Estimation of Life Tables in the Latin American Data Base (LAMBdA): Adjustments for Relative Completeness and Age Misreporting

Alberto Palloni Guido Pinto[y] Hiram Beltran-Sanchez[z]

October 27, 2016

1 Background

The methodologies described in this paper belong to a small subset of a broader set of methods developed to produce adjusted estimates of adult mortality for countries in the Latin American and Caribbean (LAC) region covering 150-160 years, from 1850 to 2010. This period encompasses approximately the end of colonial rule, the aftermath of wars of independence from Spanish and Portuguese domination, the establishment of nation states, integration into a world system and the world economy, and all developments that unfolded following World War II.[1] In this paper we focus only on adjustments of life tables for the post-1950 period. To do so we avail ourselves of

place at a particular time as they exclude events that, for a number of reasons, are never recorded. Since population censuses too are normally a ected by coverage problems, mortality rates computed with the raw data may contain smaller *net errors* that would be expected otherwise. In general, however, the observed mortality rates underestimate mortality levels, particularly at very young and old ages. We use the term *relative completeness* when we speak of ratios of observed to true mortality rates.

Table 1 displays estimates of relative completeness of adult (over 5 years of age) and, for comparison, those corresponding to infant (age 0) and early child (ages 1-4) death registration in a sample of LAC countries over two di erent periods of time. The gures in this table con rm that the quality of the information is poorer at very young ages and that, although there is a clear universal trend toward improvement, an important fraction of countries still show signs of de cient registration even quite recently.

Imperfect relative completeness of death registration is not the only problem a ecting estimates of mortality. An important domain of errors involves age misreporting and the most insidious manifestation is systematic over (under) reporting. Vital and census statistics in LAC countries are, almost without exception, a ected by age overstatement, particularly at ages over 40 or 45 (see below). When the (true) age distribution of a population is roughly exponential in nature |as it always is in stable and quasi stable populations|systematic age overstatement of populations induces downward biases in mortality rates at older ages. These biases are not o set when there is an equal propensity to overstate ages at death. The reason these two type of errors do not cancel each other out is that while both adult mortality rates and adult population age distributions are roughly exponential, one slopes upwards (mortality rates) whereas the other slopes downwards (population). Matters are made worse when, as is almost always the case, the rate of decrease of population with age (natural rate of increase in a stable population) is several times lower than the rate of increase of adult mortality rates (rate of senescence in Gompertz mortality regimes). The consequence is that unless the propensity to overestimate ages at death is much higher than the propensity to overestimate ages of population, observed mortality rates will contain downward biases. If left uncorrected, the resulting life tables will o er a misleading portrayal of the curvature of mortality at older ages, suggesting the existence of slower rates of senescence or heavy in uence of selection due to changing frailty composition. As the quality of vital registration and census enumeration improves, the magnitude of these biases tends to decrease and the entire history of observed life tables will erroneously suggest trends in old age patterns of mortality and even relative acceleration of the rates of mortality decline at older ages.

Unlike problems created by age heaping, distortions caused by systematic age misstatement cannot be repaired by restoring the original age distribution standard using computations that rely on safe assumptions. Systematic age misstatement is altogether di erent since it is harder to diagnose and, as we show below, its treatment requires additional knowledge of two functions: (a) the conditional (on age and gender) propensity of individuals to exaggerate (decrease) the true age and (b) the conditional (on age and gender) distribution of the di erence between the correct and declared age. To solve the problem we propose generalizations of an existing procedure to identify the presence of age misstatement, formulate a new method to estimate functions describing (a) and (b) from observables, and de ne an algorithm that adjusts observed adult mortality rates for both faulty coverage and systematic age misreporting.

33(able)relang Tabl62-13s5ectionsfuity rateervadteervadtbles7i16(rateel)28(5eel)28-261(o0-282(slople)-417vadtbf17vadtn)

Table 1: Relative completeness of deaths registration in the LAC countries: 1920-2010.

Country	Period 1900-1949				Period 1950+	
	Mid-Year	Age 0	Age 1-4	Age 5+	Mid-Year	Age 5+
Argentina	1914	0.968	0.865	0.939	1953	0.974
					2005	0.995
Brazil					1985	0.885
					2005	0.996
Chile	1925	0.867	0.829	0.852	1956	0.961
	1945	0.867	0.829	0.934	2006	0.980
Colombia	1944	0.821	0.815	0.749	1957	0.790
					2008	0.800
Costa Rica	1927	0.901	0.922	0.893	1956	0.918
	1938	0.901	0.922	0.893	2005	0.975
Cuba	1925	0.806	0.893	0.800	1961	0.890
	1948	0.806	0.893	0.870	2006	0.989
Dominican Republic	1942	0.476	0.451	0.487	1955	0.500
					2006	0.604
Ecuador					1956	0.738
					2005	0.805
El Salvador	1940	0.554	0.776	0.721	1955	0.700
					2008	0.714
Guatemala	1945	0.714	0.898	0.784	1957	0.888
					2005	0.940
Honduras	1942	0.542	0.551	0.495	1955	0.518
	1947	0.542	0.551	0.500	1989	0.750
Mexico	1925	0.843	0.822	0.752	1955	0.860
	1945	0.843	0.822	0.883	2005	0.959
Nicaragua	1945	0.526	0.545	0.498	1956	0.456
					2007	0.561
Panama	1945	0.837	0.757	0.829	1955	0.839
					2005	0.853
Paraguay					1956	0.601
					2006	0.681
Peru					1950	0.490
					2008	0.533
Uruguay	1908	0.844	0.822	0.879	1969	0.960
					2007	0.996
Venezuela	1938	0.833	0.857	0.846	1955	0.866
	1945	0.833	0.857	0.855	2006	0.895

Table 2: Biases due to age overstatement.

Country	Mid-Year	Unadjusted		Adjusted*	
		E(45)	E(60)	E(45)	E(60)
Argentina	1953	25.96	15.39	25.29	14.55
	2005	30.02	17.96	29.33	17.15
Brazil	1985	28.55	17.61	27.62	16.51
	2005	31.27	19.77	30.23	18.58
Chile	1956	24.44	14.57	23.72	13.64
	2006	33.20	20.45	32.16	19.33
Colombia	1957	27.34	16.68	26.46	15.67
	2008	35.09	22.29	33.86	20.96
Costa Rica	1956	29.08	17.55	28.10	16.46
	2005	34.96	22.40	33.78	21.13
Cuba	1961	30.13	18.15	29.18	17.08
	2006	33.46	20.94	32.56	19.95
Dominican Republic	1955	33.62	22.44	31.91	20.52
	2006	38.35	25.76	36.41	23.68
Ecuador	1956	28.75	17.98	27.77	16.83
	2005	37.42	25.23	35.94	23.62
El Salvador	1955	27.64	17.54	26.69	16.42
	2008	32.79	21.74	31.85	20.62
Guatemala	1957	24.44	15.06	23.68	14.07
	2005	31.39	20.22	30.42	19.10
Honduras	1955	30.55	20.37	29.14	18.64
	1989	37.33	25.06	35.61	23.17
Mexico	1955	26.57	16.69	25.80	15.71
	2005	33.04	21.13	31.97	19.95
Nicaragua	1956	32.09	21.05	30.61	19.37
	2007	36.23	24.05	34.71	22.41
Panama	1955	28.93	17.67	27.87	16.45
	2005	35.92	23.18	34.65	21.81
Paraguay	1956	32.97	20.81	31.73	19.44
	2006	34.84	22.17	33.60	20.84
Peru	1950	30.61	20.64	29.47	19.25
	2008	39.37	26.32	37.66	24.52
Uruguay	1969	26.72	15.47	26.11	14.69
	2007	30.35	18.17	29.85	17.57
Venezuela	1955	27.49	16.81	26.47	15.64
	2006	32.75	20.94	31.53	19.59

* Adjusted for age misreporting

adjustments procedure to deal with it are well-known. Much less is known about the nature and impact of age misreporting. In the section below we propose a methodology to identify the presence of these errors and to correct them.

