Stuart Sweeney
Institute for Social, Behavioral, and Economic Research
University of California
Santa Barbara, CA 93111

Smoothing Period Parity Progression Ratios for Survey Data from Developing Countries

Stuart Sweeney[12]
[1] *Department of Geography, University of California, Santa Barbara, CA 93106-4060*
[2] *Institute for Social, Behavioral, and Economic Research, UC Santa Barbara*

keywords: Period parity progression ratios, DHS, local likelihood hazard estimation.

Abstract

Fertility stalls are typically identified from a time series of TFR. Since stalls universally occur in developing countries, the surveys used to calculate TFR appear at irregular intervals and have complex survey designs. An alternative measure is the period parity progression ratio that can be used to identify long run patterns covering years without surveys. Standard estimation of period PPRs use each survey separately, and the overlapping estimates from different surveys will likely be misaligned and each will be biased downwards near the interview year. This paper describes and evaluates a pooled-survey estimator of period PPR based on local likelihood estimation. The estimator yields a single composite period PPR that spans the full set of survey years, and uncertainty is captured and expressed as part of the estimation process. Following the description of the statistical theory, the estimator is assessed using simulated data and applied to Guatemala.

Introduction

One of the basic precepts of demographic transition theory is that once a country commences along its development trajectory, the associated demographic trends will follow – initially declining mortality rates, and then after some delay declining fertility rates (Kirk, 1996). The transition is suppose to proceed without interruption once it commences. It is problematic for the theory – as it is for any social science theory – when a large number of examples counter to the theory are found. This has been the case with the fertility component of the transition. Stalled transitions – meaning either that fertility rates stabilized at a level well above and out of balance with mortality, or that fertility rates have increased – have been identified now in several developing countries. Of course, simply identifying stalls is not an immediately threat to the entire theory, but certainly if stalls are being found it is important to study them. And studying them requires having methods to identify the direction and pace of change in fertility, and to decompose and isolate components of changing fertility to further understanding of why a stall occurs, or perhaps why a stalled fertility transition resumes its descent.

This paper presents new method of estimation for period parity progression ratios that is primarily intended to support the study of fertility stalls in developing countries. Period parity progressions ratios (a_{jt}) measure the probability that a woman at parity j in period t will have another birth. Because a move to the next parity generally represents a choice informed by ideal family size and other considerations, the progression ratios reflect the incremental family building process as it unfolds through time. Compared to other period fertility measures (e.g. TFR, ASFR) that are produced only for the data collection year, progression ratios capture a more refined and complete temporal resolution and support a decomposition of period fertility behavior by parity.

As typically applied in published studies (Ní Bhrolcháin, 1987; Feeney and Yu, 1987; Spoorenberg, 2010) and textbooks (Hinde, 1998), period parity progression ratios are estimated for a single sample of data and can be used to characterize decades of fertility behavior prior to the year of data collection. As concern over potentially stalled fertility transitions builds and becomes the focus of academic research, it would seem that progression ratios provide a perfect means for such assessments. Recent extensions of parity progression ratios to incorporate tempo-effects (Kohler and Ortega, 2002) and applications have focused on low fertility regimes in Europe and are generally based on comprehensive data archives or registry systems. There are relatively few applications to developing countries, beyond the original work of Feeney and Yu (1987) where the data is restricted to national-level sample surveys such as the World Fertility Surveys – WFS, Demographic Health Surveys – DHS, and Reproductive Health Surveys – RHS. Existing applications such as Spoorenberg (2010) or Hinde (1998) use single DHS or RHS samples and rely on standard direct estimators that yield estimates for approximately 20 years prior to the sample up to the date of the survey.

The remainder of this paper introduces a pooled-survey estimator for period PPRs using local likelihood hazard estimation. The second section discusses the structure of DHS and RHS surveys, the sample design and associated weighting, and demonstrates – using data from several waves of Guatemala DHS/RHS – that retrospective questions used to measure birth timing also mean that the information contained in pooled data is strongly overlapping. It is the overlapping nature of the separate surveys that we leverage to estimate long term smooth patterns of change, and to reduce uncertainty in the those estimates. The second section presents the statistical theory including some details of local likelihood estimation. The third section of the paper evaluates the method using simulated data. The simulated data will be initially from parametric distributions that we think most closely reflect birthing as a recurrent event. We will also use subsampling of the simulated data in an attempt to replicate the complex survey design used in most DHS. That will be used to evaluate the properties of the bootstrap resampling estimator that we propose accounts for the differential sampling weight in DHS and RHS. The paper closes with a brief application using data from Guatemala.

Health Survey Data in Developing Countries

Since fertility stalls are found only in developing countries – at least to date – we need to consider the available data that can be used to study and characterize their occurrence. The only data available in most cases is either DHS or RHS. The DHS and RHS data are virtually identical in their construction for our purposes. Each woman in a household between the ages of 15-49 (44 in some countries or surveys) is asked a series of detailed questions about her reproductive history and date of first union. Some of the surveys are self-weighting (all weights=1) but the most are based on complex survey designs and include person weights. Survey designs and the content, or exact wording, of some surveys does change through time – currently there are six different "waves" of DHS instruments and each has an associated manual (cite). The number of surveys available in any country country vary, but it is almost always the case that spacing between survey years is irregular.

The goal of methods proposed below is to leverage the additional information and stability available by pooling surveys over years. Separate parity-specific data files can then be constructed that only include women who have achieved that parity. Our representation of birth intervals follows standard protocols for survival data. T_i records the elapsed time in months between parity j to $j+1$ measured from calendar time starting at calendar month code m_i and total observed months for the pooled survey running from 0 to M months. The basic information set for each woman is $(Y_i, \delta_i, m_i, w_i)$ where $Y_i = min(T_j, c_j)$, c_i is the duration from m_i to the survey interview date, δ_i indicates whether $T_i < c_i$, and w_i is a survey weight. The censoring times c_i vary widely since the five surveys are pooled and interview dates for any given survey may range over a year or more.

