Title
Consistent population estimates: an application to Brazil

Author
Borges, Gabriel Mendes

Publication Date
2018

Peer reviewed|Thesis/dissertation

Consistent Population Estimates: an Application to Brazil

by

Gabriel Mendes Borges

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Demography

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Kenneth Wachter, Co-chair
Professor Dennis Feehan, Co-chair
Professor William Dow

Fall 2018

Consistent Population Estimates: an Application to Brazil

Abstract

Consistent Population Estimates: an Application to Brazil

by

Gabriel Mendes Borges

Doctor of Philosophy in Demography

University of California, Berkeley

Professor Kenneth Wachter, Co-chair

Professor Dennis Feehan, Co-chair

Demographers have long been aware that the practical application of basic demographic identities often leads to inconsistent estimates due to data quality limitations. Despite the considerable effort to produce demographic estimates and reconcile inconsistent demographic data in different contexts, this issue remains unsolved. This dissertation proposes a Bayesian probabilistic approach that allows for a concurrent estimation of consistent counts of population, mortality, fertility and migration. Old and new methods are combined, thus building upon well-established demographic techniques and statistical methods.

This dissertation addresses these issues in three chapters. Chapter 3 describes a set of methods to estimate and reconcile past demographic data, including measures of uncertainty. These methods are highly flexible and applicable to different contexts, both at national and subnational levels, with varying availability and quality of data. In addition to reconciling past population estimates, the methods detailed in this dissertation are concerned with estimating fertility, mortality and migration based on a set of raw observed data. Chapter 4 presents novel estimates and analysis of population, fertility, mortality and migration for Brazil and its 27 states for the period 1980-2010, based on several methods and data sources. Chapter 5 applies the data obtained in Chapter 4 to the methods presented in Chapter 3 to produce demographic estimates for Brazil and its three selected states.

This dissertation offers several important methodological contributions to the fields of formal demography and statistical demography. The application of the proposed methods for Brazil and its states unveils new perspectives for generating higher quality demographic estimates in the country. Among other findings, the results reveal higher internal migration flows than those estimated with census data, and fertility estimates that differ from previous work.

This dissertation also offers substantial methodological contributions specifically toward techniques of fertility and mortality estimation. Chapter 4 presents a new sensitivity analysis for the Brass P/F ratio method to evaluate the magnitude of bias in the results of its application when one or more conditions of the method are not met. This leads to a refinement

in the original method to correct for these biases. This chapter also introduces a method to adjust for bias on recent deaths in the household reported in censuses. The application of the method for Brazil shows that such adjustment greatly improves the efficiency of mortality estimation, particularly at old ages, resolving one of the main limitations of these data.

To Ivy and Lia

Contents

List of Figures

List of Tables

Acknowledgments

I am grateful to many people who have advised and supported me to during my PhD study. I would first like to thank my family. This journey would not have even started without their love, encouragement and support. Thanks to my parents for their love and effort to prioritize my education, despite the adversities they may have faced. To Nina and Isa for their love, friendship and motivation. Thanks to Ivy for agreeing to go on this journey with me, for her love, support and patience in moments of stress. Thank you, Lia, to increase my motivation to do everything I do. I am also grateful to my other relatives and friends who have supported me along the way.

I could not have finished this dissertation without the help of my committee members, who have directed me while allowing me to do my own work. I would like to thank my dissertation chair Ken Wachter for his great knowledge, his guidance, patience, motivation, and support in difficult times. To my co-chair Dennis Feehan, from whom I learned a lot in conversations about this and other projects, for his availability, encouragement and interest in my work. To Will Dow, who read my work and gave me many helpful comments. I would also like to thank Josh Goldstein, who kindly agreed to serve on my orals examination committee and gave insightful comments along the way.

I am greatly indebted to the staff of the Department of Demography, in particular Monique Verrier, who made my life easier while doing my PhD.

I am also thankful to my peers, students at the Department of Demography, and many other friends I made in Berkeley, for making this experience enjoyable and unforgettable.

I am grateful to Magali Barbieri and other friends from the Human Mortality Database, who provided me an opportunity to join their team and work on this great project.

Finally, I thank IBGE and Capes for the opportunity and financial support.

Chapter 1

Introduction

1.1 Background and Significance

The relevance of demographic estimates are closely related to the main interests of demography as a field of study, which is estimating the population size, its composition and geographic distribution, in addition to understanding the dynamics involving population change.

The composition of the population by age and sex is perhaps the most valuable demographic information. Past and future population data are used for planning and monitoring public policies and private decisions, funding allocations, weighting sample surveys, in addition to being denominator of several indicators. Beyond these uses, the population by age and sex interact with the other components of demographic change (fertility, mortality and migration), which are also of interest in their own right.

These interactions are given not only by the demographic interrelationships, but also through estimation procedures. A large set of methods have been developed to estimate fertility, mortality and migration based on the population's age distribution. Indirect methods to estimate fertility and mortality also make assumptions about migration and census coverage. Indirect methods to estimate migration also make assumption about census coverage, as well as intercensal mortality. Similarly, the components of demographic change have been used to asses census quality and coverage.

Improving methods for population estimates are important for better understanding past population trends, including their underlying uncertainty, but are also essential for population projections. Better demographic estimates are crucial to the improvement of population projections accuracy, not only because they provide better base estimates, but also because they improve the understanding of demographic parameters (Bulatao and Bongaarts, 2000). Due to the importance and sensitivity of past demographic estimates to projections, a large portion of the workload in population projections is devoted to past demographic estimation and reconciliation through evaluation, adjustments and consistency checking (Chackiel, 2009; UN, 2014).

In addition to produce more accurate population estimates, their underlying errors should

be assessed and quantified so they could be expressly incorporated in the uncertainty of population projections. To date, population projections are often calculated conditioned on the presumably known point estimates for the past. However, it is often the case that past population estimates contain considerable amount of uncertainty.

This is illustrated in Figure 1.1, which shows population estimates for Angola published by the last three United Nations Revisions of World Population Prospects. The incorporation of new data has considerably changed the estimated population for the entire series. The estimates for the year 2010 in the 2017 Revision is almost 20% higher than that published by the 2012 Revision, which is similar to the uncertainty in the projected population for Angola in 2045. This indicates that neglecting uncertainty in past demographic estimates tends to underestimate uncertainty in population projections.

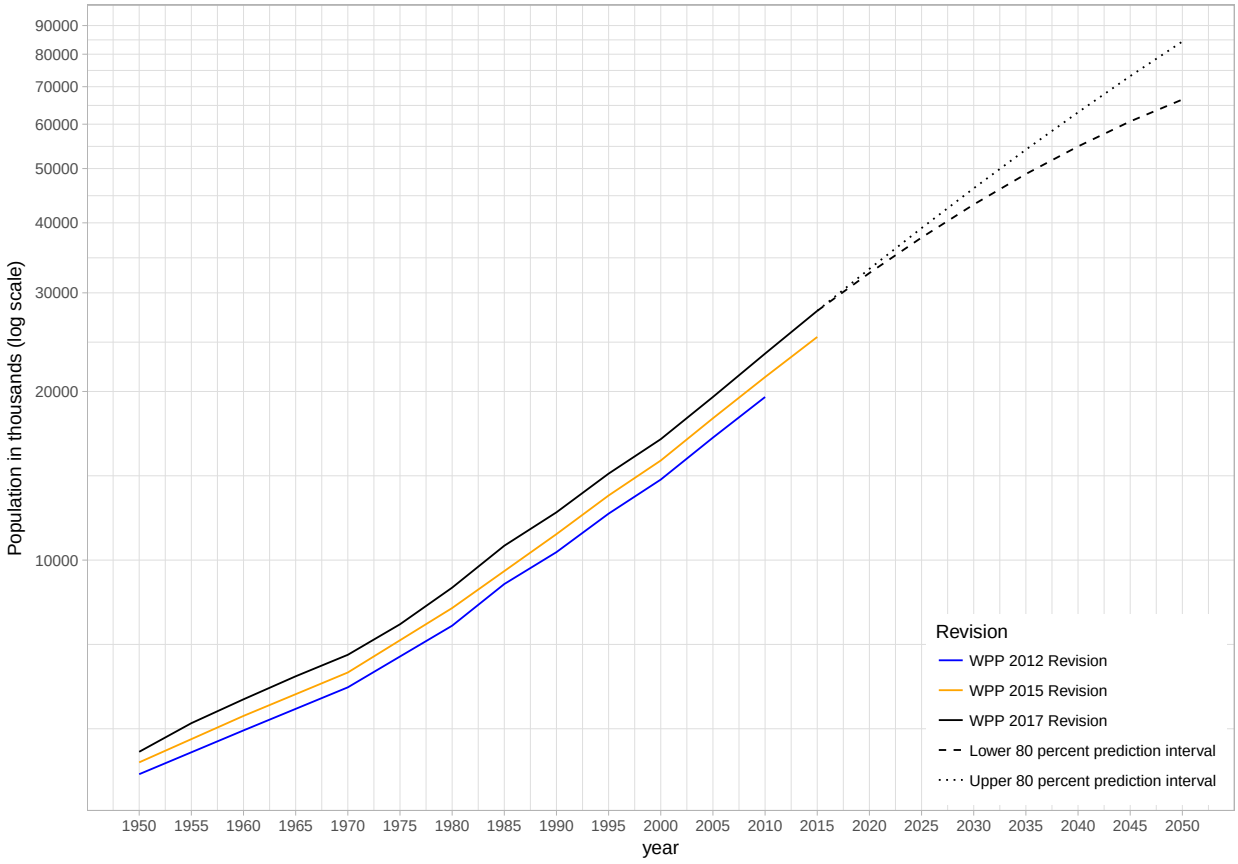

Figure 1.1: Total Population by year (thousands), Angola

1.2 Population Dynamics and Data Quality Issues in Demography

The interrelationship between demographic variables can be expressed by demographic accounting equations. The most basic identity in demography is the balancing equation, which states that the population at time $t+n$, K_{t+n}, equals the population at time t, K_t, plus the natural increase (births minus deaths) and the net migration (immigrants – emigrants) that occur between years t and $t+n$:

$$K(t+n) = K(t) + B(t,t+n) - D(t,t+n) + I(t,t+n) - O(t,t+n) \tag{1.1}$$

This relationship could be also applied to specific cohorts c:

$$K_c(t+n) = K_c(t) - D_c(t,t+n) + I_c(t,t+n) - O_c(t,t+n) \tag{1.2}$$

The youngest cohort at time $t+n$, which was not born at time t can be expressed by the number of births $B(t,t+n)$ between t and $t+n$ instead of $K_c(t)$.

This equation could be also defined by the well developed and satisfactory methods used to make population projections over discrete steps based on matrices and vectors. The key formula is given by:

$$K(t+n) = AK(t) \tag{1.3}$$

where A is the transition matrix containing the information on the probabilities of surviving and giving birth, called Leslie matrix (Leslie, 1945; Wachter, 2014). This is the basis of the cohort-component method for population projections, and can also include migration.

These demographic identities are overdetermined and, if one piece of information is missing, it can be estimated from the known values of the others. It is often the case that one or more pieces are missing or have data quality limitations, leading to "error of closure" and consequently incompatible data (Preston, Heuveline, and Guillot, 2001). Thus, although the identities for the true values hold, the observed population counts ($K^{obs}(t)$) are and demographic parameters (A^{obs}) are often inconsistent:

$$K^{obs}(t+n) \not\equiv A^{obs} \cdot K^{obs}(t) \tag{1.4}$$

Each piece of the balancing equation comes from a different data source and has its own limitations, requiring particular methods for assessment and further adjustment.

Population Counts

Population counts $K(t)$ normally come from censuses, which are affected by coverage and content errors. Coverage errors are, in general terms, related to failure in counting persons or housing units, leading to missing or duplicated cases. Content errors refer to mistakes

in the given information of persons or housing units effectively enumerated, such as age misstatement (US Bureau of the Census, 1985).

Coverage problems such as census undercount affect particular groups differentially. Children are known to be one of the most affected groups. Somehow surprisingly, this phenomenon has occurred in all sorts of population censuses, regardless the census design and cultural or socioeconomic characteristics of the country (O'Hare, 2015). Another characteristic that affects population distribution by sex and age is the differential undercount by sex, especially among young adults, in which the male population tends to be more undercounted that the female (Ewbank, 1981; Lee, 1982).

In addition to coverage errors, there are several problems that normally affect information about the age statement of people in a census. An important problem that affects information about the age of the population is the preference for ages with terminal digits, accumulating statements in numbers ending in 0 and 5. An additional and recurrent error observed in many populations is that the elderly tend to declare themselves older than they really are, an increasing trend with growing age. The shape of the population pyramid at old ages itself can convert random age heaping to net overcount among these age groups (Coale and Caselli, 1990; Del Popolo, 2000; Preston and Elo, 1999; Romero and Freitez, 2008).

Death counts

The natural sources of death counts and mortality estimations are the Civil Registration and Vital Statistics (CRVS) systems. However, in many cases, especially in low and middle income countries, the registration systems cover only part of the population and suffer from registration completeness and data quality problems (Mikkelsen et al., 2015). Random and systematic age misreporting is also a common problem in death counts, although this is thought to be less problematic for the reported deaths than for the population (Hill, 2017).

Indirect demographic methods have been proposed to estimate mortality in regions that present these data quality issues, particularly incompleteness of death registration. Since infant deaths are normally more underreported than adult deaths, special methods have been developed to estimate infant mortality, such as the those based on data about children ever born and surviving in censuses and surveys (Brass, 1971; UN, 1983).

Other methods for evaluating data quality and estimating adult mortality are: i) death distribution methods (Bennett and Horiuchi, 1981; Brass, 1975; Hill, 1987; Preston and Hill, 1980) (2) methods based on intercensal survival (UN, 1983), and (3) methods that use indicators of mortality levels based on survival of close relatives, such as parents (Brass and Hill, 1973) and siblings (Hill and Trussell, 1977), or other personal networks (Feehan, Mahy, and Salganik, 2017).

Birth counts

CRVS systems should also be the main source for fertility estimation, but they also suffer from coverage and quality problems, especially in developing countries. (Moultrie, 2013b).

These limitations stimulated the development of demographic methods for fertility estimations, often based on censuses and surveys.

One of the earliest method is the the own-children method, which consists of a reverse-survival technique that uses the population of children in a census to estimate fertility (Cho, Retherford, and Choe, 1986; Grabill and Cho, 1965). Fertility can be also estimated directly through data containing birth histories collected in surveys such as the Demographic and Health Survey (DHS). Another approach, perhaps the most used in developing countries, uses the recent and lifetime fertility measures routinely collected in censuses and surveys to estimate. These methods reconcile information from recent fertility, with the total parity by age group, which is thought to be more reliable. The most used technique is the PF ratio method (Brass, 1964; UN, 1983), which can also be applied to adjust observed recent birth counts from vital statistics.

Migration

Migration is probably the most difficult piece to measure in the balancing equation, due to the lack of administrative data. Indirect techniques have been also used to estimate migration, for instance by using intercensal residual methods or including surveys and census questions about relatives or household members living abroad (Hill and Dorrington, 2013).

Censuses often identify the non-native population as well, which combined with the country of origin, allows an estimation of the stock of immigrants and emigrants. Questions about the place of residence five years before the census date are also collected in some countries and give a better idea of recent immigration flow.

1.3 Balancing equation with adjustment factors

The balancing equation applied to a cohort (Equation (1.2)) can be adapted to include adjustment factors for the census data and registered birth and death counts to take into account the above-mentioned errors. Thus, it is possible to define a relationship between the observed data with the true, but unobserved parameters of interest. The true population at time t, $K_c(t)$, is expressed in terms of the observed population, e.g. in a census, $K_c^{obs}(t)$ and a factor $\kappa_c(t)$ that accounts for census coverage. Similarly, the true number of deaths $D_c(t, t+n)$ in the intercensal period is given by the registered deaths $D_c^{obs}(t, t+n)$ divided by the completeness of registered deaths, $\delta_c(t, t+n)$:

$$\frac{K_c^{obs}(t+n)}{\kappa_c(t+n)} = \frac{K_c^{obs}(t)}{\kappa_c(t)} - \frac{D_c^{obs}(t, t+n)}{\delta_c(t, t+n)} + I_c(t, t+n) - O_c(t, t+n) \tag{1.5}$$

The number of births $B(t, t+n)$ between t and $t+n$, equivalent to $K_0(t)$, is also expressed in terms of registered births $B_c^{obs}(t, t+n)$ and completeness of births $\beta_c(t, t+n)$.

Basic demographic identities, such as Equation 3.4, normally leads to a parameter redundant model, since it is not possible to estimate all the parameters. The resulting model is thus not identifiable and there is a family of solutions that would give consistent estimates.