3 Systematic age misreporting

3.1 Setup

We begin with a few basic de nitions. Let $^{o}_{x}$ be the average conditional probability that individuals aged x overstate their age in a census and $^{u}_{x}$ the conditional probability of understating their age. Then (1 $^{o}_{x}$ $^{u}_{x}$) is the probability of an accurate age statement. Individuals who over(under)

3.2 Observed patterns of age misreporting

What do we know about age misreporting in population and death counts in LAC and in other countries? There is an extensive literature on general errors in age reporting (Ewbank, 1981; Chidambaram and Sathar, 1984; Kamps E., 1976; Nuñez, 1984) as well as on systematic age misstatement, mostly adult age overstatement, in population counts. And while a fair number of these studies uncover evidence of overstatement in low income countries (Mazess and Forman, 1979; Grushka, 1996; Bhat, 1987, 1990; Del Popolo, 2000; Dechter and Preston, 1991) or in US migrant (Hispanic or Hispanic origins) groups (Rosenwaike and Preston, 1984; Spencer, 1984), there is a body of literature that identifies patterns of age overstatemet in high income countries as well (Horiuchi and Coale, 1985; Coale and Kisker, 1986; Condran et al., 1991; Preston et al., 2003; Elo and Preston, 1994). In the US, for example, age overstatement is one of the factors that could explain the so called Black-White mortality crossover, whereby African American mortality rates dip below those of their White counterparts at very old ages (over 70). And while the recurrent idea of heavy selection due to frailty has not been completely discarded, the most recent investigations suggest that overstatement of ages in the population (and also deaths) among African American more so than among Whites accounts for a substantial part of the mortality crossover (Elo and Preston, 1994). The Black-White mortality crossover is just an extreme example of the damage that age misreporting can inflict on estimates of adult mortality. As others before us have done (Dechter and Preston, 1991; Grushka, 1996; Bhat, 1987, 1990), we will show that age overstatement is also an important source of error in LAC countries.

Partial information on the matrix has been obtained mostly from studies involving record

binary variable set to 1 when there is over (under) statement and zero otherwise. Initially the model specifies a vector of covariates including age, age squared, urban/rural residence, gender, and education. The sample includes individuals aged 50 and over since at younger ages there are only traces of systematic age misstatement (mostly in the form of heaping). Because gender and age are the only covariates that can be used at a national level, we simplify the model to include only these two traits as predictors. Finally, after verifying that the effects of age squared and gender were statistically insignificant, the final model conditions only on `true' age of individuals. Table 3 displays estimated parameters for over and under stating ages using the weighted sample.

ii *Estimation of conditional probabilities of over(under) stating ages by* $1 < n \quad 10$ *years,* $\pi^{o}_{x}(j)$ *and* $\pi^{u}_{x}(j)$: We estimate a multinomial model with 9 categories that includes gender and (true) continuous age as independent variable. The resulting estimates reveal that the effects of gender are always statistically insignificant, that those of age show no clear pattern and, in addition, that their magnitude is quite small in 6 out of 8 cases for overstatement models and in 5 out of 8 contrasts for age understatement. To simplify we estimate a null model predicting

Table 3: Estimated parameters of best logistic models for age misreporting.

Variable	Overreporting Coe (se)	Underreporting Coe (se)
True age[1]	0.014(.0036)	0.002(.0040)
Constant	-2.127(.271)	-1.846(.297)
N	6290	6290

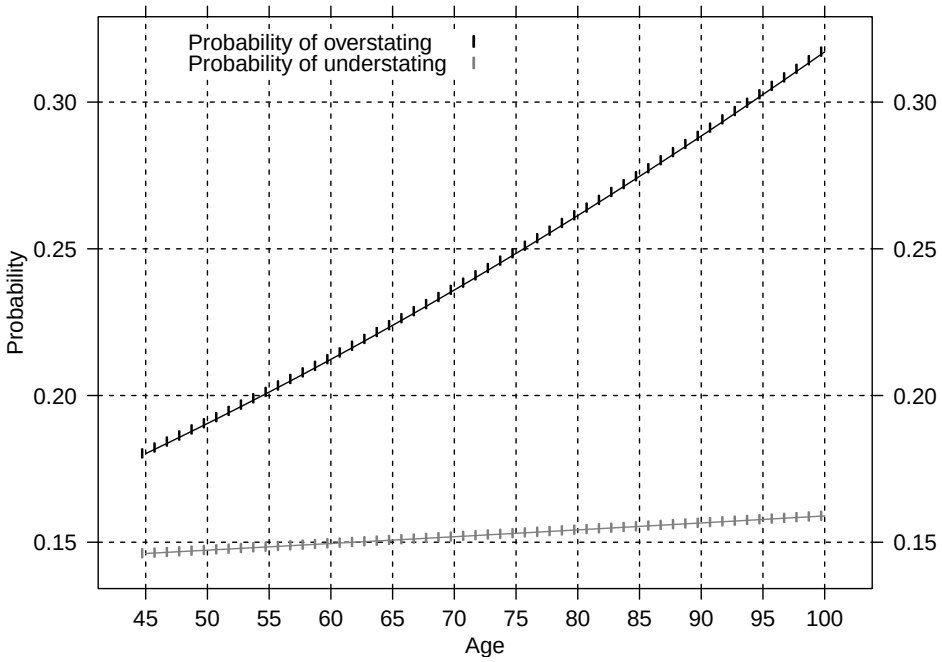
[1] Regressions estimated using sampling weights. Sample includes population with true age 60 and older and excludes ambiguous cases and foreign citizens.

Table 4: Average (conditional) probabilities of overreporting ages.

	Probability [1]	
n	Overstating	Understating
1	0.621	0.510
2	0.191	0.128
3	0.079	0.091
4	0.040	0.052
5	0.023	0.041
6	0.015	0.035
7	0.009	0.028
8	0.007	0.026
9	0.005	0.013
10+	0.009	0.060

