0159$	$f_1(0)$
$y_1 = 1$	$\left\{ \begin{array}{c} 0.9943 \\ 0.9901 \\ 0.9985 \end{array} \right\} 0.0740$	$\left\{ \begin{array}{c} 1.1258 \\ 1.1267 \\ 1.1248 \end{array} \right\} 0.2952$	$\left\{ \begin{array}{c} 1.1722 \\ 1.1739 \\ 1.1706 \end{array} \right\} 0.0209$	$f_1(1)$
$y_1 = 2$	$\left\{ \begin{array}{c} 1.1511 \\ 1.1330 \\ 1.1708 \end{array} \right\} 0.0113$	$\left\{ \begin{array}{c} 1.2826 \\ 1.2894 \\ 1.2754 \end{array} \right\} 0.0122$	$\left\{ \begin{array}{c} 1.3290 \\ 1.3433 \\ 1.3133 \end{array} \right\} 0.0085$	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

parameter estimates reported in Table 2. For brevity we excluded the numbers for the margins. The numbers reported in each cell are of the form given on the right hand side of equation (17). For each cell we report the estimates of the reflation factors  $w(\cdot)$  associated with all three functional forms (respectively linear, convex, concave; shown inside braces) followed by the fraction obtained from the balanced panel. Reflation factors below (above) one mark labor market states which are overrepresented (underrepresented) in the balanced panel. Note that for some states the bias induced by attrition/substitution is practically zero [see  $(y_1 = 1, y_2 = 0)$ ] but for others it is substantial [e.g.  $(y_1 = 2, y_2 = 2)$ ]. The findings from our sensitivity analysis are typical, in that functional form does not make much of a difference.

Table 4: A 3x3 RAN Model - Adjusted and [Unadjusted] Joint and Marginal Probabilities  
Annual Transitions Between 2001-Q1 and 2002-Q1  
 $x = \text{age 15 and over}$

	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	0.4540 [0.5052]	0.0584 [0.0566]	0.0172 [0.0160]	0.5296 [0.5778]
$y_1 = 1$	0.0736 [0.0740]	0.3323 [0.2952]	0.0245 [0.0209]	0.4305 [0.3902]
$y_1 = 2$	0.0130 [0.0113]	0.0156 [0.0122]	0.0113 [0.0085]	0.0399 [0.0320]
Col. sum	0.5406 [0.5905]	0.4063 [0.3640]	0.0530 [0.0454]	1

Table 4 provides the unadjusted joint probabilities and marginals obtained from the balanced panel (shown in brackets) along with the adjusted versions obtained from the linear RAN model.

Table 5: 3x3 RAN Model - Adjusted and [Unadjusted] Transition Probabilities

Annual Transitions Between 2001-Q1 and 2002-Q1 $x = \text{age 15 and over}$				
	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	0.8573 [0.8744]	0.1102 [0.0980]	0.0325 [0.0276]	1 [1]
$y_1 = 1$	0.1710 [0.1898]	0.7720 [0.7565]	0.0570 [0.0537]	1 [1]
$y_1 = 2$	0.3250 [0.3525]	0.3917 [0.3813]	0.2833 [0.2662]	1 [1]

The magnitudes of the biases in the balanced panel [discrepancies between  $f(y_1, y_2|D = 1, C = 3, x)$  and  $f(y_1, y_2|x)$ ] range between -24 and 11 percent. Six of the 9 cells have biases of 10% or more in absolute value.

In Table 5, the associated forward transition probabilities are shown. As in the previous table, the numbers in brackets are the unadjusted ones. Almost surely someone who views the evidence will argue that the differences between unadjusted and adjusted magnitudes are not large enough to warrant correction. It is worth noting that even though the picture of labor dynamics that emerges might not be different by some measure of closeness, the correction is still warranted because it produces a version which is fully consistent with the cross-section estimates. This capability of the RAN model is especially important in the case of statistical agencies like TURKSTAT, who refuse to exploit the short panel dimension of the HLFS on the grounds that there is no weighting method that can reconcile dynamic and static estimates.

As we argued above, the non-parametric feature of the RAN model is attractive, but it has the usual shortcomings that data based methods have. To illustrate the possible pitfalls, we consider another example, where  $x$  denotes males aged 35-54 who have high school education and reside in urban areas of Turkey. RAN model estimates for this partition of the sample are reported in Table 6. In this case the statistical evidence favors the hypothesis that attrition/substitution is ignorable. Note that the sample sizes are small, and consequently bootstrapped standard errors based on 100 replications are large. In fact in some cases the bootstrapped means are very different from the estimated parameter value (see  $\theta_{12}$  and  $\theta_{22}$  for the concave case). This finding exposes the well-known fragility of the bootstrap method when sample sizes are too small. In such cases it would be advisable to increase the number of bootstrap replications (say to 1,000) before passing