Most of the methods that deal with demographic estimation are deterministic simulations. Census and vital registration data, even when adjusted, rarely result in sensible demographic estimates and the parameters in these models are normally inconsistent. Thus, it is necessary to design a framework to estimate and reconcile the existing data that violate the equalities stipulated by the population dynamics model. Developing these methods that incorporate measures of uncertainty is particularly relevant in the context of widespread need and use of probabilistic projections (see, for instance, Raftery et al., (2012)).

The methods proposed in this dissertation, in addition to reconciling past population estimates, are concerned with estimating population counts, fertility, mortality and migration based on a set of raw observed data. This study proposes a more formal way to incorporate demographic methods to produce adjustment factors for completeness of registered births and deaths and census coverage. For this purpose, this dissertation makes use of the known regularities in demographic rates as part of the estimation process. Finally, old and new methods are combined, exploring the use of demographic techniques that have been developed for many decades and statistical methods that have more recently been enhanced by the more intensive use of computing power.

1.4 Organization of the dissertation

This dissertation is divided in five chapters, including this introduction (Chapter 1). Chapter 2 reviews the relevant literature on the demographic techniques and methods that aim to produce consistent demographic estimates. Chapter 3 presents the general framework and the specific methods used to produce past demographic estimates. Chapter 4 presents estimates of population, fertility, mortality and migration for Brazil and its 27 states for the period 1980-2010, based on several methods and data sources. Chapter 5 applies the data estimated in Chapter 4 to the methods presented in Chapter 3 to produce demographic estimates, including population counts, fertility, mortality and migration, for Brazil and three selected states.

Chapter 2

Literature Review

2.1 Introduction

This section discusses briefly previous works that have tried to developed methods of demographic estimates dealing with inconsistencies in the data that are relevant to this dissertation.

2.2 Review of literature

Death Distribution Methods (DDM)

DDM are demographic techniques developed to estimate adult mortality through the estimation of completeness of registered deaths relative to census undercount. These methods were built on the demographic knowledge about the relationships between the distribution of deaths, the age structure and the rates of change in the populations. This set of methods can be divided into two major approaches: i) the Growth Balance methods and ii) the Synthetic Extinct Generations (SEG) methods (Hill, You, and Choi, 2009).

These indirect demographic techniques assume that the population is closed, the completeness of censuses and death registrations are constant by age and there is no age misreporting. In practice, these methods are more sensitive to some assumptions than to the others (Hill, You, and Choi, 2009).

These methods have been widely used, but they rely on strong assumptions about the pattern of census undercount and deaths under-registration, in addition to assuming closed population, which are unlikely to hold in practice. Chapter 4 will explore these methods and their limitations more carefully.

Inverse Projection and Back Projection

Inverse Projection and Back Projection are methods that, in general terms, intend to reconstruct past populations, using birth, death and population counts as inputs in order to estimate demographic rates (Lee, 1974, 1985, 2004; Wachter, 1986).

Reconstructing past populations has been done for different purposes. In a series of six "studies on the use of demographic census for reconstructing the movement of the Brazilian population", published in 1941 and 1942, Mortara, (1941) reconstructed the Brazilian population dynamics for the previous century based on census data and other assumptions about demographic components. Lee, (1974, 1985) reconstructed English demographic history from the 16th to the 19th centuries based on these techniques.

These approaches provide important insights to this dissertation, since they make use of limited information to reconstruct the entire history of populations.

The Demeny-Shorter method and its variations

Demeny and Shorter, (1968) were one the first authors to propose a comprehensive method to adjust the age distribution of two successive censuses. The method they propose assumes closed population and known intercensal mortality estimate, and then adjusts the age distribution based on an iterative process that adopts the same age-adjustment coefficient for two consecutive censuses ($\kappa_c(t+n) = \kappa_c(t)$):

$$\frac{K_c^{obs}(t+n)}{\kappa_c(t+n)} = \frac{K_c^{obs}(t)}{\kappa_c(t)} - D_c^{obs}(t, t+n) \tag{2.1}$$

Das Gupta, (1975) extends the idea underlying the Demeny-Shorter method to propose a method for estimating age-reporting errors from two consecutive census. This method relaxes the assumption that the age distortion is equal between the two censuses and adds assumptions about the pattern of age misreporting.

Ntozi, (1978) also builds on the Demeny-Shorter method, extending it for three censuses and leaving out the assumption that the size of age errors are constant at successive censuses. It is replaced by an assumption of regularity of change in error, that is, that a geometric mean relationship holds between error components of the same age group in the three censuses.

Lee, (1982) and Lee and Lam, (1983) apply the Demeny-Shorter with some refinements to adjust English Censuses. The authors maintain the main assumptions present in the original method though. One of the refinements is that the authors propose simultaneously estimating intercensal mortality and the census age adjustment coefficients, rather than estimating mortality separately.

Preston, Elo, et al., (1998) use a similar strategy to estimate the African American population by age group at census dates from 1930 to 1990 for cohorts for which the extinct-generation method and the regular demographic analysis based on birth and death counts cannot be performed. The method requires the information on death registration and migration to estimate the "true" cohort sizes through a model containing an age effect and a period

effect. The preferred model does not allow for interactions, assuming that the age-specific error is constant over time and the period-specific error is constant over age.

The main limitation of these techniques it that, in addition to the specific assumptions of each method, all of them depend heavily on the accuracy of the mortality schedule assumed or on the quality of the intercensal registered deaths.

Demographic Reconciliation

A procedure called "demographic reconciliation" has been used for many decades to construct population projections in the Latin American and Caribbean (LAC) Demographic Centre, which is the base for the United Nations population estimates for LAC countries (CELADE, 1968; Chackiel, 2009). This is also the method used by several National Statistical Office (NSO) in Latin America, for instance, in Brazil (IBGE, 2013b), Mexico (CONAPO, 2012) and Argentina (INDEC, 2013).The United Nations use a similar technique to reconstruct populations between 1950 and 2010, integrating all demographic components and obtaining the base population for their projections (Gerland, 2014).

This approach aims to obtain adjusted structures by age and sex from the enumerated population in the censuses using the knowledge about the parameters that represent the country's demographic dynamics. These are time-intensive techniques that use a trial and error method to reach a satisfactory solution, and depend on many subjective decisions carried out by the specialists preparing the population estimates. Alkema, Raftery, Gerland, et al., (2012) have also expressed concerns about the difficulty to reproduce estimates using these approaches.

Luther and Retherford, (1988) propose a method for simultaneously correcting two or more censuses and intercensal births and deaths. The method requires a set of initial correction factors for births and age-specific correction factors for death and census counts. Given these preliminary correction factors, not necessarily consistent, the method seeks for an optimal set of final consistent correction factors yielding a comprehensive set of estimates that deviate as little as possible from the preliminary ones and at the same time ensure internal consistency. New alternatives have been proposed to deal with the limitations of the methods currently available. Wheldon et al., (2013, 2016), for instance, propose a method that reconstructs historical demographic parameters using a Bayesian hierarchical approach, estimating age-specific population counts and fertility, migration and mortality rates. In addition to simultaneously estimating demographic parameters, considering the uncertainty associated with historical demographic data, this method innovates by incorporating measurement errors.

The methods described in the previous paragraph seem to work well to reconcile existing estimates. However, they are highly dependent on the preliminary correction factors or bias-reduced initial estimates. These factor are assumed to be known and their initial estimates are not part of the model. The adjustment factors for the deaths would come, for instance, from the application of the death distributions methods. In fact, these methods only resolve the inconsistencies that remain in the already adjusted initial set of estimates.

2.3 Summary

There has been a considerable effort to develop methods that produce consistent demographic estimates. They have produced useful insights about the relationships between the demographic parameters and the regularities in the demographic rates and data quality limitations. However, it remains unclear how to combine these different techniques and the accumulated demographic knowledge to produce population estimates using imperfect data.

Chapter 3

Methods

3.1 Introduction

As discussed in the previous chapters, although mechanistically simple and straightforward, demographic equations rarely produce consistent estimates due to the lack or inaccuracy of data. Most of the techniques that deal with these inconsistencies in demographic estimation are deterministic simulation methods. Gerland, (2014) and UN, (2014), for example, manage this by using a trial and error method, in which an initial set of estimates is provided, the results are validated, checked for consistency and then adjusted and re-estimated until satisfactory results are achieved. These are time-intensive tasks and depend on several decisions about data sources and population groups to be privileged. Furthermore, the application of these techniques rarely produces uncertainty measures.

Improving methods for population estimates are important for better understanding past population trends, including their underlying uncertainty, but are also essential for population projections. Better demographic estimates are crucial to the improvement of population projections accuracy, not only because they provide better base estimates, but also because they improve the understanding of demographic parameters (Bulatao and Bongaarts, 2000).

In addition to produce more accurate population estimates, their underlying errors should be assessed and quantified so they could be expressly incorporated in the uncertainty of population projections. To date, population projections are often calculated conditioned on the presumably known point estimates for the past. However, it is often the case that past population estimates contain considerable amount of uncertainty. A more formal way to make inference about the parameters of interest in the presence of uncertainty are probabilistic approaches, which are consistent with the well stablished literature on probabilistic methods for population projections (Bulatao and Bongaarts, 2000; Lee, 1998; Lee and Carter, 1992; Raftery et al., 2012).

Probabilistic approaches have been used to estimate fertility and mortality for contexts with incomplete vital registration systems. Alkema, Raftery, Gerland, et al., (2012) and Liu and Raftery, (2017) developed methods to incorporate the uncertainty of past TFR by

using a method for estimating the bias and variance of different sources of data with varying data quality, mostly censuses and surveys. These approaches take into account sampling and non-sampling errors, which are evaluated through comparison with official estimates.

The idea of incorporating multiple sources to produce point estimates and measures of uncertainty has been also applied to child (Alexander and Alkema, 2018; Alkema, New, et al., 2014; Rajaratnam et al., 2010) and maternal mortality (Alkema, Chou, et al., 2016).

As pointed out by Hill, (2017), these are promising approaches, but have not been widely applied to adult mortality. Recent studies have used Bayesian hierarchical methods to produce mortality estimates for small areas, taking into account the high stochastic variation in death counts at these levels (Alexander, Zagheni, and Barbieri, 2017). In addition to account for sample variation, some contexts also require the incorporation of uncertainty measures to account for incomplete or underreported count data. A few studies have proposed Bayesian approaches to deal with this type of problem, in which the true counts are often assumed to be Poisson distributed, with another distribution representing the probability of the event being reported (Anderson, T. Bratcher, and Kutran, 1994; Dvorzak and Wagner, 2016; Moreno and Girón, 1998; G. L. Oliveira, Loschi, and Assunção, 2017; Schmertmann and Gonzaga, 2018).

Migration is probably the most difficult piece to measure in the balancing equation due to the lack of administrative data, leading to high levels of uncertainty. In addition to random variation, migration estimates may also be subject to measurement and sampling error. A few studies have tried to incorporate uncertainty in these estimates. Kintner and Swanson, (1993) estimate net migration by residual, comparing two consecutive censuses, and incorporate measures of uncertainty due to random variation in mortality rates and to census measurement errors. Passel, (2007) also use the "residual method" to estimate the number of unauthorized migrants in the United States, which involves comparing an analytic estimate of the legal foreign-born population with a survey-based measure of the total foreign-born population. Because the residual estimate is based in part on sample data from the CPS and in part on a demographic estimate, the resulting figure is subject to both sampling and estimation error. A. J. Garcia et al., (2015) estimate migration flows in sub-Saharan Africa through census questions that asks where individuals lived previously, and model these outcomes according to socioeconomic variables known to have relationships to migration. To estimate international migration in European countries, Raymer et al., (2013) use a Bayesian model that incorporates limitations of multiple and sparse data sources collected through different systems and designs. The study also incorporates covariate information and expert opinion about the effects of undercount, measurement, and accuracy of data collection systems.

Despite advancements in techniques for fertility, mortality and migration estimation independently, there remains a need for stochastic integrated population models, particularly in contexts of faulty data. Robinson et al., (1993) propose a framework to estimate the coverage of the United States census by age, sex and race based on Demographic Analysis by attributing random variation to adjustment factors of the demographic components (births, deaths and migration).

More recent studies have tackled this problem by using Bayesian approaches that are closely related to what is proposed in this dissertation. Bryant and Graham, (2013, 2015) derive population estimates from multiple administrative data sources, with application for subnational levels in New Zealand. The authors use several data sources that are proxies of the demographic stocks and flows and model these by incorporating measures of uncertainty that captures how well the data represents the true demographic processes. This is an interesting approach to be used in contexts of high quality administrative data, but it is more difficult to apply in to regions that lack this type of data. Wheldon et al., (2013, 2016) also propose Bayesian hierarchical models to incorporate measurement error in fertility, mortality, migration and population counts and reestimate past populations using expert opinion errors affecting each component. This approach often requires bias-adjusted estimates which should be performed by the analyst prior to the application of the method. By combining this information with prior standard deviations of the initial estimates, the method resolves the inconsistencies that remain in the already adjusted initial set of estimates.

The methods proposed in this dissertation, in addition to reconciling past population estimates, is concerned with estimating population counts, fertility, mortality and migration based on a set of raw observed data. By modeling the observed counts, often as a Poisson distribution, the proposed method also take into account stochastic variation that are particularly relevant for small population groups. Finally, the Bayesian Melding approach is used, allowing for specification of prior distributions for both inputs and outputs of the model.

Modeling population and the components of demographic change simultaneously as an integrated population model has several benefits, and this approach has been increasingly adopted by population ecologists (Kéry and Schaub, 2012; Pizarro and Villa, 2005). Since population and demographic rates are directly connected through demographic identities, the information about one or more processes improves the estimation of the others, by either increasing the precision or avoiding spurious accuracy. If the population of children, for example, is consistent with births and infant mortality estimates for that same cohort, the combination of the three estimates should provide more accurate measures. Conversely, if there is certain inconsistency in multiple sources of data, for example if the difference in the population of a certain cohort between two years is not completely explained by the mortality and migration in this period, the incorporation of different data sources and their uncertainty allows for more accurate estimates.

The appeal of using Bayesian analysis in these contexts lies in its potential to overcome the challenges of combining information from different sources and dealing with high stochastic variation, measurement errors and lack of identifiability in the models.

One method that deals with these challenges is the Bayesian melding approach (Poole and Raftery, 2000), which is adopted in this study to pool several demographic information.

3.2 Bayesian Melding

This section describes the method used to make inference about the demographic parameters of interest, namely the adjustment factors for censuses, completeness of death and birth counts and migration estimates, which in combination with the observed counts, would result in consistent demographic estimates. This is done by making use of the demographic balancing equation, indirect demographic techniques and the knowledge on the regularities of demographic rates, estimated via Hamiltonian Monte Carlo (HMC) algorithm and the Bayesian melding approach.

Bayesian melding is a probabilistic approach that allows for the specification of uncertainty in both inputs and outputs of the model. It is based on a deterministic model M that relates a vector of input variables θ to a vector of output variables ϕ: $M : \theta \rightarrow \phi$. It was first proposed to model population dynamics models in ecology (Poole and Raftery, 2000), but its use has been expanded to model land use and transportation policy (Sevcikova, Raftery, and Waddell, 2007), HIV prevalence (Alkema, Raftery, and Clark, 2007; Clark, J. R. Thomas, and Bao, 2012) and mortality (Sharrow et al., 2013).

In this study, the deterministic model that links the inputs and outputs is given by the demographic balancing equation, which states that the population from cohort c at time $t + n$, $K_c(t + n)$, equals the population at time t, $K_c(t)$, minus the deaths plus the net migration (immigrants − emigrants) that occur between years t and $t + n$:

$$K_c(t+n) = K_c(t) - D_c(t, t+n) + I_c(t, t+n) - O_c(t, t+n) \tag{3.1}$$

The youngest cohort at time $t + n$, $K_0(t + n)$ which was not born at time t, can be expressed by the number of births $B(t, t + n)$ between t and $t + n$ instead of $K_0(t)$.

This equation could be also defined by the well developed and satisfactory methods used to make population projections over discrete steps based on matrices and vectors. The key formula is given by $K(t + n) = AK(t)$, where A is the transition matrix containing the information on the probabilities of surviving and giving birth, called Leslie matrix (Leslie, 1945; Wachter, 2014). This is the basis of the cohort-component method for population projections, and could also include migration.

$$\underbrace{\begin{pmatrix} K_0(t+n) \\ K_1(t+n) \\ K_2(t+n) \\ \vdots \\ K_\omega(t+n) \end{pmatrix}}_{\mathbf{K(t+n)}} = \underbrace{\begin{pmatrix} F_1 & F_2 & F_3 & \dots & F_\omega \\ S_1 & 0 & 0 & \dots & 0 \\ 0 & S_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & S_\omega & 0 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} K_1(t) \\ K_2(t) \\ K_3(t) \\ \vdots \\ K_\omega(t) \end{pmatrix}}_{\mathbf{K(t)}} \tag{3.2}$$

where S_c is the ratio of the population at time $t + n$, $K_c(t + n)$, to the population at time t, $K_c(t)$, from the same cohort c. It can be also interpreted as the combination of survival and migration rates. The terms on the first row of the Leslie Matrix A represent the fertility

and infant mortality rates, which, when multiplied by the number of women at reproductive ages, generate the population under age 5. Mathematically, this term is defined as:

$$F_c = B_c^f \cdot S_0 \tag{3.3}$$

where B_c^f is the number of girls born from women of cohort c between years t and $t+n$ and S_0 is the survival and migration rates of these girls during the same period.