Statistical Theory

Period Parity Progression Ratios: Direct Estimation

The method as described in Ní Bhrolcháin (1987); Feeney and Yu (1987); Hinde (1998) closely parallels occurrence-exposure rate definitions for period life table construction with parity specific birth rates defined assuming uncensored birth intervals. Defining $v = \{0, ..., M\}$ and pairs $\{v', v''\} \in v$ with $v'' \geq v'$, the direct birth rate is,

$$q_{v,v'} = \frac{\sum_i I[(m_i = v') \ \& \ (T_i = v''-v')]}{\sum_i I[(m_i = v') \ \& \ (T_i \geq v''-v')]} \tag{1}$$

where $I[\ ]$ is an indicator function taking the value 1 if true and 0 otherwise. The numerator is the number women (births) progressing to parity $j+1$ in period v'' who had birth j in period v', and the denominator is a measure of the exposure remaining among the initial "cohort" defined as women having birth j in period v'. The period parity progression ratio can be defined as $a_{j,v'} \approx \prod_{v''}(1 - q_{v',v''})$ (true) or as $a_{j,v''} \approx \prod_{v'}(1 - q_{v',v''})$ (synthetic). Both approximations rely on discrete estimates $(q_{v',v''})$ to the period parity-specific birth hazard. The final estimates combine the true and synthetic estimates: $a_{\tilde{v'}} = \{a_{v' \leq median(v')}, a_{v''>median(v')}\}$.

In uncensored and self-weighting survey data, the rate estimate (1) would be correct. The estimate from a complex survey with case weights is more problematic especially when samples are pooled. The equivalent weighted estimated based on censored data would be,

$$q^s_{v',v''} = \frac{\sum_i w_i * I[(m_i = v') \ \& \ (Y_i = v''-v') \ \& \ (\delta_i = 1)]}{\sum_i w_i * I[(m_i = v') \ \& \ (Y_i \geq v''-v')]} \tag{2}$$

Note that in (2) the numerator is correct but the denominator is biased upward because women present at the start of the interval may be censored prior having their $j + 1^{st}$ birth. This problem is recognized in the literature with the recommendation being to trim the sample to exclude any birth j too close to the censoring date. This is also the justification for introducing the synthetic estimates. Censoring is only one problem with the direct estimates. As with any hazard estimate, the population at risk is shrinking as durations increase. Since each rate birth rate is a binomial trial, the variance increases with interval length because the number at-risk ('trials') is smaller. In addition to downward bias introduced in any estimate of (2) from censoring, there is a trade-off in direct estimates of $a_{\tilde{v'}}$ between bias introduced by excluding longer durations and increased variance from including longer durations.

The desire to recover smooth functions from direct occurrence-exposure estimates has a long tradition in demography (see Hoem et al., 1976). Smoothing and pre-smoothing estimators for hazard rate estimation have received considerable attention in the biostatistics literature and methods have evolved in conjunction with methods for density estimation (Cao et al., 2005; Loader, 1996, 1999; Müller et al., 1997; Wang, 2005; Wang et al., 1998). Application areas within demography have generally focused on smoothing age-specific rates, and mortality specifically.

Life table smoothing approaches provide a way to deal with the bias-variance trade-off alluded to at the end of the previous section (Müller et al., 1997; Wang, 2005; Wang et al., 1998). In Wang et al. (1998) local polynomial models are used to smooth rate estimates and weights proportional to the population at risk are used to deflate high variance in the tails. Also, similar to Wang et al. (1998), we are interested in estimating entire hazard surfaces rather than isolated hazard functions. Specifically, for each parity we would like to recover the hazard rate surface formed by T month-specific hazard rate functions. In Wang et al. (1998) this is accomplished using a two-step estimator: estimating smooth hazard curves for each cohort life table in the first step and then smoothing across cohorts in the second step. In their work, they were constrained to work with direct estimates because their basic units of data were life tables.

In our case we have access to individual records and while time is measured discretely, relative to the scale of the process under study we can work with the process as if it is unfolding in continuous time. Specifically, we use a local polynomial approximation to the log-likelihood of the hazard surface Loader (1996),

$$\begin{aligned}\mathcal{L}_{t,x}(b) = & \textstyle\sum_{i=1}^{n} W\left(\frac{Y_i-t}{h}, \frac{x_i-x}{h}\right) \langle b, B(Y_i-t, x_i-x)\rangle \\ & -\textstyle\int_{i=1}^{n} N(u) W\left(\frac{Y_i-t}{h}, \frac{x_i-x}{h}\right) e^{\langle b, B(Y_i-t, x_i-x)\rangle} du\end{aligned} \qquad (3)$$

where t is duration since birth k, x is the period at birth k, $W()$ is a kernel weighting function[1], h is the kernel bandwidth, and b is vector of locally weighted coefficients associated with the polynomial covariates $B(t,x) = (1\ t\ x\ t^2\ xt\ x^2)^T$. At each temporal pair (t,x) we recover an estimate of the hazard surface as $\hat{\lambda}(t,x) = e^{\langle \hat{b}, B(0,0)\rangle}$. We agree with Loader's (1996; 1999) stance arguing against

[1] We use tricube weights.

the use of plug-in estimators for the bandwidth and instead using a variety of diagnostic plots combined with priors to select the degree of smoothing. In two applications of the method by the author (Sweeney and Grace, 2013; Grace and Sweeney, 2013) we use nearest-neighbor methods that allow the bandwidth the increase in the higher variance tails of the distribution (when t is large) but capture details of the hazard surface shape when the population at risk is still large. Automatic selection of the degree of the polynomial was also used in those application and is used below in the evaluation against simulated data.