Table 6: Another 3x3 RAN Model - Parameter Estimates

Annual Transitions Between 2001-Q1 and 2002-Q1					
$x = \text{male, ages 35-54, high school education, residing in urban areas}$					
Parameter	$\theta_{00}$	$\theta_{11}$	$\theta_{12}$	$\theta_{21}$	$\theta_{22}$
(i) $w(\cdot)$ linear:					
Estimate	0.9472	-0.0234	0.1348	0.0688	0.3507
Bootstrap mean	0.9465	-0.0003	0.2731	0.0524	0.3638
Bootstrap std. error	0.1271	0.2530	0.5869	0.2220	0.4227
(ii) $w(\cdot)$ convex:					
Estimate	-0.0540	-0.0231	0.1191	0.0697	0.3127
Bootstrap mean	-0.0699	0.0028	0.1813	0.0646	0.2934
Bootstrap std. error	0.1345	0.2620	0.4179	0.2281	0.3471
(iii) $w(\cdot)$ concave:					
Estimate	-0.0518	-0.0239	0.1560	0.0681	0.4101
Bootstrap mean	-0.0378	-0.0037	2.6577	0.0415	2.6068
Bootstrap std. error	0.1340	0.2570	9.0427	0.2262	8.5999
Sample Sizes:					
Balanced panel	460				
First period cross-section	1,416				
Second period cross-section	1,440				

judgement on ignorability.

When the narrower objective of producing dynamic statistics consistent with the cross-section statistics is adopted, the correction can proceed despite our cautionary remark. In fact, the reflation factors for the subsample under examination reported in Table 7 point to a surprisingly consistent picture regardless of choice of functional form. Interestingly, small cell sizes that produced the fragility in the bootstrap stage rescues the reflation stage. For example, consider cell  $(y_1 = 2, y_2 = 2)$  for which substantial differences are observed across parametric forms (see the second digits after the decimal point reported inside braces). In the balanced panel there are only 4 individuals in this cell. Under the alternative parametric assumptions the adjusted fractions for this cell are respectively 0.012464, 0.012694 and 0.012198. Thus, as long as small cell sizes yield a small magnitude for  $p_{jk}$  ( $< 0.01$ , say), the differences in RAN model reflection factor estimates by functional form do not translate to comparable differences in the magnitudes of the adjusted fraction.

Table 7: Another 3×3 RAN Model - Reflation Factors

Annual Transitions Between 2001-Q1 and 2002-Q1				
$x = \text{male, ages 35-54, high school education, residing in urban areas}$				
	$y_2 = 0$	$y_2 = 1$	$y_2 = 2$	Row sum
$y_1 = 0$	$\begin{Bmatrix} 0.9471 \\ 0.9474 \\ 0.9468 \end{Bmatrix} 0.0978$	$\begin{Bmatrix} 1.0160 \\ 1.0158 \\ 1.0162 \end{Bmatrix} 0.0196$	$\begin{Bmatrix} 1.2978 \\ 1.2952 \\ 1.3011 \end{Bmatrix} 0.0087$	$f_1(0)$
$y_1 = 1$	$\begin{Bmatrix} 0.9237 \\ 0.9258 \\ 0.9214 \end{Bmatrix} 0.0652$	$\begin{Bmatrix} 0.9926 \\ 0.9927 \\ 0.9924 \end{Bmatrix} 0.7543$	$\begin{Bmatrix} 1.2744 \\ 1.2657 \\ 1.2843 \end{Bmatrix} 0.0239$	$f_1(1)$
$y_1 = 2$	$\begin{Bmatrix} 1.0819 \\ 1.0672 \\ 1.0989 \end{Bmatrix} 0.0109$	$\begin{Bmatrix} 1.1508 \\ 1.1443 \\ 1.1583 \end{Bmatrix} 0.0109$	$\begin{Bmatrix} 1.4326 \\ 1.4591 \\ 1.4021 \end{Bmatrix} 0.0087$	$f_1(2)$
Col. sum	$f_2(0)$	$f_2(1)$	$f_2(2)$	1

## 5 Findings From a Broader Investigation

As can be inferred from our second example, in our broader empirical investigation we exposed the parametric features of RAN model to a torture test by choosing  $x$  to identify smaller and smaller segments of the population. This exercise is warranted, because statistical agencies often publish official statistics broken down by a high dimensional  $x$ . The question is whether RAN model can rise to the challenge of yielding their dynamic counterparts.

The covariates we studied included sex (male, female), location (urban, rural), education (4 categories) and age (5 groups). Notably, the RAN model yielded extremely robust results as long as cell counts in the balanced panel remained within acceptable ranges for the sample sizes under investigation. In extreme cases when cell sizes were extremely small, we ran into occasional convergence problems during the bootstrap stage. This problem was attributable to the fact that some bootstrapped samples yielded zero cell counts, in which case correction could not proceed. Clearly zeros encountered during bootstrapping are random as opposed to structural zeros. We were able to fix the problem by adding an observation to the empty cell and adjusting the sample size accordingly.