Complete demographic information for cohorts are not always available. Mortality, fertility and migration for the period $[t; t+n]$ are often calculated by age groups instead of birth cohorts, with the denominator being the the person-years lived between years t and $t+n$. This requires a few special adaptations to the functions in the Leslie matrix, which will be detailed in Chapter 4.

The deterministic demographic equations 3.1 and 3.2 are straightforward in the absence of measurement errors. In fact, they are overdetermined and, if one piece of information is missing, it can be estimated from the known values of the others.

However, true population counts and demographic events are rarely known and need to be estimated based on observed data. Even though the equations of the true demographic parameters, by definition, match perfectly, observed data are often inconsistent due to measurement errors. Therefore, demographic parameters need to be estimated in order to reconcile the inconsistent data.

It is possible to define a relationship between the observed data with the true, but unobserved parameters of interest. The true population at time t, $K_c(t)$, is expressed in terms of the observed population, e.g. in a census, $K_c^{obs}(t)$ and a factor $\kappa_c(t)$ that accounts for census coverage. Similarly, the true number of deaths $D_c(t, t+n)$ in the intercensal period is given by the registered deaths $D_c^{obs}(t, t+n)$ divided by the completeness of registered deaths, $\delta_c(t, t+n)$:

$$\frac{K_c^{obs}(t+n)}{\kappa_c(t+n)} = \frac{K_c^{obs}(t)}{\kappa_c(t)} - \frac{D_c^{obs}(t, t+n)}{\delta_c(t, t+n)} + I_c(t, t+n) - O_c(t, t+n) \tag{3.4}$$

The number of births $B(t, t+n)$ between t and $t+n$, equivalent to $K_0(t)$, is also expressed in terms of registered births $B_c^{obs}(t, t+n)$ and completeness of births $\beta_c(t, t+n)$.

Basic demographic identities, such as Equation 3.4, normally lead to non-identifiable models, since there is a family of solutions that would produce consistent estimates. Therefore, statistical methods are required to address this estimation problem.

In this dissertation, inference about the parameters of interest is done in two steps. First, the individual pieces that make up the balancing equation 3.4, namely population, mortality, fertility and migration are estimated. Secondly, the left and right-hand sides of the equation need to be reconciled, since they are likely to have inconsistent probability distributions. In other words, the probability distribution of the population at time $t+n$, $K_c(t+n)$, estimated based solely on the $t+n$ census and its census coverage is harmonized with the the probability distribution induced by the population at time t, $K_c(t)$, and the demographic events during

the period $[t; t+n]$. Poole and Raftery, (2000) named this procedure of combining the distributions of the output with the distribution induced by the inputs "Bayesian Melding".

The debate about combining information from the inputs and outputs is equivalent to the extensively discussed issue of aggregating expert opinions. The two main methods used for this purposes are the linear pooling and the logarithmic pooling. Linear pooling is simply a weighted average of the probability densities and logarithmic pooling is equivalent to a normalized weighted geometric mean, which in turn is identical to applying the single average to the logarithms of the densities (Genest and Zidek, 1986; Givens and Roback, 1999; Poole and Raftery, 2000).

Logarithmic pooling has been preferred to linear pooling given some of its properties. First, logarithmic pooling is typically unimodal and less dispersed than linear pooling and thus is more likely to indicate consensual values or the overlap of different probability distributions and not only a simple representation of the diversity in the distributions. Second, and most importantly in a Bayesian framework, is that fact that this operation is insensitive to the order in which the prior is pooled and updated, which is called "external Bayesianity" (Genest, McConway, and Schervish, 1986; Genest and Zidek, 1986; Givens and Roback, 1999; Poole and Raftery, 2000).

In the model described previously, the two distributions are combined by a pooling weight α which should reflect the reliability of each information. As in the linear pooling, there is no clear normative basis for choosing the pooling weights (Genest and Zidek, 1986), although there have been some attempts to propose more formal ways to estimate these parameters (de Carvalho et al., 2015). In this dissertation, pooling is performed by giving equal weights to the different priors $\alpha = 0.5$, since there is not reason to prioritize one over another.

The next sections describe the setup of the models used to define the likelihood and prior distributions of each component of the demographic balancing equation.

3.3 Modeling population counts

There are several ways to model population K_c [1]. One option is to use a normal or log-normal distribution, under the assumption that errors are randomly centered around zero.

However, observation errors in population estimates are often systematic, i.e., they have nonzero mean, and the model should be able to correct for this bias. If the observed population derives from a census, it is usually smaller than the true population, meaning that census undercount exceeds overcount. The census counts K_c^{obs} could then be assumed to follow a binomial distribution:

$$K_c^{obs} \sim Binomial(K_c, \kappa_c) \tag{3.5}$$

[1] For simplicity, the term t is omitted in this section, since the procedure to model population at different years is the same

where K_c is the true but unobserved population and κ_c is the census coverage for cohort c [2].

Even though the most common use of the binomial distribution is to estimate the probability of success κ_c given the number of successes K_c^{obs} in a series of experiments K_c, statisticians have also tried to make inference about the true but unobserved parameter number of trials (K_c). This has been done both from a frequentist (Carroll and Lombard, 1985; Olkin, Petkau, and Zidek, 1981) and a Bayesian (Blumenthal and Dahiya, 1981; Draper and Guttman, 1971; Raftery, 1988; Smith, 1991) approach.

This issue is often called the "binomial n problem" and has been also addressed in the context of estimating total population through capture and recapture models in wildlife (Otis et al., 1978) and human populations (Wolter, 1986).

In human populations, K_c^{obs} normally comes from censuses and κ_c and K_c are parameters to be estimated. The Post-Enumeration Survey (PES) is the natural data source to model κ_c in equation 3.5, if the survey was carried out in the own country under analysis and the results are available. Information about K_c is much harder to obtain.

The binomial model is limited because it contains only one free parameter and the variance is determined by the mean. When estimating census counts, for example, both moments are calculated from the coverage estimation of the PES.

More importantly, the model in equation 3.5 only accounts for random sampling errors. Sampling errors in the PES are relatively easy to control and quantify and depend mostly on the sample size. This tends not to be a problem for large population groups, as those used in this study. With increases in the sample sizes of the PES, sampling errors have been dominated by uncertainties due to systematic non-sampling errors. Non-sampling errors are harder to identify and measure and arise from many sources, such as correlation bias, processing and matching errors, among others (Wachter and Freedman, 2000) [3]. Freedman and Wachter, (2003) suggest that large PES sample sizes not only increase the relative importance of non-sampling errors, but also make them more problematic, since bigger samples are harder to manage so that systematic errors are made more difficult to control and measure.

To take these issues into account, an over-dispersed version of the binomial distribution is required. In a Bayesian framework, the most used one is the beta-binomial distribution, where the probabilities of success, in this case κ_c, follow a beta distribution. The beta distribution is defined in the interval [0, 1] and parametrized by two shape parameters, a_c^K and b_c^K:

$$\kappa_c \sim Beta(a_c^K, b_c^K) \tag{3.6}$$

This mixture gives an algorithm for simulating from the beta-binomial: draw from the prior distribution $\kappa_c \sim Beta(a_c^K, b_c^K)$ and then draw $K_c^{obs} \sim Binomial(K_c, \kappa_c)$ (Gelman et

[2] If there is evidence that κ_c can assume values greater than one, i.e., there is census overcount, K_c^{obs} has to be modeled with a different distribution. An alternative is to adopt the same procedure used to model death counts, which is described in section 3.4

[3] see discussion about PES errors in Chapter 4

al., 2013).

Section A.2 in the Appendix A provides more details about the beta distribution, its properties, estimation methods and conjugacy with the binomial and Poisson distributions.

A conceptual difficulty with Bayesian analysis in the "binomial n problem" is to find sufficiently flexible and tractable family of prior distributions for the discrete parameter "n", in this case K_c (DasGupta and Rubin, 2005; Raftery, 1988). This has also some practical problems, since some statistical programs and modeling languages do not perform inference for discrete unknown parameters and these discrete parameter models need to be re-expressed as mixture models with continuous parameters (Team, 2017).

A natural alternative to overcome these difficulties is to model population counts with a Poisson distribution, which captures a feature often observed in count data, that is the observation error increases with the population size. More precisely, in the Poisson distribution, the variance is equal to the mean.

Thus, this dissertation assumes that the true population counts K_c follow a Poisson distribution with rate λ_c:

$$K_c \sim Poisson(\lambda_c) \tag{3.7}$$

As previously mentioned, prior information on λ_c is much harder to obtain, and a non-informative prior on this quantity is often required. Section A.1 in the Appendix A discusses non-informative priors for the Poisson distribution. This dissertation uses a uniform prior on the positive real line:

$$\lambda_c \sim Uniform(0, \infty) \tag{3.8}$$

$$\lambda_c \propto 1 \tag{3.9}$$

This is an improper prior, but leads to a proper prior for K_c^{obs}, as shown in Section A.2 of Appendix A.

Section A.2 also shows that the distributions described above (3.5, 3.6, 3.7, 3.8) lead the following hierarchical structure:

$$K_c^{obs} \sim Poisson(\lambda_c \cdot \kappa_c) \tag{3.10}$$

$$\lambda_c \propto 1 \tag{3.11}$$

$$\kappa_c \sim Beta(a_c^K, b_c^K) \tag{3.12}$$

The posterior distribution of equation 3.10 results from the combination of the likelihood for K_c^{obs} and the priors for κ_c and λ_c. The resulting posterior distribution for K_c provides information about the plausibility of different values for the total population given the observed data and the prior knowledge about these parameters.

Table 3.1 summarizes the hierarchical structure to model population data, showing the distribution for each variable, a brief description, the data source / estimation strategy, the

type of information (data, parameter or hyperparameter) and the sections they are discussed in different chapters of this dissertation.

Variable	Distribution	Description	Data source Estimation strategy	Type	Section
K_c^{obs}	$\sim Poisson(K_c \cdot \kappa_c)$	observed population counts	census	data	3.3 4.2, 5
K_c	$\sim Poisson(\lambda_c)$	true but unobserved population counts	-	parameter	3.3
λ_c	$\propto 1$	rate parameter	ignorance prior (uniform prior)	hyper-parameter	3.3 A.1
κ_c	$\sim Beta(a_c^K, b_c^K)$	census coverage	method of moments	parameter	3.3, A.2 4.2, 5
a_c^K, b_c^K	–	shape parameters	PES/ expert opinion	hyper-parameters	3.3, A.2 4.2, 5

Table 3.1: Population Model Summary

To illustrate this model, take the population of children observed in the 2010 Census in Brazil ($K_{0-4}^{obs} = 13796$ thousands) and assume that census coverage for this age group is beta-distributed as follows: $\kappa_{0-4} \sim Beta(15.3, 1.7)$ [4].

Figure 3.1 shows the posterior distributions of K_{0-4} and κ_{0-4} for 50000 HMC draws and Table 3.2 shows the 10th, 50th and 90th percentiles and the mean estimates of the posterior distributions for both parameters. The prediction interval shows that 80% of the posterior distribution of K_{0-4} lies between 14134 and 17225. This is a relatively wide interval that is largely driven by the choice of the prior for κ_{0-4}. In the full model, this information will be combined with other probability distributions, e.g. fertility and infant mortality, and the this distribution might change depending on how consistent this estimate is with other information.

The prediction interval shows that 80% of the posterior distribution of K_{0-4} lies between 14134 and 17225. This is a relatively wide interval that is largely driven by the choice of the prior for κ_{0-4}. In the full model, this information will be combined with other probability distributions, e.g. fertility and infant mortality, and the this distribution might change depending on how consistent this estimate is with other information.

[4] see Section A.2 for details about estimation of these hyperparameters

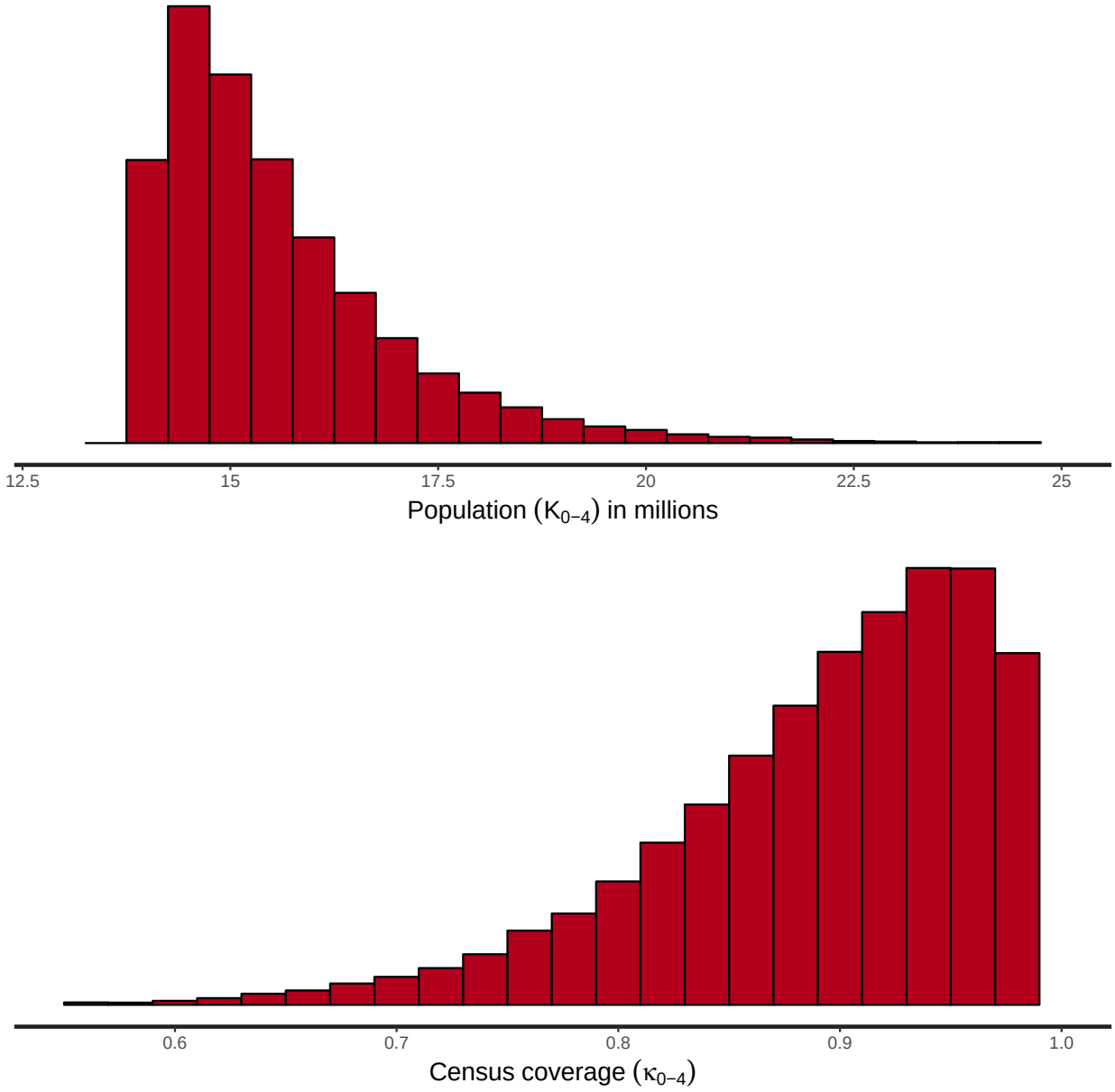

Figure 3.1: Illustration of the posterior predictive distribution of the population aged 0-4 (in thousands) and census coverage for the same age group. Brazil, 2010

Parameter	Percentile			Mean
	10th	50th	90th	
K_{0-4}	14134	15074	17225	15446
κ_{0-4}	0.8011	0.9154	0.9755	0.8996

Table 3.2: Summary statistics of the posterior predictive distribution of the population aged 0-4 (in thousands) and census coverage for the same age group. Brazil, 2010

3.4 Modeling death counts and mortality

Modeling death counts

Demographic events such as births and deaths, even when derived from complete vital statistics systems, are subject to random variation and may be assumed to follow a Poisson distribution. Brillinger, (1986) presents a reasoning for a Poisson distribution for the number of deaths, based on the assumptions of a Poisson birth process and independent lifetimes. Thus, the deaths D_c in a population K_c from cohort c are distributed as follows:

$$D_c \sim Poisson(K_c \cdot h_c) \tag{3.13}$$

where h_c is the mortality rate for cohort c.