In comparison to (2), the estimated hazard surface approach provides a solution to heteroskedastic rate estimates and incorporates censoring directly into the estimator. Estimation of period parity transition probabilities are recovered using numerical integration at each period, $\hat{a}_{jx} = 1 - \int_0^\infty \hat{\lambda}_j(u, x)du$; the continuous analog of the discrete direct estimate $1 - \prod_{v''}(1 - q_{v',v''})$. The censoring issues near the last survey are less sever in pooled estimator approach because smoothing in the direction of x results in predictions for the tails of hazard functions near to M. Still, censoring is an issue and hazard rate functions are simply truncated for long durations. Similar to above we can define 'synthetic' progression ratio estimates by defining a diagonal along the rate surface such that period, x is held constant at the point of the $j+1$ birth; thus $\hat{a}_{jx'} = 1 - \int_0^\infty \hat{\lambda}_j(u, x-u)du$. A final estimator can be defined, similar to the direct estimator, that switches from the 'true' to the 'synthetic' estimates at their smallest point of separation. Unlike direct estimates from single surveys, we only have to deal with the rigid end of period censoring once as opposed to for each survey. Censoring around the interview dates in the other surveys is handled directly by the method and of little impact because information is shared across surveys.

The degree of overlap for an example set of pooled-DHS/RHS surveys can be seen using real data from Guatemala (see Figure 1). In this case there are five DHS or RHS surveys at irregular intervals; 1987, 1995, 1998, 2002, and 2008. The colored bands indicate the period when the survey enumerators where in the field collecting data. Data shown is for birth intervals at parity 5, and the durations are arranged into blocks horizontally associated with each survey year. The point here is that the information about birth intervals collected from women in surveys from 1995 and later adds to the complete set of information that can be used to estimate period PPRs in years prior to the 1987 survey. Also, if we were restricted to using only the 1987 survey alone, the resulting estimate would have a wider variance and would suffer because of censored intervals at 1987. The same logic hold for other survey years except 2008.

The hazard surface estimation incorporates sampling weights as part of our inferential framework. We use a weighted pairs bootstrap to extract samples of observation identifiers prior to creation of the parity-specific data files. Weights are proportional to the case weights and rescaled so that the sum of weights within each DHS/RHS year are equal to the number of observations. We compute hazard surfaces for each parity and subgroup and the associated progression ratios for 1000 bootstrap samples. Confidence envelopes, shown in the figures referenced in the application section, are defined by the 0.025 and 0.975 quantiles of the bootstrap distributions at each month. Our weighted estimate of the progression ratios are the medians of those distributions. The evaluation of the estimator using simulated data uses the same strategy. However, with simulated data we can compare results using a complex survey design and case weights against a design without case weights. More information about the construction of the simulated data is provided below.

Period Completed Fertility and Decomposition Analysis

In the application we use the period PPRs to construct period completed fertility. Period completed fertility is defined similarly to Ní Bhrolcháin (1987) and Feeney and Yu (1987). That is,

$$F_x = L_x a_{mx} \sum_{j=0}^{8} \prod_{i=0}^{j} a_{ix}$$

where L_x is the ratio of total 1^{st} births to 1^{st} births after first unions and a_{mx} is the probability that a woman will be in union by age 25. For L_x we fit a locally polynomial Poisson regression using total 1^{st} births in period t as an offset (so a log-rate model) and then take the inverse of the predicted value at each x. For a_{mx} we fit a hazard surface as above but with the duration, Y_i, measured from age 10 to the date of union or interview. The probability is then recovered as above by numerically integrating for each x over $0 \leq t \leq 180$.[2]

We compare plots of F_x from 1970 to 2005 against the period TFR and tempo-adjusted period TFR calculated for each survey. The F_x and plots of each parity transition are used to identify the time and duration of stalls and resumption of declines. We are also interested in decomposing the the components of a_{jx}. We use Horiuchi et al. (2008) approach based on the line integral model

[2] The duration is 0 at age 10 and 180 at age 25.

of decomposition. Parity specific decomposition effects on a_{jx} were constructed for 6 month time steps and then aggregated into larger epochs.

In addition to providing confidence envelopes for our estimates of $\hat{a}_{jx}$ and $\hat{F}_x$, we would also like to formally test whether the differences in observed period completed fertility between population subgroups are statistically significant. The relevant subgroups in the application to Guatemala are indigenous and Ladino. This is accomplished using a randomization test. Defining our subgroups as indigenous (A) and Ladino (B) our test controls for the possible confounding effects of a three-level education factor variable (E) and a two-level urban/rural factor (U). For each subgroup we define $5 \times 3 \times 2$ cross-classification tables $N^A_{Y \times E \times U}$ and $N^B_{Y \times E \times U}$. R replicates from the full sample are drawn without regard to ethnicity but matching the observed DHS/RHS year by education by urban record counts. For each pair of replicates, we define $D(y)_{\{R\}} = F(y)^A_{\{R\}} - F(y)^B_{\{R\}}$. The 0.025 and 0.975 quantiles of $D(y)_{\{R\}}$ at each value of y are used to evaluate the observed difference, $D(y)$, against 95% confidence envelopes. Values of $D(y)$ falling outside the confidence envelopes are interpreted as significant differences between Ladino and indigenous period completed fertility that are not attributable to differences in educational or rural/urban composition between the two groups.