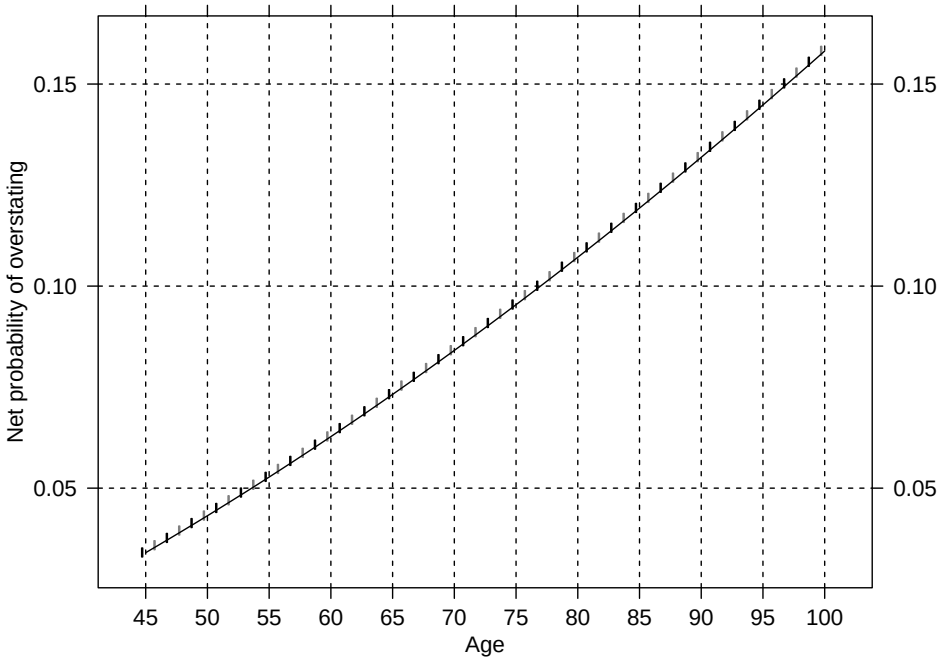
[1]

Figure 1: Predicted probabilities of over(under) stating ages.



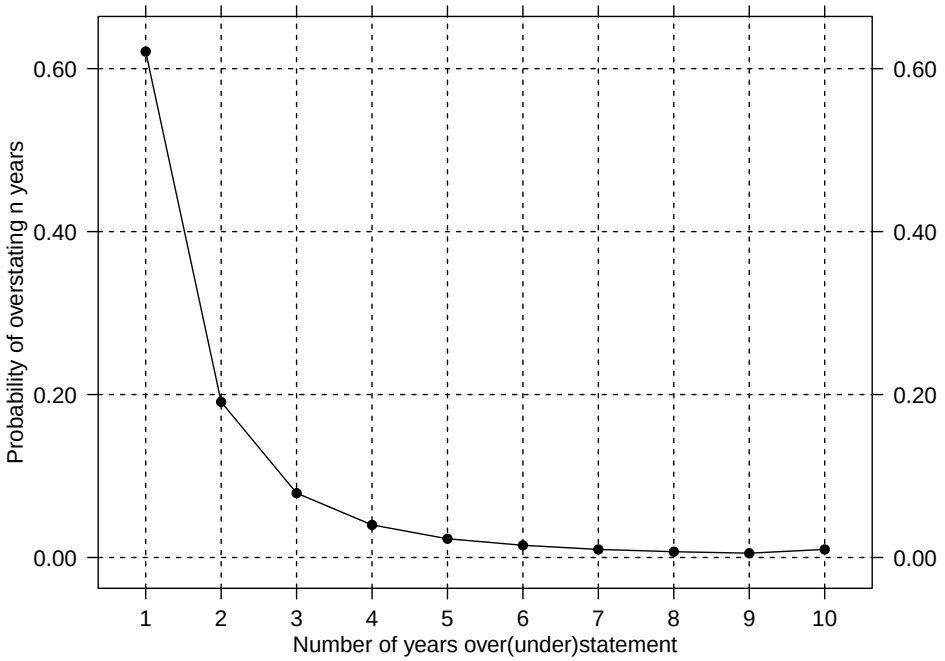
Source: Costa Rica Special study of 2000 population census.

Figure 2: Predicted probabilities of net overstating ages.



Source: Costa Rica Special study of 2000 population census.

Figure 3: Conditional probabilities of overstating age by *n* years.



Source: Costa Rica Special study of 2000 population census.

based on record linkages show that there is age misreporting of ages at death as well, albeit of lower magnitude than that found in population counts, and that it also tends to be in the direction of overstatement (Rosenwaike and Preston, 1984). This is con rmed by the application of indirect techniques designed to detect age at death overstatement in a number of low and high income countries (see below). It follows that expressions analogous to (3.1) and (3.2) must be applicable for death counts as well. To make the problem tractable one needs an empirical approximation to a matrix analogous to but now specialized to ages at death. To our knowledge no such matrix has ever been estimated in LAC or anywhere else and we are unaware of any national data that could be used for such purpose. In what follows we assume that the *standard age pattern of age misstatement of death counts is identical to that of age misstatement of population counts*, although its level may be di erent. This assumption enables us to de ne the nal model of age misreporting as a set of two equations with two unknown parameters:

$$ {}^{o} = {}^{no} \wedge S \quad {}^{T} \tag{3.4} $$

$$ {}^{o} = {}^{no} \wedge S \quad {}^{T} \tag{3.5} $$

where ^{T} and ^{O} are the true and observed distributions of death counts and ^{no} is the magnitude of net overstatement of ages at death relative to the standard pattern. In closed populations equations (3.4) and (3.5) are naturally (see below) related and it is unlikely that there is always a unique solutions for ^{no} and ^{no} unless we either x the value of one of them or, alternatively, retrieve solely their ratio. A brief proof of lack of identi cation is in Appendix B and solutions for empirical estimation are in section 4.2.

4 Identification and correction of errors due to systematic age misreporting

In this section we propose a methodology to identify and then adjust mortality statistics for age misreporting. The methodology is only applicable when age misreporting is produced following the model outlined in the previous section.

4.1 Identification of systematic age misreporting

A key component of our analysis is the detection and identification of patterns of age misstatement in the population and death counts. As shown in a previous section, the distortions associated with age misreporting in population and death counts is more complex than those involving only faulty completeness. Detection of the problem is difficult since its manifestations are quite subtle and, in the absence of overt and striking phenomena such as the US Black-White cross over, is likely to remain concealed and undetected. There are two well-tested methods to identify the existence of age over(under) statement in either population or death counts. The first method requires an external data source with correct dates of birth or ages in a population at a particular time that can be compared to age-specific census counts at approximately the same time. An example of this is the utilization of Medicare data in the US, a source of information that, as a rule, contains both population exposed and mortality data. Because Medicare data are linked to Social Security records and these are known to register age with high precision, mortality rates computed from Medicare data are a gold standard against which conventional mortality rates could be contrasted and their quality evaluated (Elo et al, 2004). If one ignores the existence of a population not covered by Medicare records, it is also feasible to link individual census records to Medicare records and investigate more precisely the nature of patterns of age misreporting in census counts. If, in addition, Medicare records are linked to the US National Death Index (NDI) it is then possible to repeat the same operations and assess the quality of reporting of age at deaths. In all cases one must assume that the coverage of population in both sources is complete or, if incomplete, identical[4]. Record linkage from multiple sources such as those illustrated above has rarely been used as it is expensive and involves resolution of complicated confidentiality issues.