When death counts are underreported, which is the case for many regions of the world (Mikkelsen et al., 2015), the number of registered deaths, D_c^{obs}, may be modeled using a binomial distribution, following the same strategy used to model population counts:

$$D_c^{obs} \sim Binomial(D_c, \delta_c) \tag{3.14}$$

where δ_c is the probability of a death being reported.

Equations (3.13) and (3.14) can be combined into a single Poisson distribution for the registered deaths, similar to what was done for modeling population counts (Section 3.3) and is shown in Section A.2 of Appendix A:

$$D_c^{obs} \sim Poisson(K_c \cdot h_c \cdot \delta_c) \tag{3.15}$$

Most studies that deal with similar issues of making inferences about the parameters in equation 3.15 have adopted a Bayesian approach due to identification problems that arise in a likelihood based analysis, since there is a range of values of K_c, h_c and δ_c that maximizes the likelihood. In other words, the likelihood that derives from 3.15 only allows inference about the product $(K_c \cdot h_c \cdot \delta_c)$, and gives no possibility to estimate the parameters K_c, h_c and δ_c individually, which are ultimately the measures of interest. Identification requires additional information, for example prior information about δ_c or h_c (Dvorzak and Wagner, 2016; Moreno and Girón, 1998).

It is often the case that the population at risk, K_c is taken as known even in a Bayesian context, where several types of uncertainties are incorporated into the estimates (Alexander,

Zagheni, and Barbieri, 2017; Schmertmann and Gonzaga, 2018). Census data are normally used as a proxy of the true population, which is a very strong assumption, as discussed in the previous sections.

The model proposed here relaxes this assumption by modeling mortality jointly with population counts, according to the procedure described in Section 3.3. There are at least two alternatives to integrate this information with equation 3.15.

The first one is to incorporate equation 3.15 explicitly into the hierarchical structure from Section 3.3 and express the number of observed deaths in terms of the true, but unknown, population counts K_c.

Alternatively, the observed death counts may be modeled in terms of the observed population counts (K_c^{obs}). Let δ_c^* be the completeness of registered deaths relative to the census coverage:

$$\delta_c^* = \frac{\delta_c}{\kappa_c} \tag{3.16}$$

$$\delta_c = \delta_c^* \cdot \kappa_c \tag{3.17}$$

Since $K_c = \frac{K_c^{obs}}{\kappa_c}$, it follows from equation 3.15 that:

$$D_c^{obs} \sim Poisson\left(\frac{K_c^{obs}}{\kappa_c} \cdot h_c \cdot \delta_c^* \cdot \kappa_c\right) \tag{3.18}$$

$$D_c^{obs} \sim Poisson(K_c^{obs} \cdot h_c \cdot \delta_c^*) \tag{3.19}$$

The latter approach is preferable because the completeness of registered deaths relative to the census coverage, δ_c^*, is a more precise interpretation of results from demographic techniques developed to estimate underregistration of deaths, namely Death Distribution Methods (DDM). Furthermore, this approach allows for independence between the mortality rates and completeness of registered deaths with the census coverage. Notice that δ_c^* is estimated from DDM, which make no use of the Post-Enumeration surveys, the main data source for estimating κ_c. Had equation 3.15 been used, δ_c would have been estimated by using κ_c. This would have led to a duplicated use of κ_c in the model.

Prior distributions for δ_c^*, which are assumed to be constant by age, come from the application of DDM. The hyperparameters of the distribution can be based on the application of DDM for different age groups and using different methods or on expert opinion. This is the approach adopted by Schmertmann and Gonzaga, (2018). Alternatively, it is possible to obtain prior distributions from previous studies. Murray et al., (2010), for example, find that the application of these methods can result in errors around $\pm 20\%$.

The following equations show the hierarchical structure for the mortality model.

$$D_c^{obs} \sim Poisson(K_c^{obs} \cdot h_c \cdot \delta_c^*) \tag{3.20}$$

$$\delta_c^* \sim Gamma(a_c^D, b_c^D) \tag{3.21}$$

Notice that δ_c^* is modeled by a gamma distribution, since it can assume values greater than one, which will occur when completeness of registered deaths is higher than census coverage. In the full model, the actual completeness of deaths δ_c is still restricted to the interval $[0; 1]$. The strategy used to estimate h_c will be detailed in the next section.

Mortality rate for the female population in Brazil aged 30-34 (h_{30-34}) is calculated to illustrate the estimation process in this section. The data required is the census population of this group ($K_{30-34}^{obs} = 8026855$) and the number of registered deaths in 2010 ($D_{30-34}^{obs} = 7837$). Prior distribution for census coverage is again assumed to be beta-distributed as follows: $\kappa_{30-34} \sim Beta(15.3, 1.7)$ and completeness of registered deaths is assumed to follows a gamma distribution $\delta_{30-34}^* \sim Gamma(162, 180)$.

Figure 3.2 shows, in the diagonal, the histograms of the posterior distributions for five parameters in the model: K_{30-34}, D_{30-34}, κ_{30-34}, δ_{30-34}, h_{30-34}. Notice that δ_{30-34}, by definition, only assumes values lower than one, even though the prior of δ_{30-34}^* has significant probability mass above one. The posterior distribution for h_{30-34} shows that under these priors the mortality rate for female aged 30-34 is likely to be between 10‰ and 12‰.

The lower panels show the scatterplots of half of the HMC chains and the upper panel the other half. The comparison between both panels, in addition to the other diagnostic parameters, indicate that the model converged satisfactory. The plots also show that there is a high correlation between the parameters, except between h_c and the pair $(D_{30-34}; \delta_{30-34})$, which reinforces the independence issue between these parameters discussed above.

This illustration shows the simplified version of the model for only one age group. The full model includes an additional piece that models h_c by incorporating the regularities in mortality rates across ages. This procedure is discussed in the next section.

Modeling mortality schedules

Demographers have developed methods to estimate demographic parameters in a wide range of contexts, from situations with extremely limited data, to contexts with sufficiently reliable data.

The development of model life tables (Coale and Demeny, 1966; UN, 1955, 1982) and relational models (Brass and Coale, 1968), for example, enabled mortality estimation for the entire age distribution using information for just one or two age groups. Clark and Sharrow, (2011) and Wilmoth et al., (2012) are more recent propositions along similar lines. These approaches allow mortality estimation for all ages using only infant mortality information calculated indirectly from censuses and surveys (Brass, 1971; Hill, 2013). Similarly, the same procedure can be adopted if there exist an estimation of adult mortality from, for instance, questions about siblings alive and death (Hill and Trussell, 1977; Timæus, 2013a). These

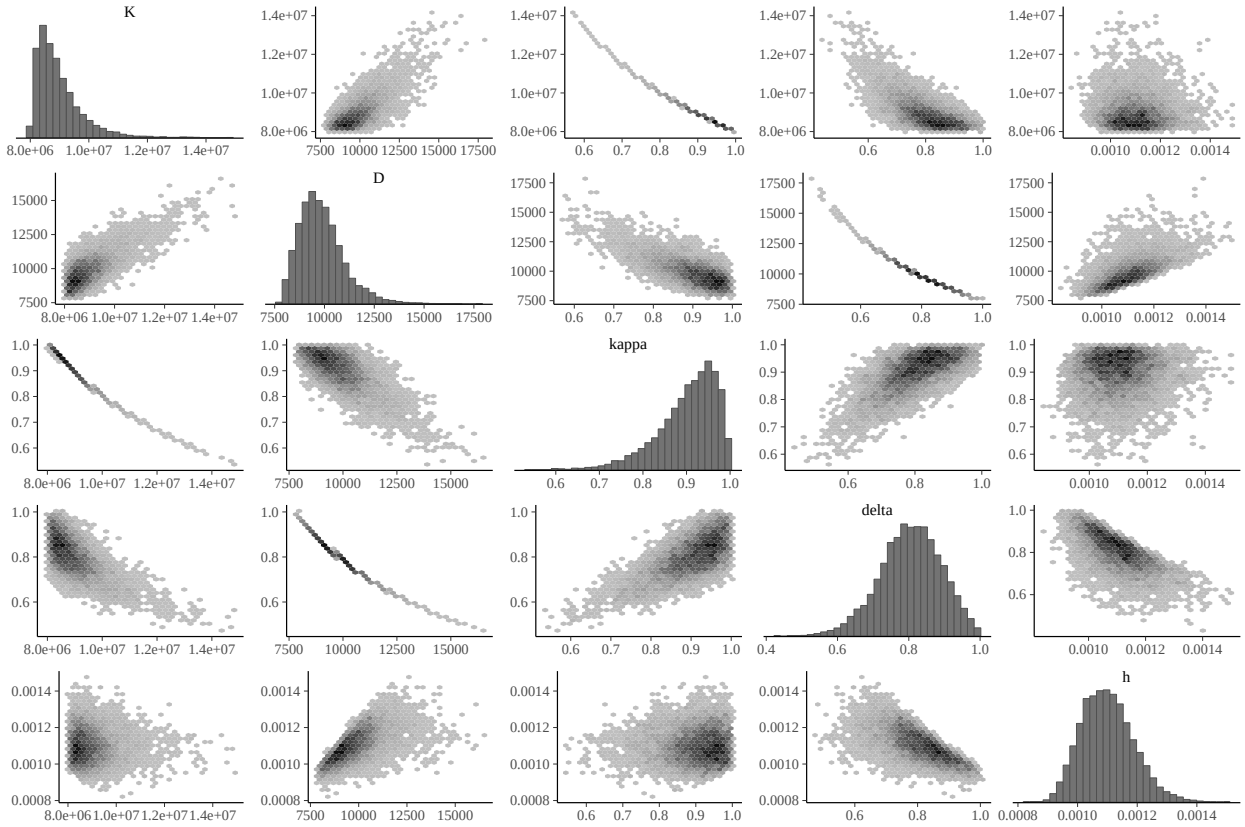

Figure 3.2: Illustration of the posterior predictive distributions for the parameters in the mortality model. Population aged 30-34. Brazil, 2010. The lower panels show the scatterplots of half of the HMC chains and the upper panel the other half.

approaches have been extensively used for mortality estimation in contexts where no reliable vital registration systems exist.

With the development of Civil Registration and Vital Statistics (CRVS) systems, several methods that use registered deaths, albeit incomplete, have been proposed. These techniques are preferred because they use information of all ages from the region being analyzed and make no assumptions about the age pattern of mortality (Bennett and Horiuchi, 1981; Brass, 1975; Hill, 1987; Preston and Hill, 1980). Dorrington, (2013) highlights, however, that "if completeness appears to be less than 60 per cent then the uncertainty is large and this should be taken into account when interpreting the results." Given the nature of these methods, low estimates of coverage means that the uncertainty is high, even if results from different methods and age trims are similar. Furthermore, the results of low completeness may also be biased, since one of the main assumptions of the method, that deaths are equally underreported at each age, is likely to be violated. This occurs because completeness of registered deaths is associated to low quality information reported, such as age.

These two sets of mortality estimates are useful and may be used jointly, since each one has its pros and cons. Model life tables have the advantage of being smooth across ages and represent the entire mortality age structure with a limited number of parameters. Furthermore, they require no data from vital registration systems. Their main limitation is that these models may miss special features of the mortality age schedule that are different from the models. On the other hand, mortality estimates calculated from death distribution methods capture these specific characteristics, but are often more noisy and require reasonably reliable vital registration systems.

The next section proposes a framework to combine the two sets of estimates: from model life tables and from death distribution methods.

TOPALS relational model

The regularities in mortality rates have led to the development of functions that capture the effect of age in mortality, called "mortality laws". The best known early contribution of a mathematical expression for the graduation of the age pattern of mortality is probably that of Gompertz, (1825), who proposed a model in which mortality rates increase exponentially with age, which seems to fit well mortality for older ages, starting around age 40. Some authors have also proposed models for the entire age range, such as Siler, (1979, 1983) and Heligman and Pollard, (1980). To avoid the rigidity of these models, non-parametric smoothing approaches have been proposed (Camarda, 2012; de Beer, 2012).

One of these approaches is the TOPALS (tool for projecting age-specific rates using linear splines) relational model. TOPALS relates the probabilities of dying of a given region to a standard mortality schedule by using splines and was first proposed by de Beer, (2012) to smooth and project probabilities of dying.

Splines are functions often used to smooth a set of irregular observations. They use piecewise polynomials to interpolate between two points, called "knots". This procedure guarantees that the smooth function is continuous. The spline function will have as many different polynomials as the number of intervals between knots. The simplest type of piecewise polynomials are linear splines, which interpolate linearly between the knots.

TOPALS uses linear spline to model the risk ratio function, r_c, that relates a vector of observed age-specific mortality rates, h_c, of a given region with the standard mortality schedule, h_c^{std}, as follows:

$$h_c = h_c^{std} \cdot r_c \tag{3.22}$$

The risk ratio r_c is a piecewise linear function connected at ages chosen as "knots". The interpretation of r_c is straightforward: it simply indicates how different the mortality rates to be smoothed are from the standard schedule for each successive age group. When $r_c > 1$, the standard scheduled is adjusted upwards, whereas when $r_c < 1$ the observed mortality rates are lower than h_c^{std}. Although modeled in its simplest form as a first-degree spline, TOPALS is flexible enough to provide smooth mortality rates by age (de Beer, 2012).

de Beer, (2012) assumes that r_c is constant for ages below 20 and uses ten-year intervals between knots for the other age groups. The parameters of the splines functions are then estimated in a way that the values of the spline equal the observed rates at the knots. He further uses these estimates to project mortality rates modeling the path of the risk ratios towards a "best-practice" mortality in the future.

Gonzaga and Schmertmann, (2016) propose a variant of TOPALS that models the logarithms of the age-specific mortality rates, leading to an additive term that relates the standard schedule to the mortality rates being modeled. The authors also use B-splines for constructing the linear spline functions. B-splines is a way to express splines curves in terms of certain basis functions, which is detailed below. The model they proposed are as follows:

$$log(h_c) = log(h_c^{std}) + \boldsymbol{b'_c} \cdot \boldsymbol{\Phi} \tag{3.23}$$

where h_c is the set of age-specific mortality rates to be modeled, h_c^{std} is the standard age schedule and $(\boldsymbol{b'_c} \cdot \boldsymbol{\Phi})$ is the piecewise linear function that connects both schedules, with seven knots at ages 0,1,10,20,40,70 and 100. The knots are the places where the segments join to make the function piecewise continuous. In this case, the knots are sensible choices that reflect changes in the underlying mechanisms of mortality. The offset function $\boldsymbol{\Phi}$ is a vector of seven parameters to be estimated and represent the differences between h_c and h_c^{std} at the knots. The higher the value of $\boldsymbol{\Phi}$, the more different are the two mortality schedules. These values are equivalent to the risk ratio r_c at the knots in the de Beer, (2012) formulation. The B-spline constant $\boldsymbol{b'_c}$ is also a vector with seven fixed values of the basis spline that gives the piecewise linear interpolation between two knots for each age c. The pair $(\boldsymbol{b'_c} \cdot \boldsymbol{\Phi})$ at age c is essentially a weighted average between two consecutive values of $\boldsymbol{\Phi}$, weighted by the proximity to the knots.

Gonzaga and Schmertmann, (2016) estimate $\boldsymbol{\Phi}$ iteratively in order to maximize its likelihood. The authors also include a penalty term on the second-order difference of adjacent B-splines coefficients to avoid irregularities for very small populations, but they argue that this term has almost no effect for moderate to large populations.

Schmertmann and Gonzaga, (2018) adapt this approach to a Bayesian framework and use a weak prior on $\boldsymbol{\Phi}$ ($\boldsymbol{\Phi} \sim Normal(0, 4)$) to stabilize the estimates in small populations. Likewise to the penalty term described above, the difference between two adjacent values of $\boldsymbol{\Phi}$ is also modeled. They also add prior information to take into account the probability of a death being registered, which is similar to the approach used in this dissertation discussed previously.

This formulation of the TOPALS model is used in this dissertation to combine and model the two sets of age-specific mortality rates described above (3.4). The use of TOPALS has two main goals in this context. First, it smooths mortality information across ages that might be irregular due to sample variation and data quality problems, such as age misstatement. Another strength of this approach is the possibility to combine two independent mortality schedules: the unsmooth death rates and the standard schedule. In this dissertation, the the

standard schedule is obtained from the application of relational model / model life tables using partial data.

This dissertation adopts a similar approach to that used by Schmertmann and Gonzaga, (2018), with a few modifications. First, instead of using a single standard age schedule, several standards are used, one per region/year. Each schedule should reflect, in a certain way, the mortality rates of the region/year being modeled, which is more informative than using a general standard only for smoothing purposes. Second, a similar approach taken by de Beer, (2012) to project mortality rates is adopted, but instead of modeling the path to achieve a "best-practice" mortality in the future, what is model is the distance to the standard mortality schedule. This is done by using indicators of the reliability of the calculated mortality rates. Finally, since this dissertation only deals with large populations, there is no need to include penalty terms, and the TOPALS offsets are only model by $\mathbf{\Phi} \sim Normal(0, 4)$. These procedures will be detailed in the next sections.