Computation

All computation was carried out in R version 3.0 (R Core Team, 2013). Hazard surfaces were estimated using the package **locfit** (Loader, 2013). All other components of methodology were encoded in R functions written by the author.

Simulation Data

This part of the paper still needs to be completed. I have pulled the relevant articles from the literature on simulating survival data (Crowther and Lambert, 2013; Metcalfe and Thompson, 2006; Hess et al., 1999; Burton et al., 2006). There are existing libraries of code in R for simulating survival data but most make simplifying assumptions about recurrent events and I expect that I will end up writing my own set of functions. I have also been considering different functional forms and their ability to simulate realistic birth intervals. Both the log-logistic and sickle rate models with starting thresholds look promising (Billari, 2001a,b). Once the simulated data is generated, subsampling relative to some group indicators will be used to impose features of a complex survey design. Code for the estimator is already complete so the evaluation against simulated data will proceed quickly once the data simulation is complete.

Application: Guatemala's Fertility Stall and Resumption of Transition

Sweeney and Grace (2013) is already under review in a journal and focuses on the fertility stall and resumption of transition in Guatemala. The application section will reference figures from that paper. Figure 2 shows the estimated period completed fertility with period TFR overplotted. It demonstrates the ability of the estimator to deliver smooth estimates spanning non-survey years with error bounds, and that it accords well with period TFR estimated in survey years. Figure 3 shows the underlying hazard surfaces estimated for the indigenous population in Guatemala. These can be considered a pre-smoothing stage prior to estimation of the period PPPRs, which are derived as described in the methods section above. Figure 4 contains the estimated period PPPRs with uncertainty bounds. Figure 5 provides an example of the resampling test of the difference in period completed fertility for two groups while controlling for confounding. Figure 6 is an illustration of decomposition analysis – attributing change in period completed fertility to specific parities.

References

Francesco C Billari. A sickle transition-rate model with starting threshold. *Statistical Methods and Applications*, 10(1-3):139–155, 2001a.

Francesco C Billari. A log-logistic regression model for a transition rate with a starting threshold. *Population studies*, 55(1):15–24, 2001b.

Andrea Burton, Douglas G Altman, Patrick Royston, and Roger L Holder. The design of simulation studies in medical statistics. *Statistics in medicine*, 25(24):4279–4292, 2006.

Ricardo Cao, Ignacio López-de Ullibarri, Paul Janssen, and Noël Veraverbeke. Presmoothed kaplan–meier and nelson–aalen estimators. *Journal of Nonparametric Statistics*, 17(1):31–56, 2005.

Michael J Crowther and Paul C Lambert. Simulating biologically plausible complex survival data. *Statistics in medicine*, 2013.

Griffith Feeney and Jingyuan Yu. Period parity progression measures of fertility in China. *Population Studies*, 41(1):77–102, 1987.

K. Grace and S. Sweeney. Stalling fertility transitions in West Africa: A parity-specific analysis of the region with decomposition. Refereed article (submitted September 2013), 2013.

Kenneth R Hess, Dan M Serachitopol, and Barry W Brown. Hazard function estimators: a simulation study. *Statistics in medicine*, 18(22):3075–3088, 1999.

Andrew Hinde. *Demographic methods*. Hodder Arnold Publication, 1998.

Jan M Hoem, Niels Keiding, Hannu Kulokari, Bent Natvig, Ole Barndorff-Nielsen, and Jørgen Hilden. The statistical theory of demographic rates: A review of current developments [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 169–185, 1976.

Shiro Horiuchi, John R Wilmoth, and Scott D Pletcher. A decomposition method based on a model of continuous change. *Demography*, 45(4):785–801, 2008.

Dudley Kirk. Demographic transition theory. *Population Studies*, 50(3):361–387, 1996.

Hans-Peter Kohler and José A Ortega. Tempo-adjusted period parity progression measures, fertility postponement and completed cohort fertility. *Demographic Research*, 6(6):91–144, 2002.

Catherine Loader. *locfit: Local Regression, Likelihood and Density Estimation.*, 2013. URL `http://CRAN.R-project.org/package=locfit`. R package version 1.5-9.1.

Clive Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

Clive Loader. *Local regression and likelihood*. Springer, 1999.

Chris Metcalfe and Simon G Thompson. The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in medicine*, 25(1): 165–179, 2006.

Hans-Georg Müller, Jane-Ling Wang, and William B Capra. From lifetables to hazard rates: The transformation approach. *Biometrika*, 84(4):881–892, 1997.

Máire Ní Bhrolcháin. Period parity progression ratios and birth intervals in England and Wales, 1941–1971: A synthetic life table analysis. *Population Studies*, 41(1):103–125, 1987.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

Thomas Spoorenberg. Fertility transition in India between 1977 and 2004. *Population (english edition)*, 65(2):313–331, 2010.

S. Sweeney and K. Grace. Ethnic dimensions of Guatemala's stalled transition: A parity-specific analysis of Ladino and indigenous fertility regimes. Refereed article (submitted September 2013), 2013.

Jane-Ling Wang. *Smoothing hazard rates*, pages 1–11. John Wiley & Sons, Ltd., 2005. doi: 10.1002/0470011815.b2a11069.

Jane-Ling Wang, Hans-Georg Müller, and William B Capra. Analysis of oldest-old mortality: Lifetables revisited. *The Annals of Statistics*, 26(1):126–163, 1998.