A second method is less data demanding, considerably less expensive and is simple to apply but can only *reveal* the existence of age misreporting in one of the two sources and provides few clues about its nature. The procedure was proposed by Preston and colleagues (Rosenwaike and Preston, 1984; Elo and Preston, 1994; Bhat, 1990; Grushka, 1996) and has been applied in countries of North America, Western Europe and in Latin America (Condran et al., 1991; Grushka, 1996; Dechter and Preston, 1991; Palloni and Pinto, 2004; Del Popolo, 2000). In a nutshell the method consists of comparing cumulative population counts in a census in year t_1 to the expected cumulative population counts in a second population census in year t_2. The computation of expected quantities requires both an initial census opening the intercensal interval, a second census counts at time t_2 closing the intercensal interval, and age specific deaths counts in the intercensal period spanning an interval of $k = (t_2 - t_1 + 1)$ years. The ratio of observed to expected population is an indicator of age misstatement:

$$cmR^o_{x;[t_1,t_2]} = \frac{cmP^o_{x+k;t_2}/cmP^o_{x;t_1}}{1 - (cmD^o_{x;[t_1;t_2]}/cmP^o_{x;t_1})} \tag{4.1}$$

[4] The assumption is more restrictive than we made it sound: if population coverage is not complete in either source, then the subpopulations missed in each census must be random relative to their true and reported age.

where $cmP^{o}_{x;t_1}$ and $cmP^{o}_{x;t_2}$ are cumulative populations over ages x and $x+k$ in the first and second census, respectively, and $cmD^{o}_{x;[t_1;t_2]}$ is the cumulative deaths after age x during the intercensal period. This expression is a simple contrast between two different estimates of the same underlying quantity (population parameter), namely, the cumulative survival ratio: the denominator uses the complement of the observed ratio of (cumulative) intercensal deaths to (cumulative) population in the first census, whereas the numerator expresses it as the survival ratio computed from the cumulative counts in two successive population censuses. It is useful to express (4.1) in a logarithmic form, namely,

$$\ln(cmR_{x;[t_1;t_2]}) = \ln(SN^{o}_{x;x+k}) \quad \ln(SD^{o}_{x;x+k}) \tag{4.2}$$

where $SN^{o}_{x;x+k}$ is the `survival ratio' computed from two censuses and $SD^{o}_{x;x+k}$ is the survival ratio computed from intercensal deaths[5]. In the absence of migration, age misstatement and imperfect completeness of census and death counts, both estimators should yield the same number, the ratio in (4.1) should be 1, and the log expression in (4.2) should be 0 for all adult ages.

To shed light on the meaning of expressions (4.1) or (4.2) and to simplify notation and terminology we will speak of *net age misreporting* to refer to the net result of both age over and under statement. Furthermore, because we, as well as past research, uncover systematic net age overstatement of adult ages in LAC countries, we will speak of `age overstatement' or `age overreporting' even though we refer to the net result of age under and over reporting. In Appendix C we show that when the assumption of absence of age misreporting is violated, we can approximate (4.2) as

$$\ln(cmR_{x;[t_1;t_2]}) \quad \ln\left[\frac{h(x+k)}{h(x)} \quad \frac{g(x)}{h(x)}\right] \quad 1 \quad 1 + I^{T}_{x;x+k} \tag{4.3}$$

where $I^{T}_{x;x+k}$ is a true integrated hazard analogue between ages x and $x+k$ (and hence strictly positive), $h(x)$ is an increasing function of age that depends on age overstatement of populations and $g(x)$ is an increasing function of age that depends only on overstatement of ages at death. Both $h(x)$ and $g(x)$ are functions of the propensity to overstate and the underlying population and deaths age distribution. Assume now that the propensity to overstate ages (of populations or deaths) is age invariant or increases with age and that the following three conditions hold: (a) the (true) age distribution slopes sharply downward, (b) the age distribution of deaths increases with age, and (c) the rate of decrease of population with age is smaller that the rate of increase of deaths with age. Under these three conditions, almost universally verified in all human populations, the ratio $h(x+k)/h(x)$ will always be larger than 1 and will increase with age, $g(x)$ will always be larger than 1 and increase with age, and the rate of increase in $g(x)$ will exceed the rate of increase in $h(x)$ so that $g(x) > h(x)$ almost everywhere in the age span. The following are possible scenarios[6]

1. When there is systematic age overstatement of population counts ONLY, $h(x) > 1$ and $g(x) = 1$, then expression (4.3) reduces to

$$\ln(cmR_{x;[t_1;t_2]}) = \ln\left[\frac{h(x+k)}{h(x)}\right] + (h$$

1 0 0 1 49 10.9091 Tf 4.243 0 Td [(n)]TJ/3308 3.958 Td [ai3 0 Tdo3-307(folloo)hk

The inequality results because the positive term in the expression, that is, the distortion of the survival ratio based on population counts, will be smaller than the negative term in uenced by the distortion in the second estimator based on intercensal death rates.

2. When there is systematic age overstatement of death counts ONLY, $h(x) = 1$ and $g(x) > 1$, the expression becomes

$$\ln(cmR_{x;[t_1;t_2]}) = \ln \quad \frac{h(x+k)}{h(x)} \quad + (g(x) \quad 1)(1 + I^{T}_{x;x+k}) > 0$$

and the positive sign results from the fact that all terms in the expression are positive.

3. When there is systematic overstatement of BOTH population and death counts, $g(x) > h(x) > 1$, then

$$\ln(cmR_{x;[t_1;t_2]}) = \ln \quad \frac{h(x+k)}{h(x)} \quad + \quad \frac{g(x)}{h(x)} \quad 1 \quad (1 + I^{T}_{x;x+k}) > 0$$

because, by assumption, all terms are positive.

Before we can use the above to diagnose conditions in an empirical case, two issues must be resolved. First, it is possible that there are empirical patterns of age overstatement of deaths and populations that o set each other and produce ratios close to 1 even though the underlying data are incorrect. That is, scenario (3) is such that the log of the ratio is 0 at all ages even when there is net age overstatement. Because of this possibility, a diagnostic of observed conditions based on the index (or the log of the index) can only detect consistency (including error consistency) of age declaration in population and death counts, rather than suggest accuracy (Dechter and Preston, 1991). Second, throughout we assumed that both census and death counts had perfect coverage. When one allows for defective census coverage, an identi cation problem is created since now we will have