Calculating mortality standard schedules

The underlying mortality rates h_c^{std} used as the the standard schedules for a given region is defined as a function of the estimated child mortality rate ${}_5q_0$ using a model life table $h_c^{std} = f({}_5q_0)$.

Thus, the complete mortality age schedule can be calculated by using relational models (Brass and Coale, 1968; Wilmoth et al., 2012) or model life tables (Clark and Sharrow, 2011; Coale and Demeny, 1966), given information about infant mortality. This study uses Clark and Sharrow, (2011) model life tables, because it gives a broader range of estimates based on the experience of all populations in the Human Mortality Database, which could also be interpreted as credible intervals. The inputs to calculate the standard mortality schedule are the child mortality rates (${}_5q_0$) calculated using the Brass, (1971) method based on information of children ever born and children still alive contained in censuses and surveys. This procedure could be extended to include information about adult mortality that is independent from the vital statistics, such as the results from the application of the sibling survival method.

Combining mortality schedules via TOPALS

TOPALS offers a useful tool for pooling two independent mortality schedules as described previously. Mortality rates h_c are modeled by using the underlying mortality rates h_c^{std}. The method proposed in this dissertation adds a new term, ω, to equation 3.23, which accounts for possible bias in the calculated mortality rates. The choice of this term is related to the choice of the pooling weight in the Bayesian melding approach. This weights reflect the reliability of the information, rather than precision.

$$log(h_c) = log(h_c^{std}) + \omega \cdot \boldsymbol{b'_c} \cdot \mathbf{\Phi} \tag{3.24}$$

The intuition behind equation 3.24 is that less reliable mortality estimates would make estimations from model life tables to have a higher weight. When ω is one, the model relies solely on h_c and mortality from model life tables is mostly used to smooth the age schedule. In this case, equation 3.24 becomes the traditional TOPALS model as presented in Schmertmann and Gonzaga, (2018). In the opposite extreme case, if ω equals zero, mortality rates being modeled (h_c) will be equal to the standard distribution.

The weighting term can be modeled as a parameter in the usual Bayesian framework by a beta distribution: $\omega \sim Beta(a^\omega, b^\omega)$. The parameters a^ω and b^ω are estimated as usual, based on expert opinion about the reliability of h_c.

Alternatively, following a similar approach used by G. L. Oliveira, Loschi, and Assunção, (2017), ω can be a function of some adequacy indicators about registered deaths. Szwarcwald, Leal, et al., (2002) propose the use of some indicators, such as age-standardized mortality rates and annual deviation to the three-years average. Values of mortality rates lower than 4 indicate strong coverage problems. Equally, by assuming that there should not exist strong variations in mortality from one year to another, annual variation greater than 10% is considered critical. Another indicator the authors suggest is the proportion of ill-defined deaths. Indicators of age heaping, such as the Myer's index may be also used. The specifics of this procedures to estimate ω for Brazilian data will be detailed in chapter 4.

Model summary

Table 3.3 and Figure 3.3 show a summary of the mortality method setup. The registered death counts, D_c^{obs}, are modeled based on likelihood and prior information about the completeness of registered deaths (δ_c^*), in addition to model the age structure of mortality rates h_c.

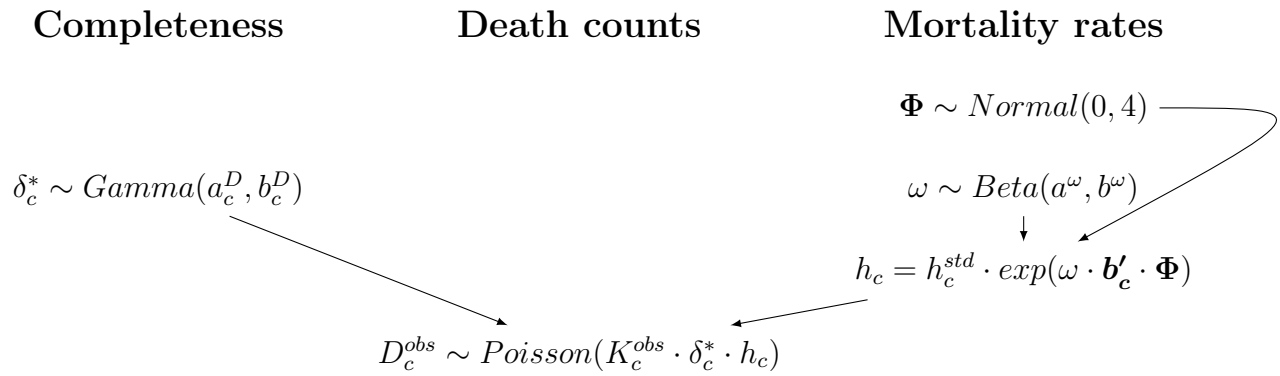

Figure 3.3: Diagram of the the relationship between priors and mortality models

Variable	Distribution	Description	Data source / Estimation strategy	Type	Section
D_c^{obs}	$\sim Pois(K_c^{obs} \cdot h_c \cdot \delta_c^*)$	observed death counts	CRVS	data	3.4, A.3 4.4, 5
h_c	$= h_c^{std} \cdot exp(\omega \cdot \boldsymbol{b_c'} \cdot \boldsymbol{\Phi})$	mortality rates	TOPALS model	parameter	3.4
$\boldsymbol{\Phi}$	$\sim Normal(0, 4)$	TOPALS offset	weak prior	parameter	3.4
δ_c^*	$\sim Gamma(a_c^D, b_c^D)$	relative completeness of registered deaths	method of moments	parameter	3.4 A.3
a_c^D, b_c^D	–	shape and rate parameters	DDM	hyper-parameter	3.4 A.3
ω_c	$\sim Beta(a^\omega, b^\omega)$	weight	method of moments	parameter	3.4, A.2 4.4, 5
a_c^ω, b_c^ω	–	shape and rate parameters	expert opinion	hyper-parameter	3.4, A.2 4.4, 5

Table 3.3: Mortality Model Summary

3.5 Modeling fertility and birth counts

The strategy adopted to model fertility and birth counts is very similar to that used to model mortality. As death counts, the total number of births by age of the mother is assumed to follow a Poisson distribution (Brillinger, 1986). Thus, the total number of births B_c women from cohort c have (K_c) is Poisson distributed as follows:

$$B_c \sim Poisson(K_c \cdot f_c) \tag{3.25}$$

where f_c is Age-Specific Fertility Rate (ASFR) for cohort c.

Underregistration of births has to be considered as well. Although less prone to under-reporting than deaths, more than 30% of the births globally are still not registered, and the improvements have been only modest (Mikkelsen et al., 2015). To take this into account, the number of registered births, B_c^{obs}, is modeled by a binomial distribution, following the same strategy used to model population:

$$B_c^{obs} \sim Binomial(B_c, \beta_c) \tag{3.26}$$

where β_c is the probability of a birth being reported.

The model for birth counts also includes uncertainty in the denominator of fertility rates, the population K_c. Furthermore, similarly to what was discussed in section 3.4, registered births may be modeled by using the actual completeness β_c, with exposure K_c (equation

5.1) or the completeness of registered births relative to the census coverage β_c^* and exposure K_c^{obs} (equation 3.28).

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{3.27}$$

$$B_c^{obs} \sim Poisson(K_c^{obs} \cdot f_c \cdot \beta_c^*) \tag{3.28}$$

Whereas in mortality the latter approach seems to have more practical use, in fertility both models may be useful, depending on the source of the estimate of completeness of birth counts. There are a few methods that estimate birth completeness relative to census coverage ((UN, 1983, chapter 2), (Moultrie and Zaba, 2013)), but it is also common to have independent measures of birth completeness, for example from capture and recapture methods.

As shown in equation 3.15 in section 3.4, mortality rates (h_c) are often unknown, and it is estimated by the number of observed deaths and their estimated completeness. When modeling birth counts, fertility rates f_c may be also modeled in the same way, but it is much more common to have independent estimates of f_c than h_c, given the number of information collected in surveys and censuses about fertility ((UN, 1983, chapter 2), (Moultrie, 2013c)). Conversely, data about completeness of registered births are relatively rare compared to those available for completeness of deaths, since the methods to estimate the latter are more developed than those used to estimate the former.

A Bayesian approach is a natural choice to deal with this kind of problem, since identification problems also arise from equations (5.1) and (3.28). The advantage of this Bayesian setup is that it is flexible enough to allow fertility estimation based on independent information of either fertility rates, completeness of registered births, or the combination of both.

This study uses the formulation in 5.1, which leads to the following hierarchical structure:

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{3.29}$$

$$f_c \sim Gamma(a_c^B, a_c^B) \tag{3.30}$$

$$\beta_c \sim Beta(a_c^B, b_c^B) \tag{3.31}$$

Notice that K_c is modeled as discussed in 3.3.

To illustrate this model, consider four independent estimates of the age-specific fertility rates f_c for Brazil in 2010 that lead to TFR of 1.76, 1.78, 1.87 and 1.93 (*include Ref*). Two other independent studies measured the completeness of the registered births in the same year and found 95.9% and 97.3% (*include Ref*). Based on this information, the methods of moments and percentiles are used to approximate distributions that represent the prior information about the TFR and β_c (see sections A.2 and A.3 in the Appendix A). The complete hierarchical structure for this example is as follows:

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{3.32}$$

$$f_c = \frac{TFR \cdot p_c}{5} \tag{3.33}$$

$$TFR \sim Gamma(534, 291) \tag{3.34}$$

$$\beta_c \sim Beta(299, 10.4) \tag{3.35}$$

where p_c is the proportion of the TFR for each age group, which in this simplified example is assumed to be constant for the four different estimates. The birth completeness β_c is also assumed to be constant across age groups. The population K_c is estimated by the model described in section 3.3. In this example, κ_c is assumed to follow a beta distribution so that $\kappa \sim Beta(144, 2.4)$ for all age groups, which has a median of 0.9858.

In a more realistic model, other information, such as infant mortality and undercount of children in the census, would be included in order to take into account the complex relationship that represents the entire dynamics of this population. This is done in this paper when all sub-models are combined (Section 3.7).

Figure 3.4 shows, in the upper panel, the distributions of the TFR and the completeness of registered births and, in the lower panel, the scatterplot of the 20000 HMC draws for both parameters combined. The ellipse approximates the 80% prediction interval for the posterior distributions and provides a range of plausible values for β_c and TFR. The posterior distribution of the TFR is highly concentrated below the prior mean(1.84), close to the minimum value of the prior (1.76). This indicates that the distribution of the TFR was shifted markedly to the left given the likelihood of population and births counts, in addition to their respective coverage priors.

Model summary

Table 3.4 and Figure 3.5 show the summary of the fertility model. The registered birth counts, B_c^{obs}, are modeled based on likelihood and prior information about coverage of census (κ_c), completeness of registered births (β_c) and fertility rates f_c. If there is no prior information about either f_c or β_c, non-informative prior can be chosen. Notice that this setup can be easily adapted to the context where information about the relative coverage of births β_c^* is available, as shown in equation 3.28.

3.6 Modeling migration

Immigration and emigration (outmigration) are modeled independently, since data sources that provide information about these two components tend to be different. Due to the lack or incompleteness of administrative data, migration estimates often rely on information from censuses and surveys, for example from questions about the place of residence five years before the census date.

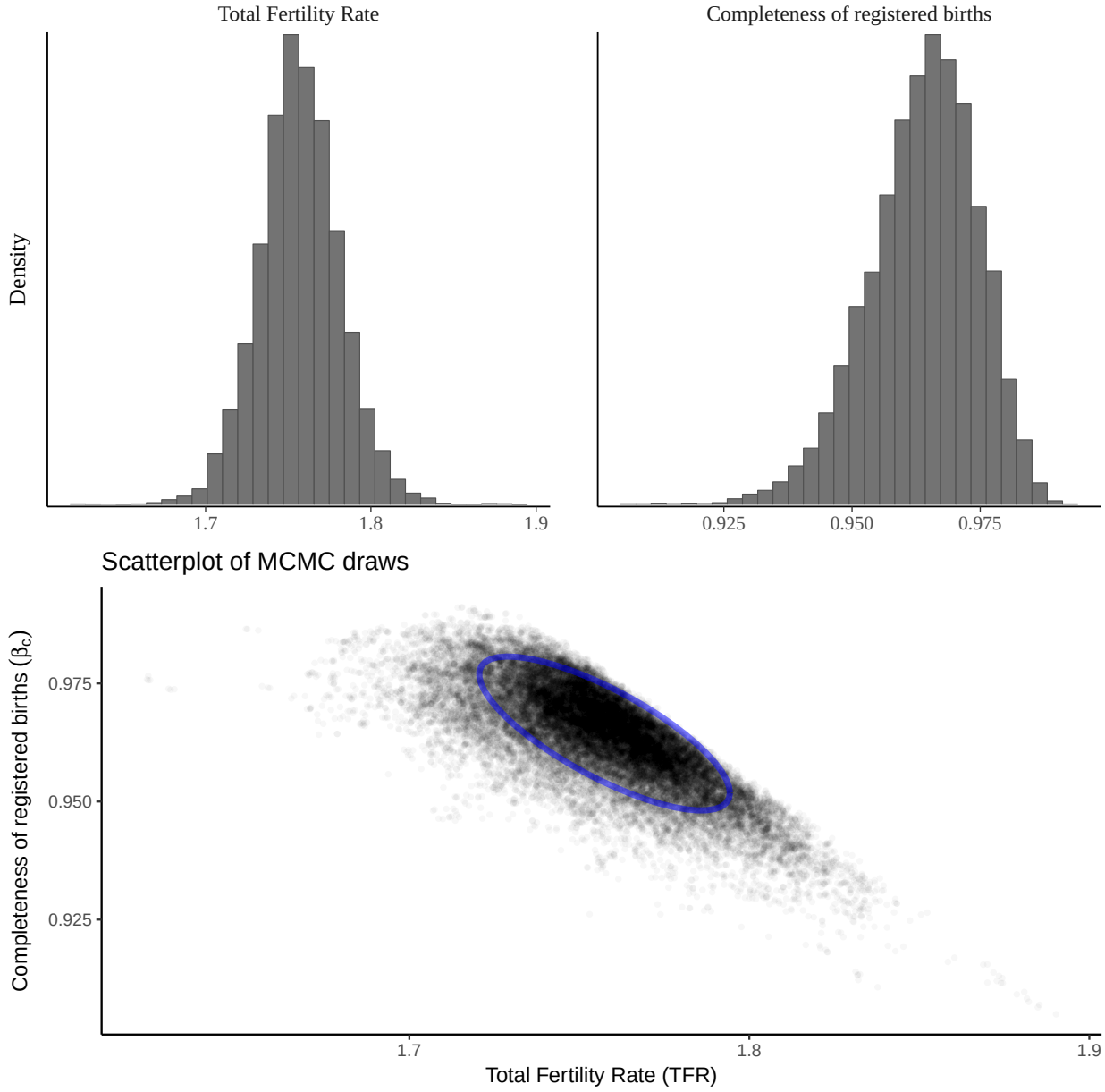

Figure 3.4: Posterior predictive distribution of the TFR and completeness of registered births, and ellipse that approximates the 80% posterior predictive intervals. Brazil, 2010

Variable	Distribution	Description	Data source Estimation strategy	Type	Section
B_c^{obs}	$\sim Pois(K_c \cdot \beta_c \cdot f_c)$	observed birth counts	CRVS	data	3.5 4.3, A.3
β_c	$\sim Beta(a_c^B, b_c^B)$	completeness of registered births	method of moments	parameter	3.4 A.3
a_c^B, b_c^B	–	shape parameters	capture-recapture	hyper-parameter	3.5 A.2
f_c	$\sim Gamma(a_c^f, b_c^f)$	fertility rates	method of moments	parameter	3.5 4.3
a_c^f, b_c^f	–	shape and rate parameters	Indirect techniques	hyper-parameter	3.5 A.2