List of Figures

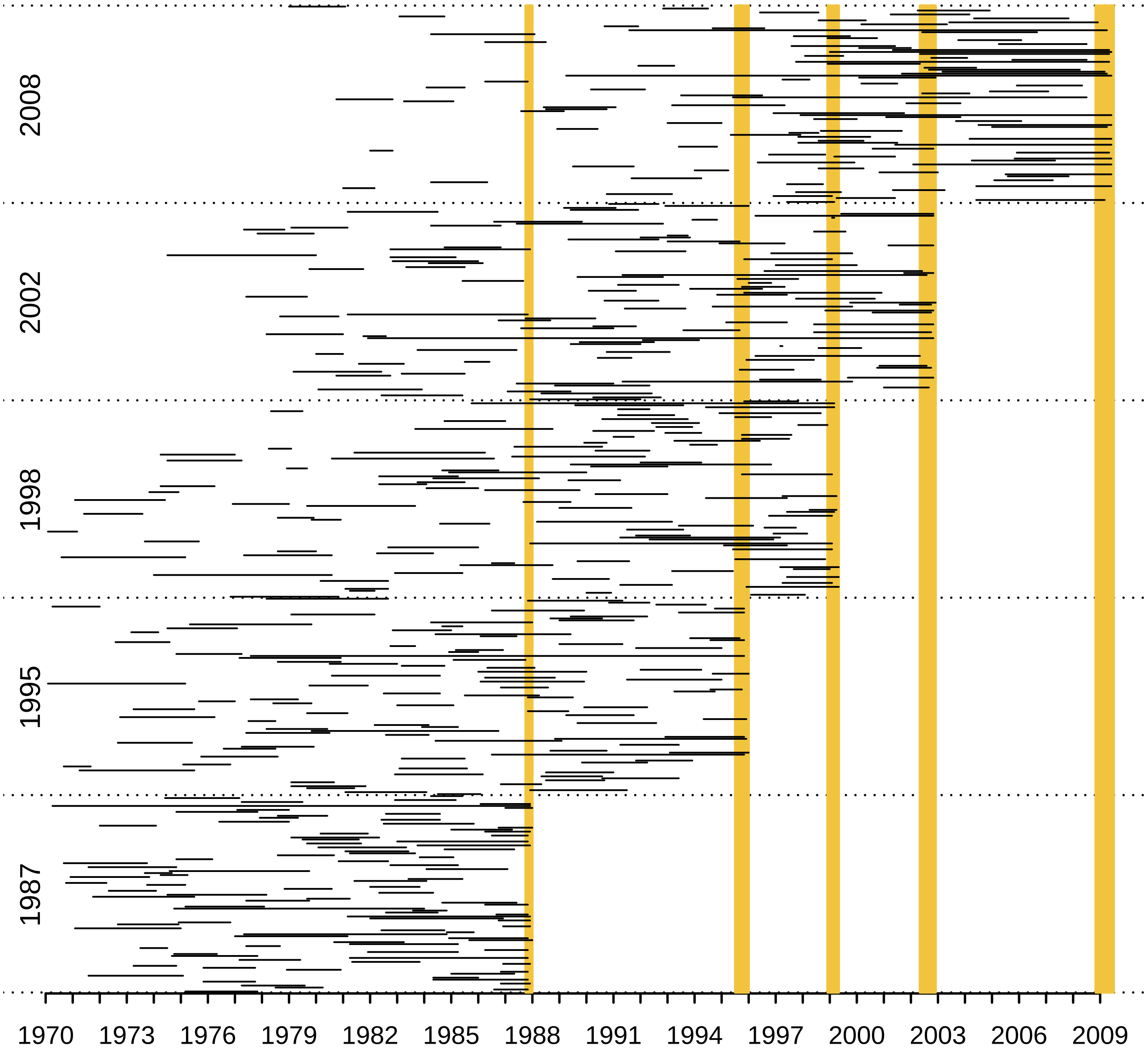


Figure 1: Overlapping samples for parity 5 from five Guatemala DHS/RHS surveys

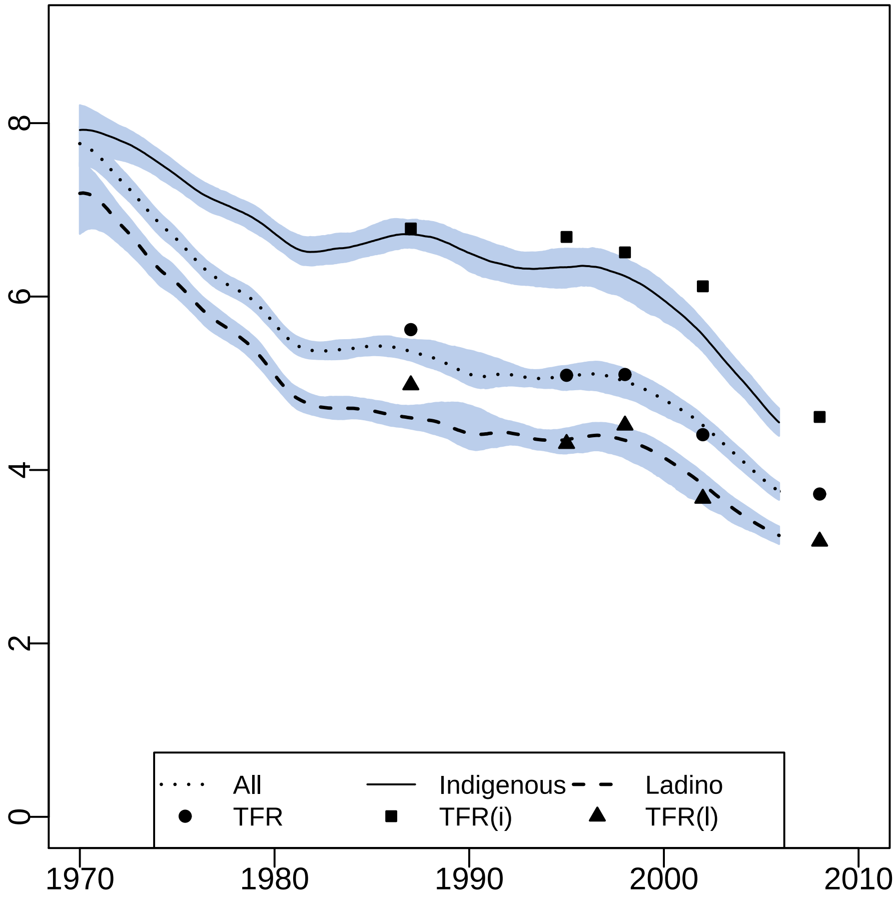
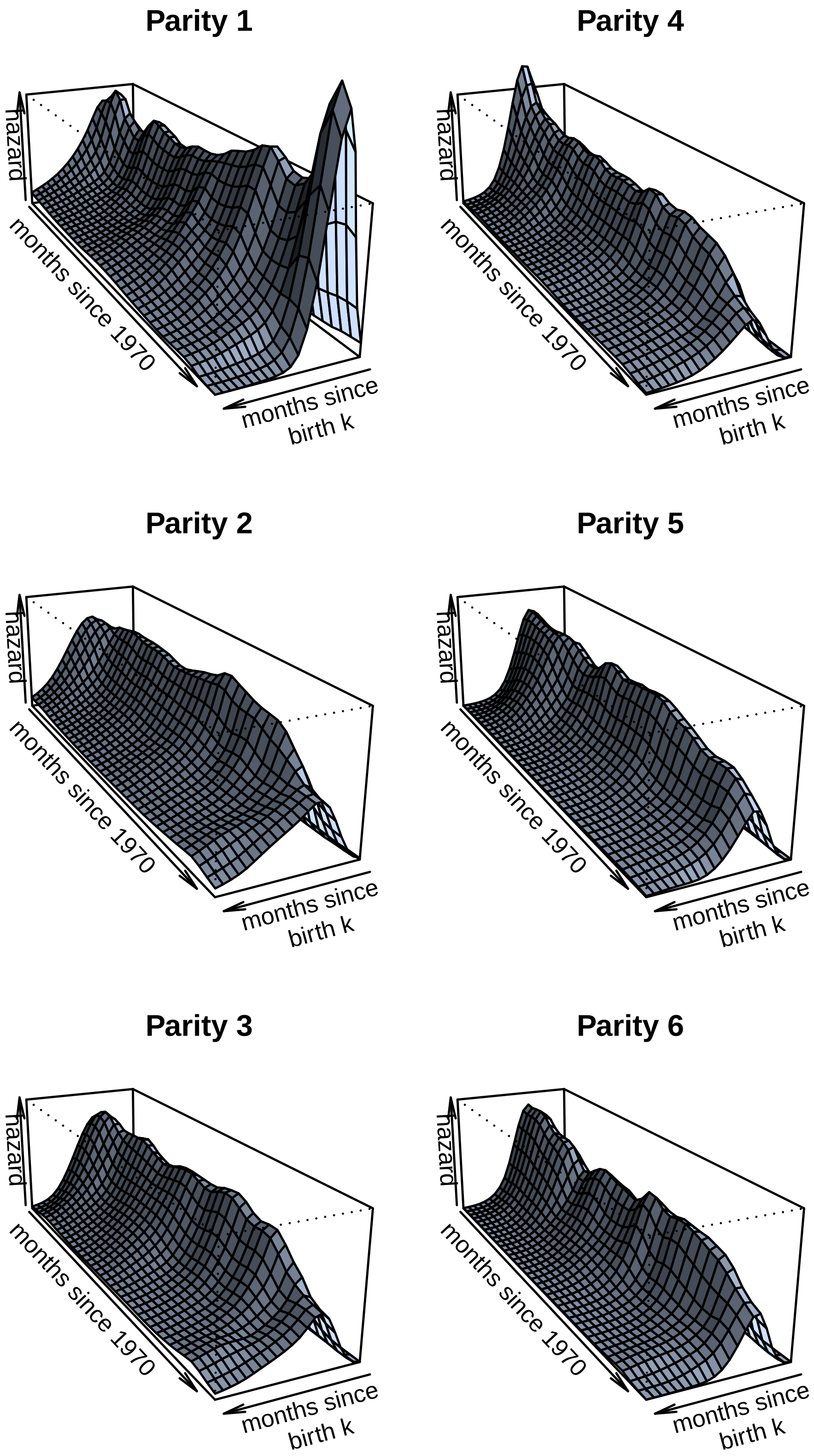


Figure 2: Period completed fertility, ethnicity – weighted

Note: Gray shading indicates the confidence envelopes defined by the 0.025 to 0.975 quantiles of the weighted pairs bootstrap as defined in the text.



Parity 1

Parity 4

Parity 2

Parity 5

Parity 3

Parity 6

Figure 3: Hazard surfaces, Indigenous women

Note: Months since birth k spans 0 to 120. Gridlines are at 6 month intervals.