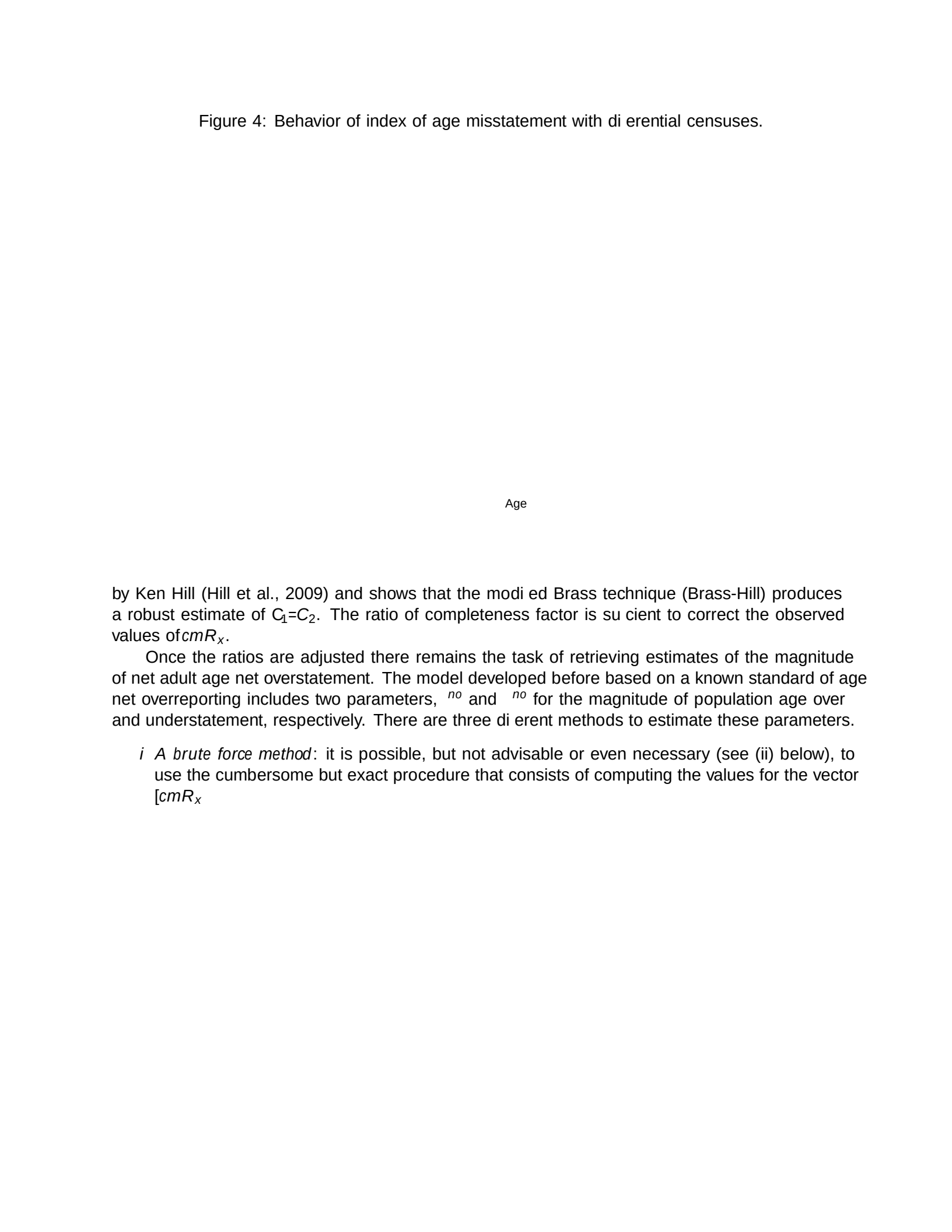
$$\ln(cmR_{x;[t_1;t_2]}) \quad \ln \quad \frac{C_2}{C_1} \quad + \ln \quad \frac{f(x+k)}{f(x)} \qquad \frac{C_3 \quad g(x)}{C_1 \quad h(x)} \quad 1 \quad (1 + I^{T}_{x;x+k}) \qquad (4.4)$$

and it is clear that we can no longer separate the role of age overstatement and completeness. In particular, even if there is no age misreporting, expression (4.4) can yield non-zero values and mimic increasing or decreasing patterns with age that result naturally from age overstatement alone. To understand better the combined in uence of defective coverage and age misreporting on observed mortality rates we need to de ne more precisely the nature of the functions $h(x)$ and $g(x)$, the nature of their dependence on patterns of age misreporting and how they interact with defective coverage. We investigate this issue in the section below.

4.2 Correction of errors due to age misreporting

As indicated before, the main tool to detect adult age misreporting is highly sensitive to relative completeness of census counts. Figure 4 displays the value of cmR_x that one obtains when there is no age misreporting at all but there is di erential completeness in census counts. Thus, one cannot learn much about patterns of age misreporting unless population census counts are rst adjusted. This requires to identify methods that provide robust estimates of completeness of one census relative to the other. As we show below, the evaluation study con rms a result rst noted

Figure 4: Behavior of index of age misstatement with di erential censuses.



by Ken Hill (Hill et al., 2009) and shows that the modi ed Brass technique (Brass-Hill) produces a robust estimate of C_1=C_2. The ratio of completeness factor is su cient to correct the observed values of cmR_x.

Once the ratios are adjusted there remains the task of retrieving estimates of the magnitude of net adult age net overstatement. The model developed before based on a known standard of age net overreporting includes two parameters, ^{no} and ^{no} for the magnitude of population age over and understatement, respectively. There are three di erent methods to estimate these parameters.

i *A brute force method*: it is possible, but not advisable or even necessary (see (ii) below), to use the cumbersome but exact procedure that consists of computing the values for the vector [cmR_x

Table 5: Regression model relating index of age misstatement and parameters of age misreporting.

Age	$_0$	$_1$	$_2$	R^2
45	1.000	-0.027	-0.004	1.000
46	1.000	-0.012	-0.005	1.000
47	1.000	-0.006	-0.005	1.000
48	1.000	-0.003	-0.006	1.000
49	1.000	0.000	-0.007	1.000
50	1.000	0.002	-0.008	1.000
51	1.000	0.003	-0.009	1.000
52	1.000	0.005	-0.010	1.000
53	1.000	0.006	-0.011	1.000
54	1.000	0.008	-0.013	1.000
55	1.000	0.010	-0.014	1.000
56	1.000	0.012	-0.016	0.999
57	0.999	0.014	-0.019	0.999
58	0.999	0.017	-0.022	0.999
59	0.999	0.020	-0.025	0.999
60	0.999	0.024	-0.030	0.999
61	0.999	0.029	-0.035	0.999
62	0.999	0.035	-0.041	0.999
63	0.998	0.042	-0.048	0.999
64	0.998	0.051	-0.057	0.998
65	0.997	0.062	-0.069	0.998
66	0.996	0.076	-0.082	0.998
67	0.995	0.094	-0.099	0.997
68	0.994	0.116	-0.121	0.997
69	0.992	0.145	-0.148	0.996
70	0.990	0.183	-0.183	0.995
71	0.986	0.231	-0.228	0.995
72	0.982	0.295	-0.285	0.994
73	0.975	0.378	-0.360	0.992
74	0.966	0.490	-0.458	0.991
75	0.952	0.638	-0.586	0.989

Table 6: Results from inverse method of age misstatement to recover parameters of age misreporting.

run	no	$\hat{}no$	no	$\hat{}no$	R^2
1	0.000	0.061	0.350	0.370	1.000
2	0.000	0.002	0.700	0.685	1.000
3	0.000	-0.059	1.050	0.999	1.000
4	0.000	-0.118	1.400	1.313	1.000
5	0.000	-0.178	1.750	1.628	1.000
6	0.000	-0.238	2.100	1.942	1.000
7	0.000	-0.298	2.450	2.256	1.000
8	0.000	-0.358	2.800	2.571	1.000
9	0.350	0.393	0.700	0.727	1.000
10	0.350	0.392	1.050	1.078	1.000
11	0.350	0.391	1.400	1.429	1.000
12	0.350	0.390	1.750	1.780	1.000
13	0.350	0.388	2.100	2.130	1.000
14	0.350	0.387	2.450	2.481	1.000
15	0.350	0.386	2.800	2.832	1.000
16	0.700	0.710	1.050	1.067	1.000
17	0.700	0.755	1.400	1.445	1.000
18	0.700	0.801	1.750	1.823	1.000
19	0.700	0.846	2.100	2.201	1.000
20	0.700	0.892	2.450	2.579	1.000
21	0.700	0.938	2.800	2.957	1.000
22	1.050	1.013	1.400	1.393	1.000
23	1.050	1.096	1.750	1.791	1.000
24	1.050	1.179	2.100	2.189	1.000
25	1.050	1.262	2.450	2.587	1.000
26	1.050	1.345	2.800	2.985	1.000
27	1.400	1.303	1.750	1.704	1.000
28	1.400	1.416	2.100	2.117	1.000
29	1.400	1.530	2.450	2.530	1.000
30	1.400	1.643	2.800	2.943	1.000
31	1.750	1.582	2.100	2.004	0.999
32	1.750	1.720	2.450	2.427	1.000
33	1.750	1.859	2.800	2.851	1.000
34	2.100	1.851	2.450	2.292	0.999
35	2.100	2.009	2.800	2.723	1.000
36	2.450	2.110	2.800	2.569	0.998