CRVS: Civil registration and vital statistics

Table 3.4: Fertility model summary

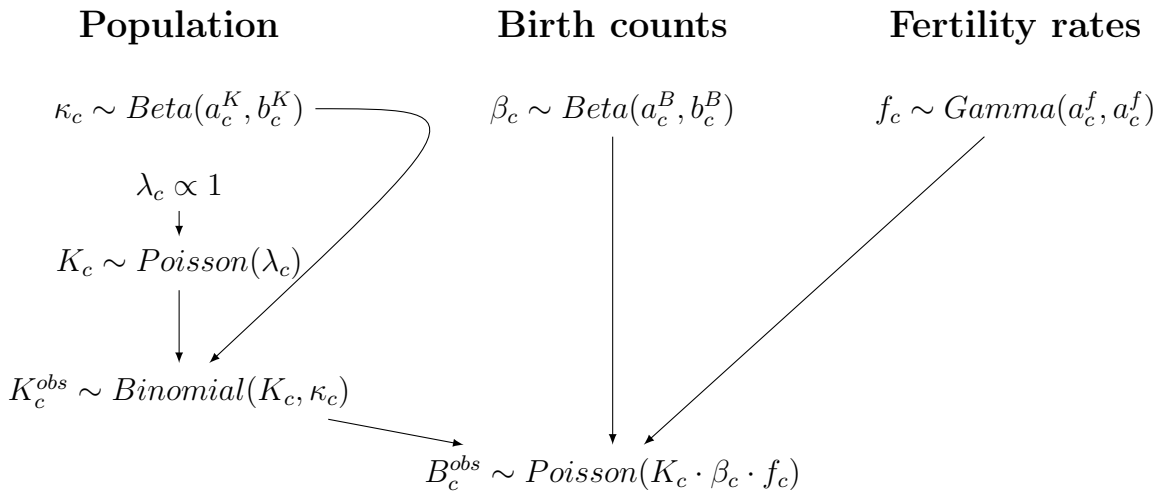

Figure 3.5: Diagram of the the relationship between priors and fertility models

A slightly different strategy is adopted to model migration, since information about the number of immigrants (I_c) and emigrants (E_c) are more difficult to obtain than births and deaths. Thus, immigration i_c and emigration rates e_c are modeled instead of counts. These rates are modeled by the following gamma distributions with parameters a_c^i and b_c^i:

$$i_c \sim Gamma(a_c^i, b_c^i) \tag{3.36}$$

$$e_c \sim Gamma(a_c^e, b_c^e) \tag{3.37}$$

Data to estimate migration come from censuses, which provides point estimates for i_c^{obs} and e_c^{obs} from the the question about place of residence five years prior to the census reference day. Uncertainty is incorporated in these estimates through the definition of the hyperparameters. The hyperparameters $a_c^i, b_c^i, a_c^e, b_c^e$ can be easily estimated by the methods of moments or quantiles given expert opinion about plausible values for the variability of these estimates. When migration data are collected in the census long-form questionnaire, sampling error can be incorporated in the same way. If no information about uncertainty in the estimates are available, weak prior distributions for i_c and e_c are:

$$i_c \sim Gamma(1, \frac{1}{i_c^{obs}}) \tag{3.38}$$

$$e_c \sim Gamma(1, \frac{1}{e_c^{obs}}) \tag{3.39}$$

Model summary

Figure 3.6 shows a diagram of the migration model. Both immigration and emigration are modeled based the estimated migration rates and their respective uncertainty, in addition to the information about population. The absolute number of migrants depends both on the model for population and migration rates. This will be important when modeling different regions which might have different census coverage levels.

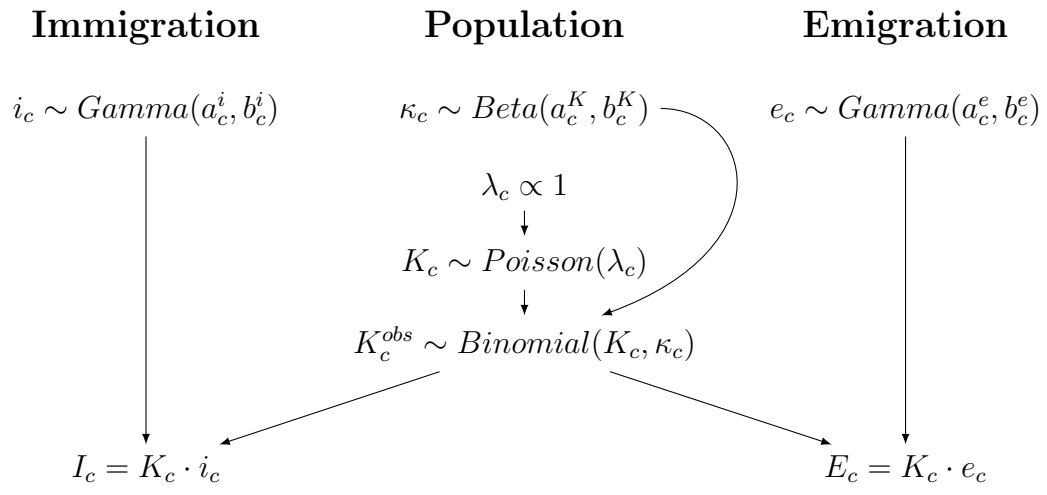

Figure 3.6: Diagram of the the relationship between priors and migration models

3.7 Simulating the posterior distribution

As previously mentioned, inference about the parameters of interest in this paper is done in two steps. First, samples from the posterior distributions of the population ($K_c(t)$ and $K_c(t+n)$), deaths ($D_c(t,t+n)$), births ($B_c(t,t+n)$) and migration ($I_c(t,t+n)$ and $E_c(t,t+n)$) are drawn. These are called the premodel posterior distributions. Samples from the posterior distributions of the parameters were draw via a Markov Chain Monte Carlo (MCMC) algorithm using the statistical softwares *R* and *Stan*. More specifically, Stan uses the HMC sampling, a form of MCMC, to explore the target distribution. The HMC algorithm tends to explore the posterior distribution in a more efficient way than the traditional MCMC. Efficiency in this context means that it requires fewer samples to describe the posterior distributions. HMC gains efficiency by reducing randomness when moving through the parameter space and exploiting knowledge of the target distribution. A practical advantage of the HMC algorithm is that, unlike Gibbs sampling and the Metropolis algorithm, it makes easier to identify problems and divergences when sampling from the posterior (Carpenter et al., 2017; McElreath, 2016; Team, 2017).

Secondly, the left and right-hand sides of the demographic balancing equation were reconciled. In other words, the premodel posterior distribution for the population at time $t+n$, $K_c(t+n)$, estimated based solely on the $t+n$ census and its census coverage information, is harmonized with the probability distribution induced by the population at time t, $K_c(t)$, and the demographic events during the period $[t; t+n]$. This is done by using the Sampling/Importance Resampling (SIR) algorithm.

The posterior distributions of this model are then approximated by drawing samples according to the following steps:

1. Draw a sample of j values of the inputs ($\boldsymbol{AK(t)}$) and outputs ($\boldsymbol{K(t+n)}$) from their prior distributions. Notice that the Leslie matrix $\boldsymbol{A}$ contains information about fertility, mortality and migration.
2. Calculate the unnormalized premodel posterior distributions of population at time t and demographic events between t and $t+n$, $q_1(\theta)$, which is formed of the sample of size j: $(\theta_1, ..., \theta_j)$.
3. Calculate the unnormalized premodel posterior distributions at time $t+n$, $q_2(\phi)$, which is formed of the sample of size j: $(\phi_1, ..., \phi_j)$.
4. Determine the induced sample on the output $\phi_i^* = M(\theta_i)$ for each of the j values generated in *Step 2* by running the model $\boldsymbol{K^*(t+n) = AK(t)}$.
5. Use nonparametric Kernel Density Estimation (KDE) to obtain estimates of the distributions of both populations at time $t+n$ to be harmonized: the induced distribution of ϕ, $q_1^*(\phi)$, and the premodel posterior distributions of ϕ, $q_2(\phi)$.
6. Form the importance resampling weights to be used for melding the two posteriors on $K(t+n)$ using logarithmic pooling with pooling weight α:

$$w_i = \left(\frac{q_2(M(\theta_i))}{q_1^*(M(\theta_i))} \right)^{1-\alpha} \tag{3.40}$$

7. Sample k values from the discrete premodel posterior distributions samples in *Step 2* with values θ_i, but proportional to the importance sampling weights w_i, calculated in *Step 6*.

3.8 Simulation Study

This section presents a simulation study conducted to examine the properties of the posterior distributions for the parameters of interest, for instance whether the known parameters could be estimated by the model with no bias and relatively narrow credible intervals. This is done by verifying that the outputs of the model can recover the ground truth from simulated data.

The fist step to conduct a simulation study is to select reasonable values for the parameters, simulating data according to the model, and then trying to fit it with the simulated data. Then, a consistency check verifies if the 90% posterior interval captures the "true" parameter values.

Data

The simulation study of this paper uses data of the female population of Sweden for the period 1/1/1975-1/1/1980. Demographic data in Sweden is considered of very high quality (Glei, Lundström, and Wilmoth, 2017) and are used here as the true estimates. Table 3.5 shows the population by ages groups on January 1st of 1975 and 1980, the number of deaths and net migration and the number of births from the cohorts aged $[x; x + n]$ in 1980. For example, the female population aged 20-24 in 1980, born between 1955 and 1959, had 40650 children during the period 1975-1979. From the 261759 women aged 15-19 in 1980 581 died. The net migration for this cohort was 11868. The life expectancy of the female population in Sweden during this period was about 78 years and the TFR was 1.7.

Simulation steps

Bayesian inference is performed by sampling from the posterior using the procedures described in section 3.7. A total of 25 samples of simulated observed data were generated for each variable. Random samples of observed counts were drawn based on the true counts (Table 3.5) and the distribution of the factors that accounts for census coverage and completeness of registered deaths and births, as follows:

Age (x)	Population (1975)	Population (1980)	Deaths	Net Migration	Female Births
0	269260	239689	1780	4515	0
5	285926	272980	370	4090	0
10	269421	287591	287	1952	3
15	261759	272977	395	3951	2888
20	278929	273046	581	11868	40650
25	326842	286680	647	8398	84395
30	282541	328990	936	3084	74195
35	229829	282894	1124	1477	28096
40	220390	229538	1490	1199	6021
45	236295	219174	2156	940	681
50	269493	233455	3559	719	25
55	244729	263853	6165	525	0
60	248069	236621	8374	266	0
65	228076	234706	13558	195	0
70	187425	206972	21318	214	0
75	135736	156844	30754	173	0
80	83347	97402	38448	114	0
85	37863	46694	36653	0	0
90	11025	14566	23297	0	0
95	1791	2649	8376	0	0
100	146	207	1584	0	0
105	4	7	139	0	0

Table 3.5: Population in 01/01/1975 and 01/01/1980; deaths, net migration and female births from 1975 to 1979. Death, births and migration refer to the events observed for the cohort aged [x; x+5] in 1980. Births are tabulated by age of the mother.

$$K_c^{obs}(1975) \sim Poisson(K_c(1975) \cdot \kappa_c(1975)) \tag{3.41}$$

$$K_c^{obs}(1980) \sim Poisson(K_c(1980) \cdot \kappa_c(1980)) \tag{3.42}$$

$$D_c^{obs} \sim Poisson(K_c^{obs} \cdot h_c \cdot \delta_c^*) \tag{3.43}$$

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{3.44}$$

$$i_c \sim Gamma(10, \frac{10}{i_c^{obs}}) \tag{3.45}$$

$$e_c \sim Gamma(10, \frac{10}{e_c^{obs}}) \tag{3.46}$$

$$\kappa_c(1975) \sim Beta(10.8, 0.65) \tag{3.47}$$

$$\kappa_c(1980) \sim Beta(110.4, 9.9) \tag{3.48}$$

$$\delta_c^* \sim Gamma(162, 180) \tag{3.49}$$

$$\beta_c \sim Beta(10.8, 0.65) \tag{3.50}$$

A series of 5000 HMC sampling were drawn for each of the 25 datasets. Figures 3.7 and 3.8 show the true values and the 80% Bayesian confidence interval of the 25 replications in simulation study, showing that, in general, the known parameters can be estimated by the model with no bias and relatively narrow credible intervals.

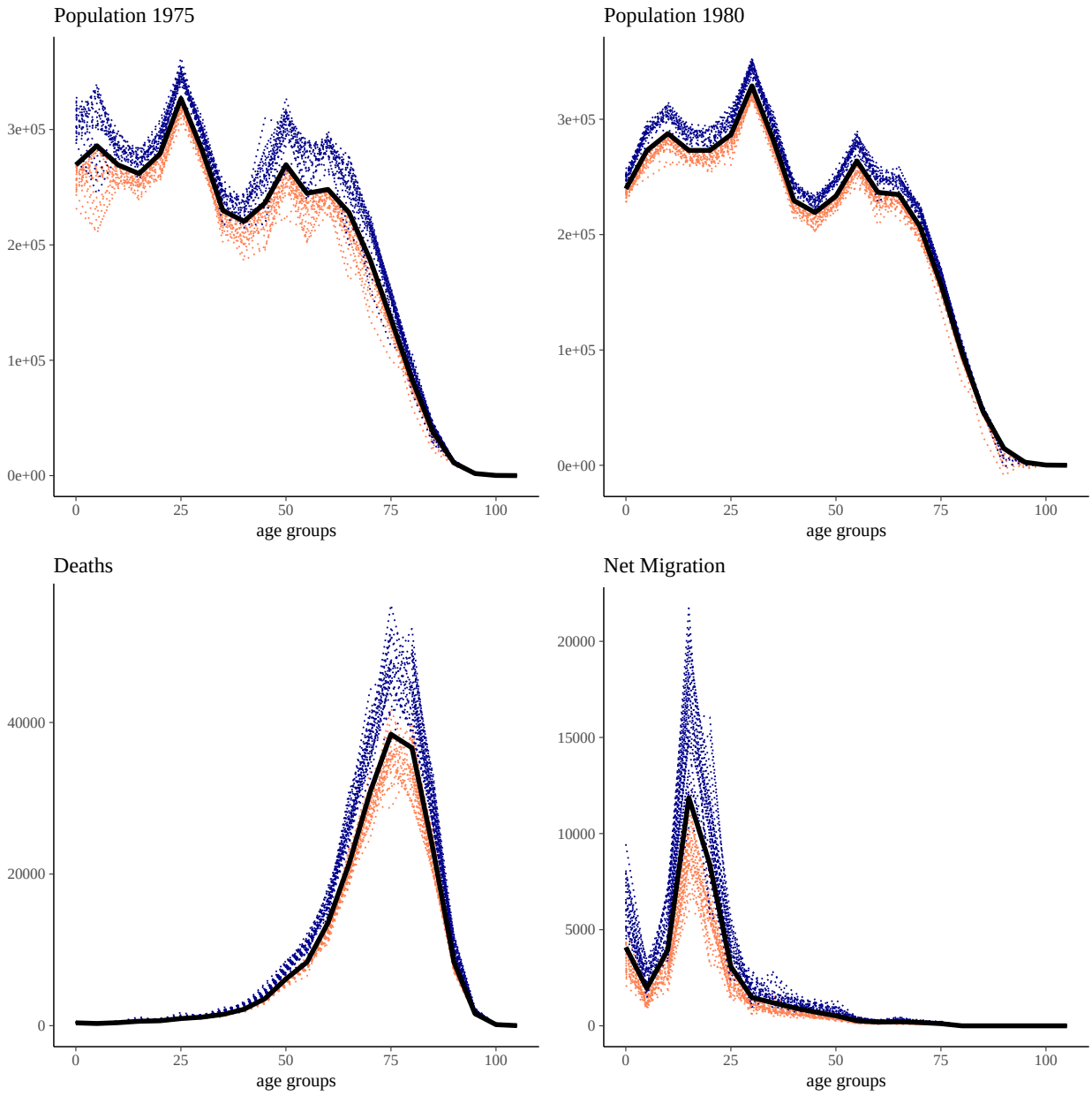

Figure 3.7: Population in 1975 and 1980, death and net migration counts. Blue and orange dots represent the 80% Bayesian confidence interval of the 25 replications and the black solid line represent the true values. Sweden, 1975-1980