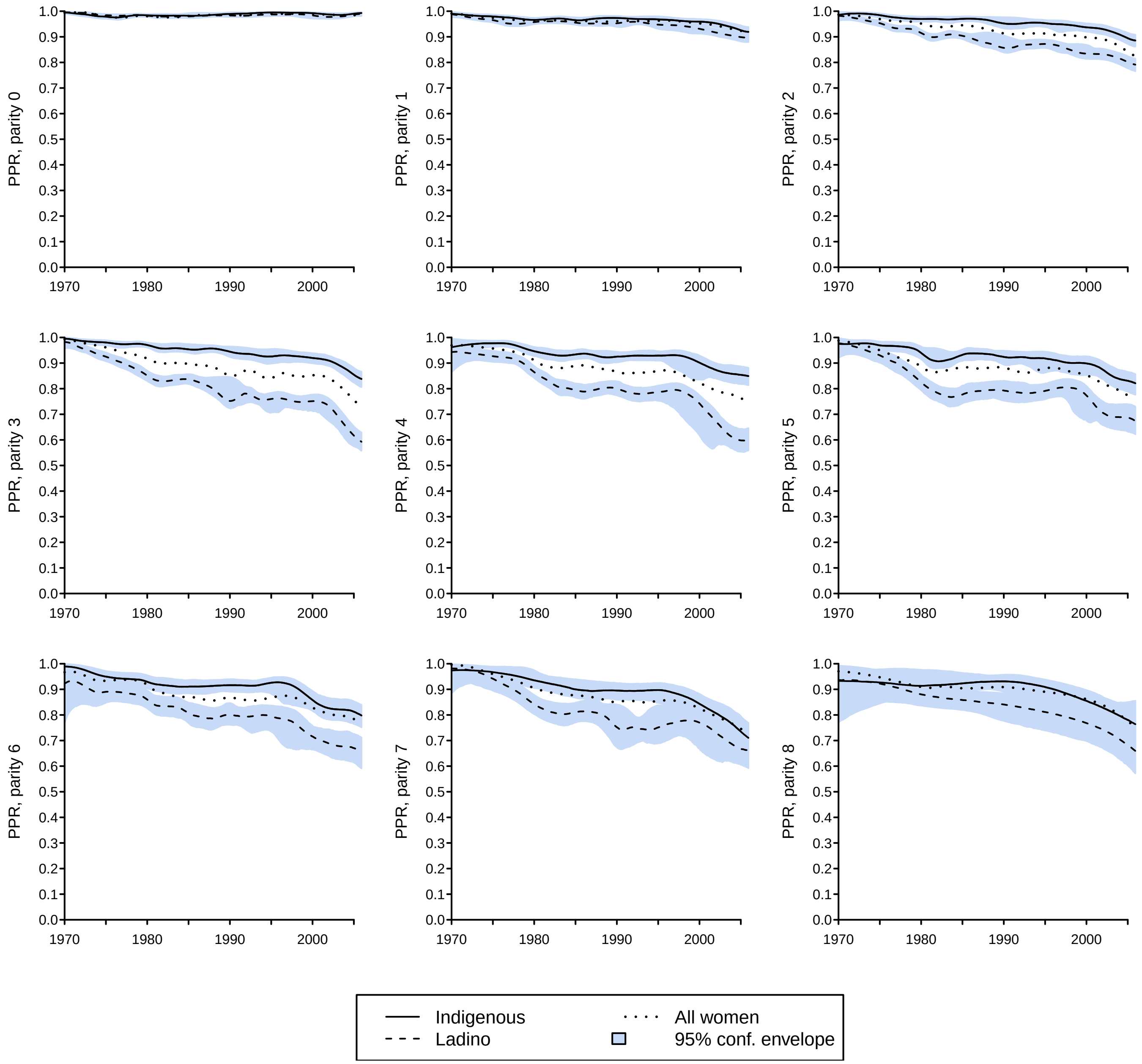


Figure 4: Period Parity Progression Ratios, ethnicity

Note: 95% confidence envelopes are based on a weighted pairs bootstrap as described in the text.