Table 7: Non-linear regression to recover parameters of age misreporting.

run	no	$\hat{}no$	no	$\hat{}no$	R^2

choose an optimal adjustment strategy we develop an evaluation study designed to identify best adjustments for relative completeness and age misreporting. The goal of the study is to generate distributions of errors associated with each adjustment procedure under a diverse set of conditions

E_0 and, additionally, that each type of demographic transition pro le preserves the age patterns of mortality and fertility. We chose the West model in the Coale-Demeny family of life tables and an age pattern of fertility identical to the one used in the computations of the Coale-Demeny stable population models (Coale et al., 1983). Information on the four classes of demographic transitions used here are in Appendix A. Finally, we construct a fth pro le of a stable population with natural rate of increase and fertility pattern equivalent to the average of LAC populations in the interval 1950-60, e.g. not yet heavily perturbed by large scale net migration as is the case in Argentina, Brazil, Cuba, and Uruguay, or early fertility changes, as in Argentina and Uruguay.

Following routine population projection calculations we produce 505 populations and associated distributions of births and deaths by single calendar year and single years of age. The simulated populations represent a very broad set of experiences, from those preserving population stability up until 1950 or thereabouts, to those shifting to quasi-stability from 1930 up to 1980, to those with little or no stability at all from the start [13].

5.2 Simulated distortions I: imperfect relative completeness of death registration

Distortions due to population or death coverage can be implemented in a straightforward matter. We de ne observed population (or death) counts by age as a fraction of the simulated (true) quantities:

$$\begin{aligned} P^o_{xt_1} &= C_1 P^s_{xt_1} \\ P^o_{xt_2} &= C_2 P^s_{xt_2};\ t_2 < t_1 \\ D^o_{xt} &= C_3 D^s_{xt};\ t = t_1; t_1 + 1; :::\quad t_2 \end{aligned}$$

for x 5, where $P^o_{xt_1}$ is the observed (distorted) population at age $(x; x+1]$ at time t_1, $P^o_{xt_2}$ is the observed (distorted) population at age $(x; x+1]$ at time t_2; and D^o_{xt} is the observed (distorted) number of deaths in year t; $P^s_{xt_1}; P^s_{xt_2}$ and D^s_{xt} are the simulated (true) quantities and $C_1; C_2$ and C_3 are the fractions of total events actually observed (completeness factors). The completeness factors for censuses were set at values in the range 0.80-1.0 in intervals of 0.5 whereas the death completeness factors varied between 0.70 and 1.0 in intervals of 0.5. Altogether we produce a total of 875 (175*5) patterns of including distorted and true demographic pro les. These de nitions are su cient to evaluate adjustment methods that require only one census and one to three years of death counts centered on the census or, alternatively, those that demand as inputs two population censuses and an array of intercensal deaths.

The above set up contains a massive assumption, namely, that completeness of both population and death counts is age invariant. At least within the age range in which the techniques are deployed (5-85), the assumption is unlikely to be met, particularly for population counts. To complete the set of reasonable distortions we add two di erent patterns of age varying completeness generating a total of 2,625 simulated populations. We show later, however, that as long as the di erence

5.3 Simulated distortions III: combining age misreporting and faulty coverage

Table 8: Methods to adjust for completeness of death registration: assumptions and required data.

Method	Assumptions	Required Data
Brass (B)	1-2-3-4-5	B
Brass-Hill (BHill		

ages. This poses a conundrum: if, as asserted before, LAC population and mortality counts

of populations de ned by selected underlying conditions and, nally, isolating two types of errors that violate basic assumptions of all methods considered here, namely, age misreporting and age dependent completeness. In section 6.2 we describe the behavior of methods to adjust for age misreporting.

6.1 Defective completeness: evaluation using pooled simulated populations

subsets. Naturally, di erent error metrics yield di erent ranking of methods but the measure we use is the preferred one in most applications of this kind.[18]

The six panels of Tables 9{11 display the mean of the proportionate absolute error for each of the six populations subsets de ned above. Table 9 refers to simulations with constant relative completeness by age and Tables 10 and 11 re ect results using two di erent patterns of age varying relative relative completeness. The errors in each population subset, $s; s = 1; 2:::6$, are $\overset{d}{s} = \sum_{j=1}^{j=K_s} \overset{d}{sj}$ and $\overset{c}{s} = \sum_{j=1}^{j=K_s} \overset{c}{sj}$, where

Table 9: Proportionate absolute errors in each of six populations subsets with age invariant relative completeness.

	A. Stable & Nonstabe			B. Stable			C. Nonstable			D. Nonstable [?]			E. Nonstable			F. Nonstable [z]		
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill) [1]	0.003	0.003	0.003	0.005	0.005	0.004	0.003	0.003	0.002	0.001	0.001	0.001	0.002	0.003	0.002	0.002	0.003	0.002
Bennet Horiuchi No 1 (BH _1)	0.242	0.304	0.265	0.199	0.263	0.224	0.251	0.314	0.273	0.215	0.294	0.251	0.010	0.014	0.011	0.212	0.256	0.117
Bennet-Horiuchi No 2 (BH _2)	0.248	0.300	0.256	0.215	0.260	0.216	0.260	0.310	0.264	0.215	0.296	0.253	0.011	0.013	0.010	0.219	0.256	0.111
Bennet-Horicuhi No 3 (BH _3)	0.240	0.303	0.263	0.200	0.264	0.225	0.247	0.312	0.271	0.212	0.293	0.249	0.010	0.013	0.011	0.210	0.255	0.115
Bennet-Horiuchi No 4 (BH _4)	0.248	0.300	0.256	0.215	0.260	0.216	0.260	0.310	0.264	0.215	0.296	0.253	0.011	0.013	0.010	0.219	0.256	0.111
Bennet-Horiuchi No 5 (2SBH _4)	0.021	0.024	0.017	0.016	0.020	0.015	0.023	0.025	0.017	0.007	0.008	0.005	0.022	0.024	0.016	0.023	0.025	0.017
Brass-Martin(BMartin) [2]	0.079	0.107	0.085	0.038	0.038	0.021	0.110	0.124	0.086	0.057	0.071	0.061	0.112	0.124	0.084	0.111	0.124	0.085
Brass Hill (BHill) [1]	0.043	0.046	0.027	0.038	0.038	0.021	0.045	0.048	0.028	0.005	0.006	0.004	0.045	0.048	0.028	0.045	0.048	0.028
Preston Bennet (PB)	0.629	0.728	0.552	0.493	0.623	0.594	0.701	0.754	0.537	0.581	0.692	0.541	0.031	0.051	0.049	0.629	0.853	0.510
Preston Hill I (PH _1)	0.340	0.388	0.297	0.275	0.381	0.375	0.356	0.390	0.274	0.358	0.388	0.267	0.203	0.226	0.146	0.325	0.308	0.175
Preston-Hill 2 (PH _2)	0.367	0.386	0.272	0.249	0.367	0.320	0.374	0.391	0.258	0.377	0.390	0.251	0.242	0.258	0.146	0.348	0.315	0.181
Preston Lahiri No 1 (PL _1)	0.406	5.911	260.880	0.336	1.498	4.478	0.449	7.014	291.655	0.452	3.434	20.699	0.021	0.023	0.015	0.423	11.192	451.144
Preston-Lahiri No 2(PL _2)	0.378	5.560	168.422	0.307	1.366	4.558	0.415	6.609	188.274	0.414	2.064	6.947	0.022	0.027	0.021	0.394	0.916	3.422
N	31,500			6,300			25,200			700			4,320			10,368		

SD, standard deviation; Med, median.