The fix we developed was also useful in higher dimensional RAN models (up to 5×5) we experimented with. Clearly empirical findings regarding the nature of attrition/substitution can, and do vary, from one time period to the other, and with choice of  $x$ . With sufficient data, a second stage analysis can be performed to shed light on the patterns (see Ikizler and Tunali, 2012 and

Gokce and Tunali, 2012 for substantive examples from  $4 \times 4$  RAN models). The fragility exposed in Table 6 suggests that the number of bootstrap replications we used (100) may not be adequate for credibly testing whether attrition/substitution is nonignorable. Nonetheless there are valid reasons for proceeding with the correction whether or not attrition/substitution is ignorable. Overall, our non-parametric approach with respect to  $x$  worked extremely well. In our systematic examination of annual and quarterly transitions over the 2000-2002 period, we discovered that the RAN model produced estimates of transition rates for commonly used partitions of the full sample (jointly by sex and location, by education, by broad age groups) that are robust to choice of functional form. Even further partitioning of the subsamples identified by sex-location pairs either by education, or by broad age groups, proved to be feasible. Thus our method is worthy of adoption for statistical and policy analysis purposes.

## 6 Conclusion

In this paper we tackle a generalized version of the attrition problem, typically associated with longitudinal data. The motivation for the generalization comes from the observation that many sustained large scale data collection efforts (CPS and the European Union Statistics on Income and Living Conditions (EU-SILC) being some well-known examples) involve multiple visits to the same address/household over a short period of time. Another feature of these efforts is the use of a rotational design whereby a fresh set of addresses/households are systematically added to, and excluded from, the sample frame according to a predetermined schedule. Notably these data sets have a short panel component that can support dynamic analyses. What stands in the way is the concern that the balanced panel which can be used for tracking the dynamics may not be representative of the population at a given point of time. The generalization we offer recognizes that proper use of such short panels requires corrections for non-response after initial response (attrition) as well as response after initial non-response (substitution). Furthermore, attrition/substitution behavior is allowed to be endogenous to the outcomes of interest.

In our empirical example outcomes are labor market states occupied by an individual. Endogeneity implies that particular labor market outcome combinations could make individuals more or less prone to exclusion from the balanced panel. The model we use exploits the set-up and key

insights in Hirano et al. (2001) but departs from it in its computational simplicity, especially when the linear version is adopted. The correction amounts to reflatting the balanced panel fractions (cell means) by factors expressed as a parametric function of the states under examination. Our empirical investigation of annual transition data from the Household Labor Force Survey in Turkey showed that attrition/substitution is a serious concern, in the sense that transition rates obtained from the balanced panel are systematically distorted. The RAN model based adjustment not only corrects these distortions, but it also reveals the attrition and substitution patterns. Based on our systematic empirical investigation, results did not display sensitivity to the parametric features of the RAN model. Thus the linear version – which is extremely simple to implement – appears suitable for empirical work.

Another attractive feature of the RAN model is the non-parametric treatment of covariates (such as sex, location, age groups, etc.). That is, each distinct covariate combination is associated with its own set of parameters and reflation factors. In a nutshell, RAN model is designed to produce estimates of transition rates which are consistent with cross-section statistics, conditional on covariates of interest. As such it is likely to gain the approval of official statistical agencies.

Furthermore, estimation does not require micro data. To implement the adjustments, it is sufficient to have the joint distribution obtained from the balanced panel that links the two legs of the short panel alongwith the marginal distributions obtained from representative data collected at each leg. To do inference, we additionally need the sample sizes that yielded the three distributions we work with. Since all of this information is readily available from statistical agencies in tabular form, the proposed methodology should appeal to a very broad audience.