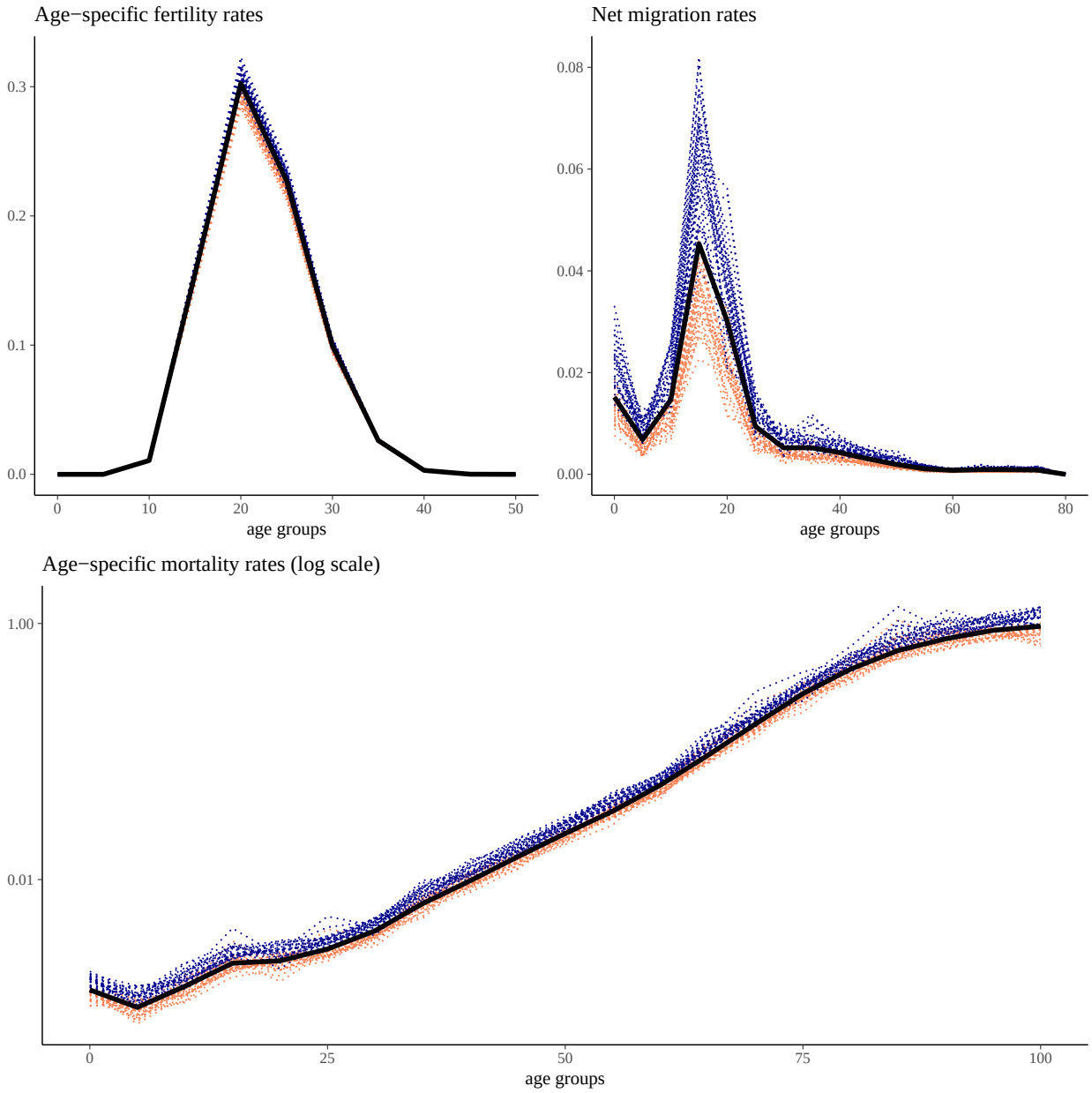

Figure 3.8: Age-specific fertility rates, net migration and probabilities of dying between ages x and x+5 for the period 1975-1979. Blue and orange dots represent the 80% Bayesian confidence interval of the 25 replications and the black solid line represents the true values. Sweden, 1975-1980

3.9 Summary

This section has described a set of methods to estimate and reconcile past demographic data including measures of uncertainty. This framework is an integrated population model that estimates all demographic components simultaneously. It is flexible enough to be applicable to different contexts, with varying data availability and quality. The model was able to recover the ground truth from simulated data for Sweden and the next sections will show an application to Brazilian data.

Chapter 4

Demographic Estimates for Brazil and States from 1980 to 2010

4.1 Introduction

This chapter presents the data to be used for producing demographic estimates for Brazil and states between 1980 and 2010.

Brazil is a large and heterogeneous country. It is composed by five regions (North, Northeast, Southeast, South and Midwest) and 27 states[1], as shown by the map in Figure 4.1.

The population size of Brazilian states vary from less than one million in states in the North region, such as Acre (AC), Roraima (RR) and Amapá (AP), to more than 40 million people in São Paulo (SP). The three most populous states are in the Southeast region: SP, Minas Gerais (MG) and RJ.

There are great regional and socioeconomic inequalities in the country. The regions North and Northeast are poorer and the regions South and Southeast are richer than the average, although there is also internal variability within each region. The Human Development Index (HDI) in the least developed states, such as Maranhão (MA) and Alagoas (AL) is as low as that in South Africa, whereas the index for the more developed states, such as the Distrito Federal (DF), is close to that of Portugal.

As the next sections will discuss, this socioeconomic inequality reflects on the quality of the vital statistics, which also varies considerably across states. Census coverage is also different by state, but it seems to have no correlation with socioeconomic development.

There is considerable amount of demographic data in Brazil for the period under analysis, which come mainly from censuses and administrative records. However, these data have several data quality issues, which should be taken into consideration when producing

[1] In fact, there are 26 states and the Federal District (Brasília), which make up the 27 Federative Units ("Unidades da Federação").

Figure 4.1: Map of Brazil, regions and states

population estimates. There has also been important development of techniques to assess and adjust imperfect data.

As discussed in Chapter 3, the main objective of this study is to resolve inconsistencies between consecutive censuses and intercensal demographic data. To illustrate the types of inconsistencies found in Brazilian data, Figure 4.2 compares the female population by age in the state of RJ enumerated in the 2010 Census with the population estimated for 2010 based on the 2000 Census and registered births and deaths between 2000 and 2010, in addition to migration estimated based on the census question about the place of residence five years

before the censuses [2].

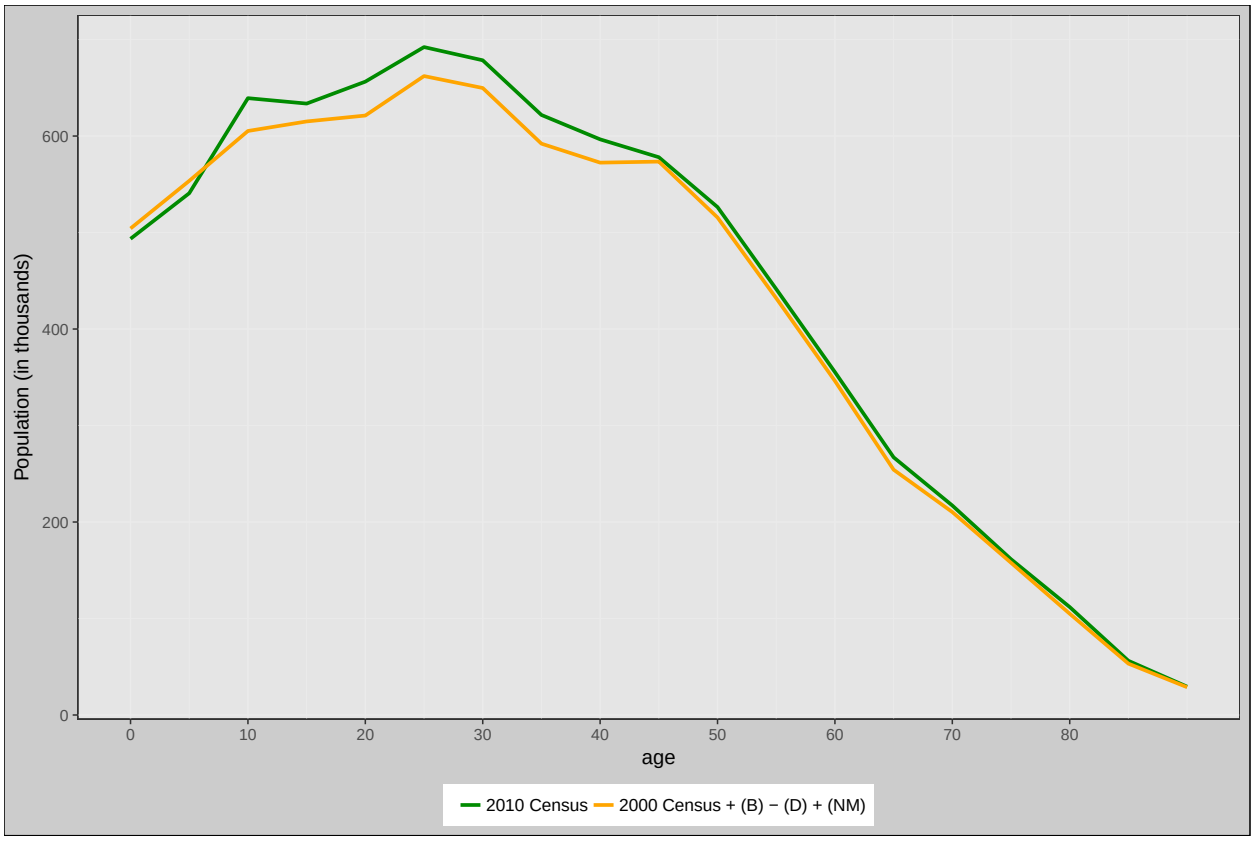

Figure 4.2: Female Population by age group from the 2010 Census compared with population projected based on the 2000 Census and intercensal demographic estimates of fertility, mortality and migration). Source: 2000 and 2010 Censuses, SIM and CR

Figure 4.2 shows that there are inconsistencies in the data for all ages. The projected population is greater than the enumerated for the two age groups below 10, probably indicating undercount of children in the 2010 Census. For the adult population, the enumerated population is greater than the projected. This may be due to errors in the migration estimates or a greater census undercount of adults in 2000 than in 2010. Differences at old ages may be explained by age misreporting.

The next sections discuss the quality of the data and the estimation procedures for the censuses (4.2), fertility (4.3), mortality (4.4) and migration 4.5.

[2]Net migration for the period 2000-2005 was estimated by logarithmic interpolation between the estimates of the periods 1995-2000 and 2005-2010

4.2 Evaluation of censuses in Brazil and states from 1980 to 2010

Censuses are the main data source for demographic analysis, particularly in contexts with incomplete vital statistics, such as in Brazil. Conducting a census is a complex task and even with all the effort spent to guarantee precise information, coverage and quality problems are inevitable in activities of this magnitude.

Nevertheless, census figures are still valuable if data limitations are understood by users and if the errors do not adversely affect the major uses of the data. In this sense, evaluation studies which examine the results, procedures and operations used in conducting a census are necessary for providing both producers and users with the information needed to assess census quality (US Bureau of the Census, 1985).

The institution responsible for the census, often the NSO, should be committed to transparency regarding the quality of the statistics produced, which comprises providing information necessary to evaluate the accuracy of census figures. According to the UNSD, (2008, p. 84), "the purpose of census evaluation is to provide users with a level of confidence when utilizing the data, and to explain errors in the census result. It is therefore important to choose an appropriate way of sending out these messages to the right group of people."

Evaluation of censuses can be used in several ways, e.g., guiding improvements in future censuses and surveys; assisting users in their interpretation of the results; adjusting the census results (US Bureau of the Census, 1985). The evaluation presented in this study serves all of these purposes, particularly for adjusting the census results to produce consistent demographic estimates [3].

In addition to producing more precise population estimates, census evaluation can help understanding patterns of census errors, which may bias mortality, fertility and other indicators that use census data. Just to cite one example, there has been considerable debate about whether some unexpected mortality patterns, such as low mortality at old ages for some population subgroups, are real or just an artifact of poor data quality, for instance age misstatement in the census (Preston and Elo, 1999).

Census errors are often divided into coverage and content errors. Coverage errors are related to failure in counting persons or housing units, leading to missing or duplicated cases. Content errors refer to mistakes in the information about persons or housing units effectively enumerated, such as age misstatement (US Bureau of the Census, 1985).

[3] Adjusting census figures is a controversial issue and is also related to the uses of censuses. Even though adjusting final census figures is often extensively discussed, this is not a common practice. The most symptomatic case of census adjustment is that of the United Kingdom (UK), where the last two censuses (2001 and 2011) were adjusted through a fully integrated coverage measurement processes, resulting in the development of the One Number Census (ONC) methodology (Abbott, 2009). In the United States (US), where census figures are highly political, census figures have never been adjusted. A more common practice is to use evaluation to adjust censuses in order to derive the base population for projections. In Latin America, this procedure has been used, for instance, in Argentina, Brazil, Colombia, Costa Rica, Chile, Ecuador, Mexico, Paraguay, Peru.

The main objective of this section is to provide an evaluation of the coverage and content errors [4] of the four Brazilian censuses from 1980 to 2010 (1980, 1991, 2000 and 2010) by age, sex and state.

Before discussing these specific issues, it is important to describe the most used methods for census evaluation and provide a brief history of Brazilian censuses and a description of their main characteristics.

Methods for evaluation of censuses

There are several techniques used for census evaluation, but two different sets of methods have emerged: direct and indirect. Direct methods are represented primarily by the procedure DSE, which is often based on a PES, in which a sample of households is revisited after the census, some data are collected again and then compared to those collected by the census in the same areas. These are also called micro-level approaches. A broader set of methods refers to performing census evaluation through DA, which are also called macro-level approaches. These consist of evaluating data using internal consistency within the same census and/or the application of demographic techniques using administrative records for deriving population estimates, used later to compare to the census results (Bryan and Heuser, 2004; US Bureau of the Census, 1985).

Dual System Estimator

The DSE method uses the same idea of capture-recapture techniques, and involves two samples: the "P-sample" and the "E-sample". The "P-sample" is a sample from the PES, independent from the census, which contains information collected from people interviewed after the census in competed in that area. This sample is further used for comparison with census records. The "E-sample" is a sample drawn from the already collected data in the census in the same area of the "P-sample". Once the two samples are collected, they are matched up against each other (Freedman and Wachter, 2001; UNSD, 2010).

Matching is what guarantees the statistical power of the DSE, not counting better (Freedman and Wachter, 2001). In fact, if more efficient enumeration is conducted in the PES than in the census, results may be distorted under the independence assumption (Chatterjee and Mukherjee, 2015).

In practice, matching is done in several phases. First, the most obvious cases are matched using information collected in both surveys, such as census block, name, sex, date of birth. After the preliminary match, it is common to carry out field reconciliation to obtain additional information to help resolve suspicious cases that remain unmatched after the initial matching phases. Such visits give an opportunity to identify erroneous census enumerations and to resolve doubtful cases in order to achieve a realistic and definitive match status for every P and E sample element, deciding whether the error should be charged against the

[4] Content errors can refer to any information collected in the census. This study focus on content errors regarding the reports about age.

census or the PES. Cases that remain unresolved are handled by statistical models that fill in the missing data (Freedman and Wachter, 2001; UNSD, 2010).

The DSE procedure is represented by the following two-by-two table (Table 4.1):.

Table 4.1: Illustration of the DSE procedure

	In the census	Out of the census
In the PES	$\hat{a}$	$\hat{b}$
Out of the PES	$\hat{c}$	$\hat{d}$

where $\hat{a}$ is an estimate of the number of people counted in both the PES and the census; $\hat{b}$ is an estimate of the number of people counted in the PES, but missed from the census; $\hat{c}$ is an estimate of the number of people counted in the census, but missed from the PES; and $\hat{d}$ is an estimate of the number of people missed by both the PES and the census.

The estimator of the total population ($\hat{t}$) under certain assumptions, such as independence between the census and the PES, and setting aside duplications and erroneous inclusions in the census, is given by:

$$\hat{t} = \frac{(\hat{a} + \hat{c})(\hat{a} + \hat{b})}{\hat{a}} \tag{4.1}$$

The estimator of the census coverage rate (or match rate), $\hat{\kappa}$, under the same assumptions, is given by:

$$\hat{\kappa} = \frac{\hat{a}}{\hat{a} + \hat{b}} \tag{4.2}$$

The assumption of no duplications and erroneous enumeration is unrealistic and, in practice, the census population in the Table 4.1 should exclude these people. Erroneous enumeration includes cases such as fabrication, out-of-scope and geographic misallocations. These cases are found through the procedures of matching and field reconciliation described above.

There are well developed statistical methods to take into account random sampling errors in the PES. Sampling errors are relatively easy to control and quantify and depend mostly on the sample size. The main problem of the PES are uncertainties due to systematic non-sampling errors. Non-sampling errors are harder to identify and measure, and may arise from many sources, such as processing error, correlation bias, and heterogeneity (Freedman and Wachter, 2001; Wachter and Freedman, 2000).

Processing errors often appear in the matching and field reconciliation phases. Despite advances in techniques and technology used for matching, inaccuracies in the information given to the PES and/or to the census, make it very problematic. The few cases that remain

unresolved, even if relatively small, can have a considerable influence on the final results. Furthermore, cases may be "resolved" incorrectly.

Correlation bias results from failures in the independence assumption between the census and the PES and refers to the tendency for people missed in the census to be missed by PES as well (Wachter and Freedman, 2000). This can result from both factors related to the census design and behavioral characteristics of the enumerated and missed population. Despite attempts to carry out a completely independent survey from the census, it is nearly impossible to do so, since both are normally carried out by the same institution, which uses the same list of households, maps and sometimes staff. As for the differences in the characteristics of the enumerated and non-enumerated populations, persons who do not wish, for different reasons, to respond to the census, will probably refuse to respond to the PES as well. Hard-to-count population groups will be probably missing from both surveys as well.

Heterogeneity refers to differences in undercount rates within demography across geography. This issue is particularly important in the debate about adjusting the census figures, since the adjustment requires population groups, at some level, be treated as homogeneous (Freedman and Wachter, 2001).