[?] $_1 = \ _3 = 0$

$C_1 = C_2$ and $C_3 < 1$

[z] $C_1 \ 6 \ C_2$ and $C_3 < 1$ and maxabs($C_1 \quad C_2$) < :10

[1] Values of errors in the Brass-Hill shown in the rst row correspond to errors associated with the ratio $C_1 = C_2$. While values of Brass-Hill in the seventh row correspond to erros associated with relative completeness of death registration.

[2] BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

Table 10: Proportionate absolute errors in each of six populations subsets with age dependent relative completeness (Scenario 1).

	A. Stable and Nonstabe			B. Stable			C. Nonstable			D. Nonstable [?]			E. Nonstable			F. Nonstable [z]	
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill) [1]	0.040	0.046	0.033	0.039	0.045	0.032	0.040	0.047	0.033	0.039	0.045	0.032	0.013	0.015	0.010	0.038	0.037	0.020
Bennet Horiuchi No 1 (BH_1)	0.267	0.319	0.279	0.273	0.331	0.296	0.265	0.316	0.275	0.217	0.295	0.253	0.016	0.020	0.018	0.208	0.256	0.125
Bennet-Horiuchi No 2 (BH_2)	0.267	0.315	0.269	0.271	0.327	0.285	0.265	0.312	0.264	0.217	0.297	0.254	0.019	0.023	0.019	0.215	0.255	0.116

Table 11: Proportionate absolute errors in each of six populations subsets with age dependent relative completeness (Scenario 2).

	A. Stable and Nonstabe			B. Stable			C. Nonstable			D. Nonstable [?]			E. Nonstable			F. Nonstable [z]		
Indicator	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD	Med	Mean	SD
Brass Hill Census (BHill) [1]	0.041	0.045	0.031	0.042	0.046	0.031	0.041	0.045	0.031	0.043	0.046	0.031	0.029	0.030	0.010	0.033	0.036	0.028
Bennet Horiuchi No 1 (BH _1)	0.244	0.311	0.286	0.249	0.327	0.300	0.243	0.307	0.282	0.218	0.280	0.259	0.015	0.019	0.016	0.234	0.245	0.143
Bennet-Horiuchi No 2 (BH _2)	0.242	0.310	0.271	0.247	0.325	0.285	0.240	0.307	0.267	0.218	0.283	0.262	0.054	0.055	0.030	0.225	0.233	0.134
Bennet-Horicuhi No 3 (BH _3)	0.244	0.309	0.284	0.249	0.324	0.297	0.243	0.305	0.280	0.215	0.278	0.257	0.014	0.018	0.015	0.233	0.244	0.140
Bennet-Horiuchi No 4 (BH _4)	0.242	0.310	0.271	0.247	0.325	0.285	0.240	0.307	0.267	0.218	0.283	0.262	0.054	0.055	0.030	0.225	0.233	0.134
Bennet-Horiuchi No 5 (2SBH _4)	0.081	0.118	0.276	0.095	0.247	0.586	0.074	0.086	0.062	0.078	0.083	0.052	0.034	0.037	0.025	0.060	0.063	0.041
Brass-Martin(BMartin) [2]	0.114	0.154	0.155	0.109	0.124	0.089	0.116	0.162	0.167	0.105	0.134	0.118	0.057	0.072	0.055	0.091	0.124	0.113
Brass Hill (BHill) [1]	0.094	0.109	0.083	0.096	0.107	0.072	0.094	0.110	0.085	0.096	0.104	0.069	0.030	0.033	0.023	0.082	0.085	0.049
Preston Bennet (PB)	0.724	0.701	0.370	0.765	0.739	0.371	0.710	0.691	0.369	0.532	0.564	0.409	0.103	0.189	0.235	0.761	0.785	0.267
Preston Hill I (PH _1)	0.414	0.521	0.490	0.380	0.528	0.496	0.418	0.520	0.489	0.426	0.512	0.473	0.139	0.180	0.153	0.343	0.383	0.201
Preston-Hill 2 (PH _2)	0.412	0.508	0.456	0.401	0.513	0.463	0.439	0.507	0.454	0.424	0.500	0.439	0.179	0.208	0.158	0.372	0.374	0.202
Preston Lahiri No 1 (PL _1)	0.486	35.550	4655.569	0.492	7.233	101.118	0.483	42.630	5204.835	0.487	7.201	69.995	0.148	0.155	0.066	0.478	4.790	78.060
Preston-Lahiri No 2(PL _2)	0.450	13.137	640.044	0.448	19.491	974.494	0.431	11.548	524.100	0.449	3.116	14.227	0.207	0.221	0.116	0.413	12.901	323.772
N	31,500			6,300			25,200			700				4,320			10,368	

SD, standard deviation; Med, median.

[?] $_1 = \ _3 = 0$

$C_1 = C_2$ and $C_3 < 1$

[z] C_1 6 C_2 and $C_3 < 1$ and maxabs($C_1 \quad C_2$) < :10

[1] Values of errors in the Brass-Hill shown in the rst row correspond to errors associated with the ratio $C_1=C_2$. While values of Brass-Hill in the seventh row correspond to erros associated with relative completeness of death registration.

[2] BMartin is a variant of Brass classic method that relaxes the assumption of stability and assumes instead past mortality decline.

Scenario 2: $C_1 = 0$:85 if age [15-35], $C_1 = 0$:75 elsewhere; $C_2 = 0$:95 if age [15-35], $C_1 = 0$:85 elsewhere; $C_3 = 0$:85 if age [15-35], $C_1 = 0$:80 elsewhere.

Search for an optimal estimate is carried out considering all prior information available and the following are general rules:

i. In the absence of any knowledge whatsoever about errors or deviations from stability, the search for best method should be concentrated on the pooled sample subset in Tables 9{11, panel A.

ii. When exogenous information suggests stability and not much else, the search should focus on the subset of stable populations in Tables 9{11, panel B. Instead, when there is prior empirical data con rming violation of stability, for example past shifts in fertility regime, but one can be agnostic about completeness and age misreporting, the search of optimal method should concentrate on the population subset in Tables 9{11, panel C.

iii. When in addition to lack of stability there is evidence of defective coverage of population and death counts but no suggestion of signi cant net age overstatement at adult ages, the search should shift to the subset in Tables 9{11, panel D.

iv. When the researcher suspect a scenario like in (iii) above but, in addition, there is evidence of age misreporting, identi cation of optimal method should be done using Tables 9{11, panel E.

v. Finally, in cases scenario (iv) is most reasonable and one can establish that completeness of two censuses is (possibly) defective but equal in both censuses, identi cation of the optimal choice must be done with Tables 9{11, panel F.