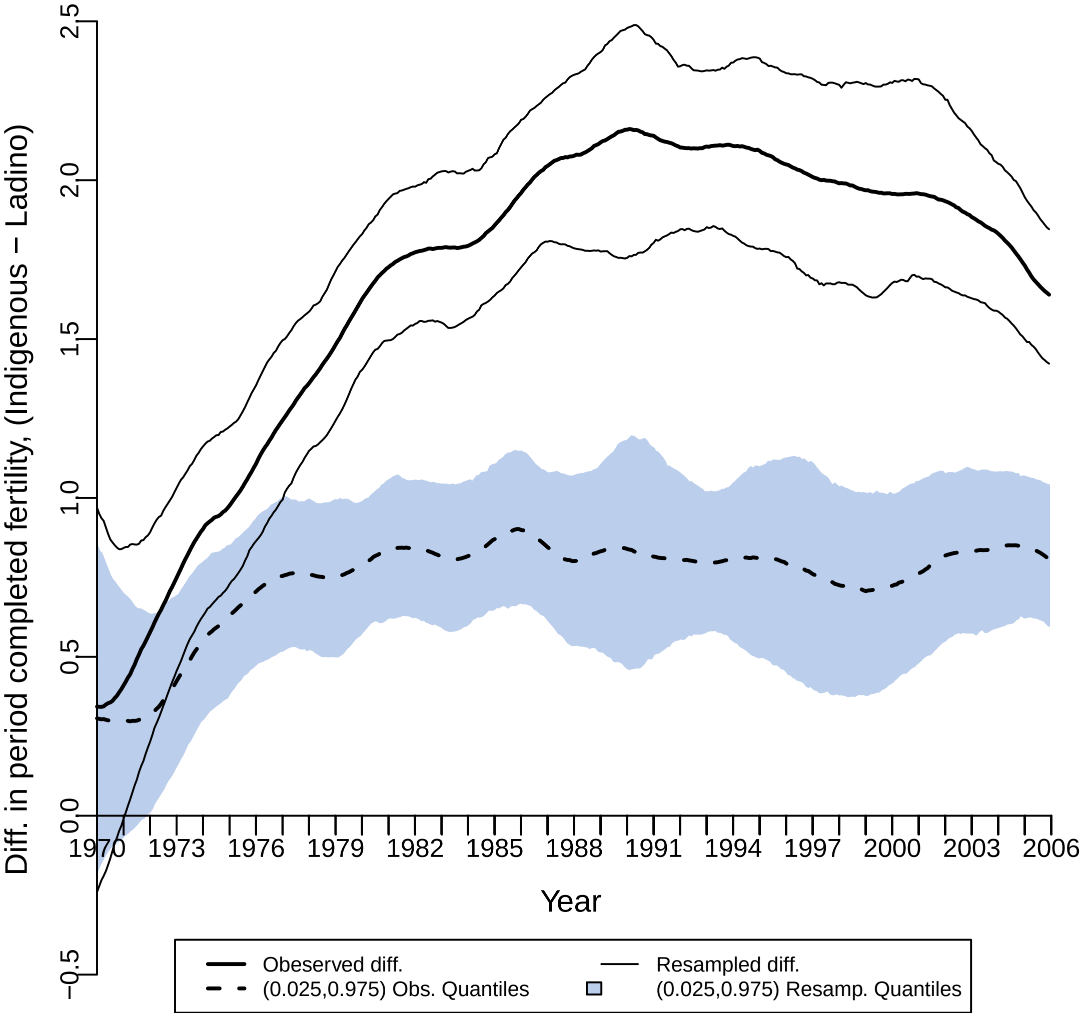


Figure 5: Resampling test of $H_o : F_t(indigenous) = F_t(Ladino)$

Note: See main text for a description of the resampling test.

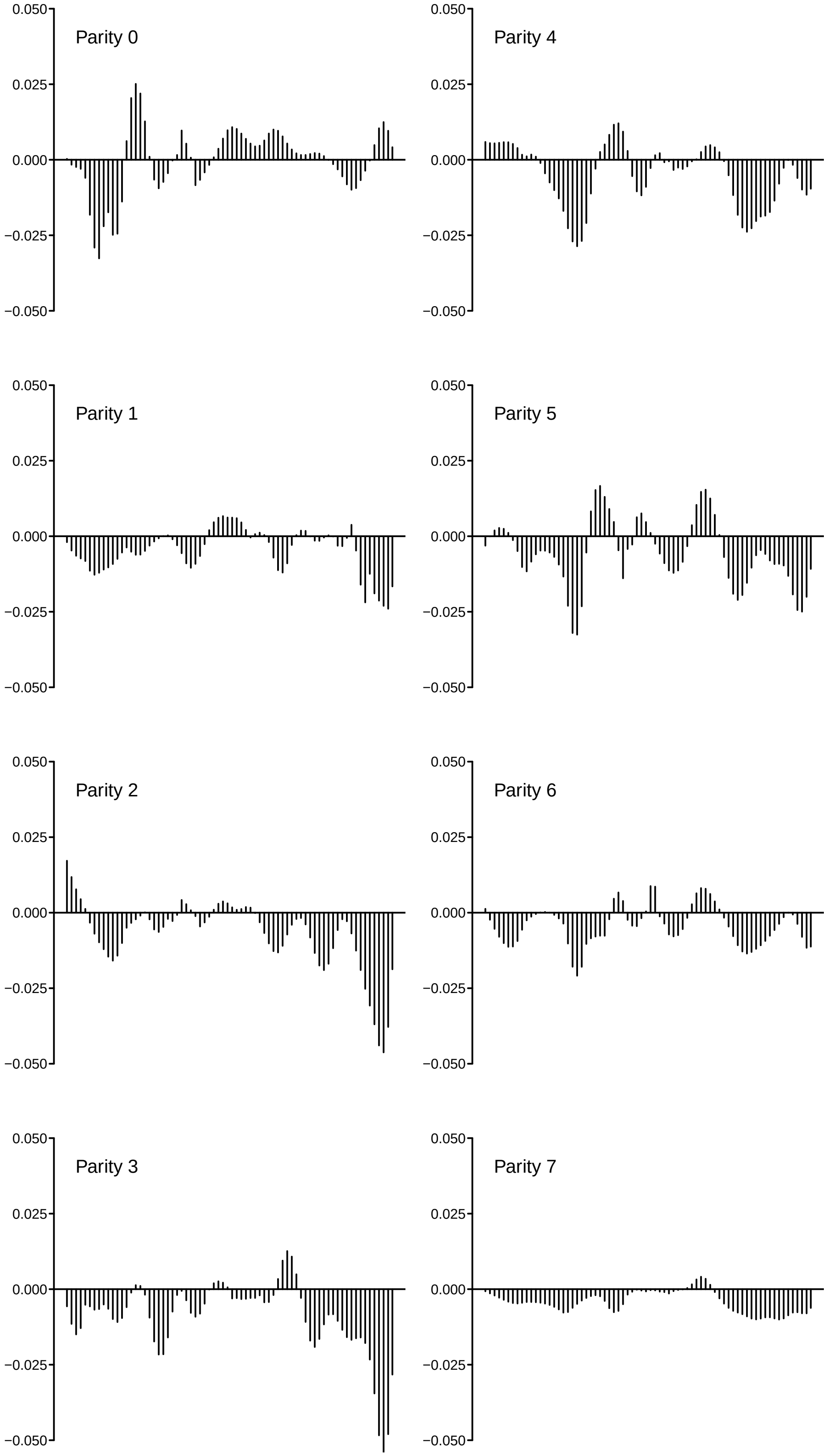


Figure 6: Contributions to ΔF_t, 6 month intervals, indigenous women