## 7 References

- Abowd, J. M. and A. Zellner (1985) "Estimating Gross Labor-Force Flows." *Journal of Business and Economic Statistics*, 3:3, 254-283.
- Bhattacharya, D. (2008) "Inference In Panel Data Models Under Attrition Caused by Unobservables." *Journal of Econometrics*, 144: 430-446.
- BLS - Bureau of Labor Statistics (2002) *Design and Methodology: Current Population Survey*. Technical Paper 63 RV, U.S. Department of Labor and U.S. Department of Commerce.
- Chen, K. (2001) "Parametric Models for Response-Biased Sampling." *Journal of Royal Statistical Society, B*, v. 63, Part 4, 775-789.
- Efron, B. (1979) "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics*, 7: 1, 1-26.
- EUROSTAT (2007) *Labor Force Survey in the EU, Candidate and EFTA Countries; 2007 Edition*. Office for Official Publications of the European Communities: Luxembourg.
- Fitzgerald, J., P. Gottschalk, and R. Moffitt (1998) "An Analysis of Sample Attrition in Panel Data." *Journal of Human Resources*, 33:2, 251-299.
- Gokce, O. Z and I. Tunali (2012) "Informality and Labor Market Mobility in Turkey: Evidence from Micro Data, 2000-2002."
- Hausman, J. A. ve D. A. Wise (1979) "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment." *Econometrica*, 47:2, 455-473.
- Heckman, J. (1987) "Selection Bias and Self-selection." In J. Eatwell, M. Milgate, ve P. Newman (Ed.), *The New Palgrave: A Dictionary of Economics*, Vol. IV. London: McMillan.
- Hirano, K., G. W. Imbens, G. Ridder, and D. B. Rubin (2001) "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica*, 69: 6, 1645-1660.
- Ikizler, H. and I. Tunali (2012) "Agricultural Transformation and Labor Mobility During The ARIP Period in Turkey: Evidence From Micro-Data, 2000-2002."

- Little, R. J. A. (1982) “Models for Nonresponse in Sample Surveys.” *Journal of the American Statistical Association*, v. 77, no. 378, 237-250.
- Little, R. J. A. (1993) “Post-stratification: A Modeller’s Perspective.” *Journal of the American Statistical Association*, v. 88, no. 423, 1001-1012.
- Little, R. J. A. and D. Rubin (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
- Little, R. J. A. and M. Wu (1991) “Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ.” *Journal of the American Statistical Association*, v. 86, no. 413, 87-95.
- Madow, W., I. Olkin, and D. Rubin (Eds.) (1983) *Incomplete Data in Sample Surveys* (3 volumes). New York: Academic Press.
- Rubin, D. (1976) “Inference and Missing Data.” *Biometrika*, 63: 581-592.
- Stasny, E. A. (1986) “Estimating Gross Flows Using Panel Data With Nonresponse: An Example from the Canadian Labor Force Survey.” *Journal of the American Statistical Association*, v. 81, no. 393, 42-47.
- Stasny, E. A. (1988) “Modeling Nonignorable Nonresponse in Categorical Panel Data With an Example in Estimating Gross Labor-Force Flows.” *Journal of Journal of Business and Economic Statistics*, 6:2, 207-219.
- Tunali, I. (2009) “Analysis of Attrition Patterns in the Turkish Household Labor Force Survey, 2000-2002.” Ch. 6 in *Labor Markets and Economic Development*, edited by R. Kanbur and J. Svejnar, 110-136. London and New York: Routledge.
- TURKSTAT (Turkish Statistical Institute) (2001) *Household Labor Force Survey: Concepts and Methods*. Ankara: State Institute of Statistics.

## Appendix

Let  $A_j$  denote the  $5 \times 5$  partition of the  $A$  matrix defined implicitly by equation (20) with the  $j$ th row removed, and let  $\underline{b}_j$  denote the  $5 \times 1$  partition of vector  $\underline{b}$  with the  $j$ th row removed,  $j = 1, 2, \dots, 6$ .

With this notation, the system with the 6th equation removed can be expressed as  $A_6\theta = \underline{b}_6$  and has the explicit form given below:

$$\begin{bmatrix} \sum_{j=0}^2 p_{0j} & 0 & 0 & p_{01} & p_{02} \\ \sum_{j=0}^2 p_{1j} & \sum_{j=0}^2 p_{1j} & 0 & p_{11} & p_{12} \\ \sum_{j=0}^2 p_{2j} & 0 & \sum_{j=0}^2 p_{2j} & p_{21} & p_{22} \\ \sum_{k=0}^2 p_{k0} & p_{10} & p_{20} & 0 & 0 \\ \sum_{k=0}^2 p_{k1} & p_{11} & p_{21} & \sum_{k=0}^2 p_{k1} & 0 \end{bmatrix} \begin{bmatrix} \theta_{00} \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{bmatrix} = \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_2(0) \\ f_2(1) \end{bmatrix}.$$

Given what we know about marginal and joint distributions, it is straightforward to see that  $\text{rank}(A_6) = 5$ . Thus the solution to the reduced system of equations is unique and is given by  $\hat{\theta} = A_6^{-1}\underline{b}_6$ . Next, we define the following 5x5 pivot matrices:

$$E_1 = \begin{bmatrix} -1 & -1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ -1 & -1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, E_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$E_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & -1 & -1 \end{bmatrix}.$$

It is straightforward to show that for  $j = 1, 2, \dots, 5$ ,  $E_j A_j = A_6$ , and  $E_j \underline{b}_j = \underline{b}_6$ . Since the pivot matrices are of full rank, this proves that all six systems are equivalent, and yield the same unique solution  $\hat{\theta}$ .