The results displayed in Tables 9{11, panels A through F contain a number of salient characteristics. First, as already suggested in the work by Hill and colleagues, Brass's methods to estimate relative completeness of the two censuses is uniformly good, regardless of population subset. Second, with the exception of Brass methods, the magnitude of errors are larger when census coverage is defective as long as completeness is NOT the same in both censuses. This is because all methods except Brass's rely on direct computations of age speci c growth rates from the observed data, a quantity that will be in error when there is di erent coverage errors in two successive censuses. Indeed, the performance of these methods improves substantially when there is accurate census coverage or, equivalently,*when coverage is the same in both censuses* (Table 9, panel D). Fourth, age misreporting a ects the accuracy of all estimates but substantially more so in some cases (Brass's methods and the second variant of Preston-Hill) than in others (Bennett-Horiuchi all variants). Fifth, the magnitude of errors obtain when relative completeness is age dependent (last two columns of panels A-F in Tables 9{10) varies sharply by technique but, in general, are lowest in the method by Bennett-Horiuchi.

The most important inference from this evaluation exercise is as follows: if one excludes population subsets with defective census completeness, the optimal choice is always one of the variants of Bennett-Horiuchi method followed by the two methods proposed by Brass, irrespective of violations of stability assumptions or age misreporting. This suggests the following strategies:

i. In the absence of exogenous information about the di erence in completeness between the two census and if the assumption of age invariant completeness holds, use Brass method;

ii. In the absence of exogenous information, whether or not age dependence of relative completeness is suspected, use a two stage procedure: rst estimate relative completeness of

census enumeration using Brass' method, adjust intercensal rates of growth and then apply Bennett-Horiuchi method.

We use both strategies in LAC and when the di erence between estimates was less than 0.05 we compute the average of Brass and Bennett Horiuchi estimates. When their di erence exceeded 0.05 we chose the estimate from strategy (ii)[20].

6.2 Defective age reporting

Do the procedures to identify and adjust for age misreporting produce robust estimates of the true population parameters? To answer this question we select the subset of simulated populations with age misreporting and defective completeness, adjusted for completeness following strategy (ii) above, we identify the existence of age misreporting, and then correct for it using techniques (ii) in section 4.2. Tables 5 through 7 display the main results. First, Table 5 contains parameters associated with expression (4.5) and reveals that the t is almost perfect and that the estimated constant is unit, as it should be. Table 6 shows that when the procedure is reversed and we regress cmR_x on the vectors $_{1x=45;100}$ and $_{2x=45;100}$ the errors of estimates are tri e. This suggests that if an observed population belongs to the space of simulated populations, we can retrieve estimates of the magnitude of age net over-reporting that are highly accurate by simply using the estimated relation between the observed cmR_x and estimates $_{1x=45;100}$ and $_{2x=45;100}$ from the simulated populations.

7 Discussion: the issue of uncertainty

By an large the methods to adjust mortality statistics reviewed here perform satisfactorily provided the key assumptions on which they rest are concordant with the empirical conditions that produce the data. This is most unlikely to be the case always or even frequently for one single assumption and much less for combinations of assumptions. The conventional strategy has invariably been to scrutinize alternative estimates and then settle for one based on explicit or, more frequently, implicit reasoning and judgments about concordance of assumptions and observables. We believe we can improve upon this practice[21]

The evaluation study generates a superpopulation of errors associated with the application of each technique under conditions that violate to di erent degrees one or several of the cardinal assumptions on which they rely. It follows that for each technique we can de ne precisely the magnitude of error|however measured| associated with conditions that depart from the combination of assumptions in *ex ante* known ways. In our simulation the base universe of populations was generated by combining di erent demographic parameters (levels and patterns of fertility and mortality) thus producing multiple instances where one could alter conditions imparting changes that violate assumptions(lack of stability, adult migration, variable completeness, age misreporting that departs from assumed patterns etc.). As a consequence, we have all the information needed to de ne the frequency distribution of errors associated with one technique under one set of simulated conditions. And, in particular, one can de ne the probability that a singular technique will

[20] It is important to note that when relative completeness is age dependent, Bennett-Horiuchi is *mean optimal*,

Hill, K., Choi, Y., and Timaeus, I. (2005), \Unconventional approaches to mortality estimation," *Demographic Research*, S4, 281{300.

Hill, K., You, D., and Choi, Y. (2009), \Death distribution methods for estimating adult mortality: Sensitivity analysis with simulated data errors," *Demographic Research*, 21, 235{254.

Horiuchi, S. and Coale, A. J. (1985), \Age Patterns of Mortality for Older Women," .

Kamps E., J. (1976), *La declaracion de la edad en los Censos de Poblacion de la America Latina; exactitud y preferencia de d gitos en los Censos de 1950, 1960 y 1970*, vol. Serie C of *Series Historicas*, San Jose, Costa Rica: ONU, CEPAL, CELADE.

Martin, L. (1980), \A Modi cationfor Use in Destabilized Populations of Brass's Technqiue for Estimating Completeness of Death Registration," *Population Studies*, 34, 381{95.

Mazess, R. B. and Forman, S. H. (1979), \Longevity and age exaggeration in Vilcabamba, Ecuador," *J Gerontol*, 34, 94{8.

Nu-nez, L. (1984), \Una Aproximacion al Efecto de la Mala Declaracion de la Edad en la Informacion Demogra ca Recabada en Mexico," Report, Direccion General del Registro Nacional de Poblacion e Identi cacion Personal.

Ortega, A. and Garcia, V. (1985), \Estudio Sobre la Mortalidad y Algunas Caracteristicas Socioeconomicas de las Personas de la Tercera Edad: Informe de la Investigacion Efectuada en los Cantones de Puriscal y Coronado del 3 al 20 de Junio de 1985," Report, CELADE.

Palloni, A. (1990), \Assessing the Levels and Impact of Mortality in Crisis Situations," in *Measurement and Analysis of Mortality: New Approaches*, eds. Vallin, J., D'Souza, S., (nalys(y)-333 G 0 g 0-332(

Palloni, A. and4-228eltrst 0029(-Sst 0029(oac)27ef)1(s)-2e

Preston, S. H. and Lahiri, S. (1991), \A short-cut method for estimating death registration completeness in destabilized populations,"*Math Popul Stud*, 3, 39{51.

Rosenwaike, I. (1987), \Mortality Di erentials among Persons Born in Cuba, Mexico and Puerto Rico Residing in the United States, 1979-81,"*American Journal of Public Health*, 77, 603{606.

Rosenwaike, I. and Preston, S. H. (1984), \Age Overstatement and Puerto Rican Longevity," *Human Biology*, 56, 503{525.

Spencer, G. (1984), \Mortality among the Elderly Spanish Surnamed Population in the Medicare Files: 1968 to 1979," .

A Appendix. De nition of demographic pro les for the simulation

Five di erent master populations were created, one stable and four nonstable populations. In each

1, there is neither population nor death age overstatement or, if there is, their e ects cancel each other out. Expression (C.2) can be simpli ed if we expand the inner log expression in a Taylor series around a value of $f(x) = g(x)/h(x) = 1$:

$$\ln R^{o}_{x;[t_1;t_2]} \quad \ln \frac{h(x+k)}{h(x)} \quad I^{N}_{x;x+k} + \frac{g(x)}{h(x)} \quad 1 \quad (1 + I^{D}_{x;x+k}) + I^{D}_{x;x+k} \qquad \text{(C.3)}$$

an expression that reduces to 0 when $h(x+k)/h(x) = 1$ and $f(x) = 1$.

Expression (C.3) is the analytic support for inferences regarding the e ects of age overstatement on the index of age misstatement $cmR_{x;[t_1;t_2]}$