Given the predominance of non-sampling errors, Freedman and Wachter, (2003) suggest that large PES sample sizes not only increase the relative importance of non-sampling errors, but also make them more problematic, since bigger samples are harder to manage so that systematic errors are made more difficult to control and measure.

Demographic Analysis

DA is a broader set of techniques used for census evaluation. There are several indicators that can be calculated to perform internal consistency checks using one or more censuses and to visually identify improbable results. More refined demographic techniques, using mainly administrative records, are available for detecting and identifying the source of error in census data. These techniques tend to be more independent from the census being evaluated (Bryan and Heuser, 2004; Moultrie, 2013a; UNSD, 2010). However, they require data that are not always available, and are more useful to contexts of high quality administrative data.

In Latin America, the procedures used to perform DA are often used in the context of producing population estimates and projections. The Latin American and Caribbean Demographic Center (CELADE), which is responsible for producing population projections for all Latin American countries, and several NSO have performed DA through a process called "Demographic Reconciliation". This procedure is a tool that aims to obtain adjusted structures by age and sex from the population enumerated in the census using the knowledge about the parameters that represent the country's demographic dynamics and then producing the most plausible demographic dynamics for the country. The functions of census error by sex and age are obtained as a by-product of this procedure (Chackiel, 2009). The essence of this technique in terms of estimating census coverage is favoring some age groups over others, e.g., the population aged 5-9 at time t is thought to be more reliable than the population

aged 15-19 at time $t + 10$. The main limitation of this procedure is that it is time-intensive and depends on many subjective decisions about the population groups to be privileged.

One advantage of DA over the PES is that results can be produced at very low cost, since it often uses data that have already been produced. Another potential advantage of DA is that it can be performed in a timely basis, producing a quick evaluation of census results (Bryan and Heuser, 2004), although more careful evaluation may take longer and depend on the publication of more recent administrative data. Methodologically, even though there is no recipe, there are well developed demographic techniques and accumulated knowledge on the expected patterns of demographic rates and census errors.

The main advantage of the PES over the DA is that it can provide estimates of census coverage by population groups that otherwise would be difficult to assess, such as subnational geographic areas. Furthermore, particularly in developing countries, such as the populations studied in this dissertation, the PES is the only independent source of census evaluation, since DA depends largely on consecutive censuses and tend to measure census undercount relative to previous censuses. Demographic rates, such as fertility, mortality and migration are also often estimated using direct and indirect methods from data collected in the censuses themselves, which could potentially create some circularity in the estimation procedures.

Census coverage and age misstatement

Coverage problems such as census undercount and age misstatement are not random and affect population groups differentially.

Differentials in coverage and quality by age and sex are one of the main sources of bias in censuses. Children, for example, are known to be one of the most undercounted population group. Surprisingly, this phenomenon has occurred in all sorts of population censuses, regardless the census design and cultural or socioeconomic characteristics of the country. Many Asian countries (Gerland, 2014; Goodkind, 2011), South Africa (SSA, 2012), the US (O'Hare, 2015), the UK (ONS, 2012) and several Latin American countries (Chackiel, 2009) have reported higher census undercount among children than the overall population.

There are several factors that may be related to this differential omission, but there is no consensus about the main causes of the problem. The leading explanations can be divided into two broad categories: i) children living in households effectively counted in the census are omitted by the respondents; ii) children are disproportionately present in households more likely to be omitted.

Among the first set of explanations is the idea that children might not be viewed as a person who should be counted and their existence may not be considered by the respondents (Chackiel, 2009). Children undercount can also be related to specific cultural characteristics of the country. In China, for instance, fertility policies provide a powerful incentive for both parents and officials to underreport out-of-quota births as well as children (Goodkind, 2011). Respondents could also think that children not registered in the civil registration systems should not be included in the census (Mortara, 1941), which would be a more problematic issue in countries with less developed vital registration systems.

Another potential explanation is related to the fact that several characteristics of the households are likely to be associated to a larger number of children, especially due to differential fertility in accordance with these characteristics. If fertility is higher in regions where enumeration is more difficult, or in other "hard-to-count" population groups, children tend to be disproportionally omitted (O'Hare, 2015; West and Robinson, 1999). Other evidences from earlier censuses in the United States show that undercount for infants is closely tied to the undercount at other ages, since large part of the infants who were not reported in the census were the children of adults who were also not listed in the census. This suggests the hypothesis that the under-enumeration of persons aged 15-19 and of children aged 0-4 might be caused by the omission of newly formed small households or families (Ewbank, 1981).

Another characteristic that affects population distribution by sex and age is the differential undercount by sex, especially among young adults, in which males tend to have higher undercount rates than females. This differential coverage may be related to family and work relationships, in which adult men have greater activity rate in the labor market than women, in addition to a greater representation in single-person households. This group is typically more mobile than the general population and more likely to form new households (Ewbank, 1981; Lee, 1982).

In addition to coverage errors, there are several problems that normally affect information about the age statement of people in a census. Reporting of age is known to suffer from several types of non-random measurement error, which will be discussed below.

Firstly, it should be mentioned that it is common that only one person provide information for all household members, resulting in less precise information.

An important problem that affects the distribution of the population by age is the preference for ages with terminal digits, for instance 0 and 5. This kind of content error may be related to operational problems with the census, such as the design of the questionnaire and training of the enumerators, but can be also relative to factors exogenous to the census, such as cultural features and the educational level of the population enumerated, which influence the knowledge and perception of the population about their age.

Age misstatement can also occur when people tend to report their ages deliberately towards one direction, primarily due to cultural factors. This type of age misstatement is often sex-specific, e.g. higher age understatement among young women than men; age overstatement at older ages, particularly among males, in cultures where status is gained with age. (Booth and Gerland, 2016)

In fact, the tendency of the elderly to declare themselves older than they really are is observed in many populations. This tendency tends to increase with with age. The pyramidal shape of the population age structure at old ages itself can exacerbate this problem, by converting random age heaping to net overcount among these age groups (Coale and Caselli, 1990; Del Popolo, 2000; Preston and Elo, 1999; Romero and Freitez, 2008).

Brief history and main characteristics of Brazilian censuses

The first census conducted in Brazil was in 1872. Decennial censuses have been conducted since 1940 in years ending in zero, except for the 1990 enumeration, which was postponed to 1991.

The IBGE, the Brazilian NSO, is the institution responsible for carrying out the population censuses in Brazil since its foundation. IBGE's headquarter is in Rio de Janeiro, but every state has an agency that is responsible for the operational implementation of the census, with a certain autonomy to coordinate the processes of the preparation and conduction of the census in their states.

Brazilian censuses have been conducted through face-to-face interviews, in which enumerators are expected visit personally all households to carry out the interviews.

The census of 1980, the first census under analysis in this dissertation, collected information of both the resident population (*de jure*) and the present population (*de facto*). This dissertation uses the results for the resident population, which is consistent with the following censuses and with the vital statistics. After the 1991 Census, only resident population was collected.

The 1991 Census was postponed from 1990, due to political problems with the central government when hiring enumerators and other census staff, and was carried out in adverse conditions.

The 2000 Census was marked by technological innovations, with an online system that allowed monitoring data collection. Furthermore, the questionnaires were, for the first time, scanned and the processes of codification, checks and tabulation were automatized.

The 2010 Census was the first to use internet as an option for respondents to fill out the questionnaire. However, only a very small proportion of the interviews were responded using this alternative. Another innovation of the 2010 Census was the use of portable electronic devices, in which enumerators used electronic questionnaires on handheld computers to conduct the interviews.

Since 1960, two different questionnaires are used: the "long form" and the "short form". The long form contains more detailed questions and these interviews are only conducted in a sample of the population. The sample size and sampling design have changed over time. In the 2010 Census, sample size varied by municipality population size, from 50% for small municipalities to 5% for large municipalities. The long form of the Brazilian censuses has provided relevant demographic information, for instance detailed questions on migration, fertility and infant mortality. The short form asks basic questions such as age and sex. There have been changes in the questions that are placed in the long and short forms. The question about deaths in the household, for example, was in the long form in the 1980 Census and in the short form in 2010 (IBGE, 2013a).

In addition to decennial censuses, intercensal population counts were conducted in 1996 and 2007. An intercensal count had been planned for the year 2015, but was postponed and then canceled due to budgetary restrictions. The 1996 Count had similar characteristics to a census, aiming to count the entire population, but with shorter questionnaire than the regular

censuses. The 2007 Count was conducted only in a selected number of municipalities, mainly the less populous ones. The population counts have had lower coverage than the decennial censuses (IBGE, 2008).

Census coverage in Brazil and states from 1980 to 2010

This section discusses the differential census coverage by age, sex, and geographic area in Brazilian censuses from 1980 to 2010.

IBGE has conducted PES to evaluate census results since the 1970 Census [5]. PES could be used potentially to measure both coverage and quality errors. The PES of the 1980 Census did an exhaustive evaluation of the quality of the information provided by the respondents, but since the 1991 Census, the PES in Brazil have focused on coverage errors solely.

Even though the results of the PES of the Census from 1970 to 2000 have been published in working documents by IBGE, their results have been rarely used by either demographers or the own IBGE in planning census activities. One of the few studies that make use of the PES in Brazil is that of Carvalho and Campos, (2006), which adjusts the census results for all ages by using the overall census coverage of the 1991 and 2000 censuses to estimate international migration in the intercensal period [6]. No study has examined the results of the Brazilian PES by state and population groups.

The remarkably limited use of the PES may be related to several political and technical issues, but it also suggests certain skepticism about their results. This study aims to explore the results of the PES published by IBGE since the 1980 Census, based on the principle that only the careful use of the PES data in combination with other demographic information can point out the plausibility of their estimates and indicate to what extent they can be useful. This evaluation would also provide useful information about the limitations and advantages of the PES, which can help improving surveys of future censuses.

The results of the PES should be interpreted carefully, as the surveys have had different territorial coverage, and used slightly different methodology and sample size. The PES of the 1991 and 2000 Censuses, for instance, did not include rural areas of the North Region.

IBGE's interpretation of the PES results is consistent with the literature in that the Achilles heel of the PES are non-sampling errors. The 2000 PES report recognizes that the results are influenced by all the technical and operational procedures adopted and implemented in all the stages of the survey. Many problems may have occurred in all different phases, many of which are difficult to measure or rectify. Therefore, the main focus in the interpretation of the results should be on the general patterns of omission, serving as a reference for a reflection about the Census and the PES itself (L. C. S. Oliveira, de Freitas, et al., 2003).

[5] The PES of the 2010 Census was conducted and its methodology is available (A. D. d. Silva, Freitas, and Pessoa, 2015), but the results have not yet been published.

[6] See Section 4.5 for a brief discussion about this procedure.

This section explores the results of the PES, which, combined with standard demographic methods, can shed some light on the overall quality of the last Brazilian censuses by age, sex and geography.

As discussed before, the PES has not been used to adjust census results in Brazil. The official measures of census undercount are derived from DA, and are simply given by the implicit difference between the census and the estimated population. The results reported in each revision of the population estimates and projection have changed with the incorporation of new information.

Table 4.2 shows the estimated undercount of the censuses from 1980 to 2010 according to both methods normally used to evaluate census coverage: the PES of the corresponding census and the implicit census undercount of the most recent DA study (IBGE, 2018).

The indicators of census coverage presented in the PES reports of different censuses are not always consistent. One indicator that has been presented in the reports of all censuses is the omission of people living in private households missed by the census but found by the PES, which is shown in Table 4.2. It is worth noting that the results of the PES shown in Table 4.2 about undercount of people living in private households omitted in the censuses fails to capture the entire features of census coverage, missing, for example, omissions of persons living in households enumerated in both the PES and the census and duplications or erroneous inclusions. Despite this limitation, these can be used for comparative purposes.

Table 4.2: Census undercount according to the PES and DA, Brazil, 1980-2010 (in %)

Year	Method	
	PES	DA
1980	4.3	2.8
1991	4.7	3.1
2000	5.8	2.4
2010	-	2.2

Source: IBGE: 1980-2000 PES; IBGE, (2018)

The results of the two methods are inconsistent, as it is the case for several other countries (Chackiel, 2009). The census undercount given by the PES is consistently higher than that resulted from the DA procedure. The temporal trends are also different, with an increase in the undercount estimated by the PES, whereas the DA indicates that the worst census in terms of coverage was the 1991 Census (3.1%), followed by the 1980 Census (2.8%). There is no published result for the 2010 Census PES, but DA indicates that it had similar coverage to that of the 2000 Census.

As previously discussed, the national average hides differentials by geography and population group.

One of the results published in the PES reports is the census undercount by urban-rural classification. In Brazil, each municipality is in charge of defining their urban and rural

areas. Nearly all municipalities have both urban and rural areas. IBGE then defines the census blocks in the each of these areas. The PES results by this disaggregation level has consistently shown higher undercount for rural areas (Table 4.3). The census undercount for rural areas in 1980 was more than twice as high as that for urban areas. For 1991 and 2000, it was almost 70% higher, and it would had been greater had rural areas in the North Region been included in the sample for these years. These differentials are possibly related to the greater logistical difficulties to map and reach rural areas.

Table 4.3: Census undercount according to the PES by urban-rural classification, Brazil, 1980-2010 (in %)

Year	Method	
	Urban	Rural
1980	3.0	7.1
1991	4.1	6.8
2000	5.2	8.8

Source: IBGE: PES of 1980, 1991 and 2000 censuses

Figure 4.3 shows the map of the undercount of the censuses 1980, 1991 and 2000 by state, based on the results of the PES for these censuses. These results should be interpreted with caution, due to the above-mentioned general limitations of a PES and the differences in operational characteristics, methodology, sample size and geographic coverage of the PES for different years. For the states of the North Region in 1980 and 1991, Figure 4.3 shows the average of the region rather the estimates for every state, which is not available from the PES publications.

As previously discussed, the PES is a complex survey and have had difficulties in all of their phases. The report of PES of the 1991 Census (L. C. S. Oliveira, Indá, et al., 1996), for instance, states that the survey went to the field without training due to lack of funding. Also due to lack of financial resources, initial matching was accepted without return to the field to check the divergencies, contrarily to what had been initially planned. The report of 2000 PES (L. C. S. Oliveira, de Freitas, et al., 2003) also reveals difficulties in the matching and field reconciliation phases.

PES in Brazil have shown methodological improvements over time, towards a more independent survey and a better sample design. Recent PES have also used better methods of for matching and field reconciliation, taking advantage of advances in computing and technology. Since the difficulties and limitations of the PES are complex, it is not clear how these improvements may have influenced the final results.

Despite the limitations, interesting results arise from Figure 4.3. The three states of the South Region (RS, Santa Catarina (SC) and Paraná (PR)), one of the most developed states in the country, show consistently low undercount in the three censuses under analysis, always below 4.0%. DF also had relatively low census undercount in all censuses, as well

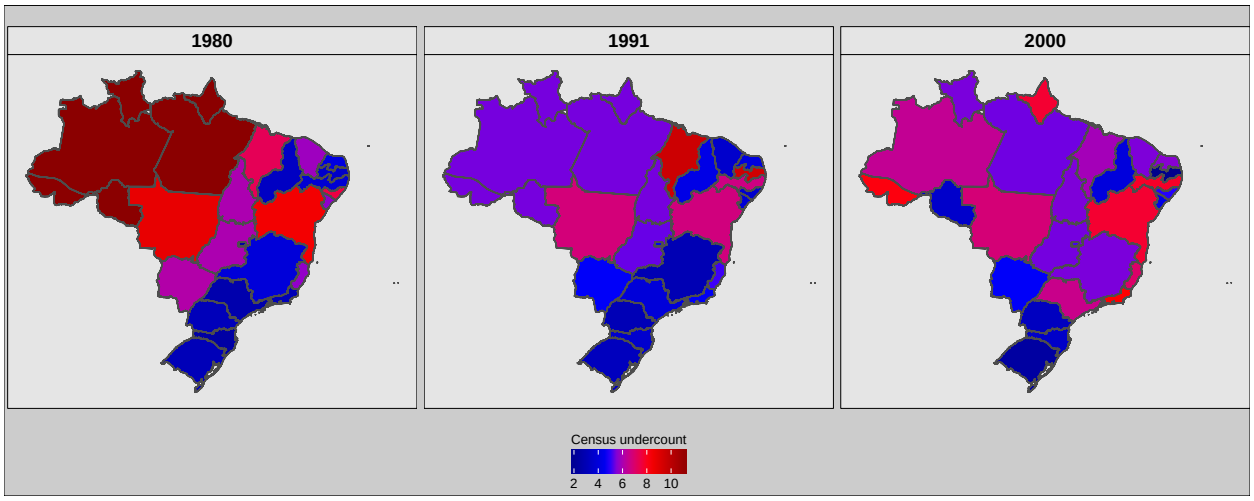

Figure 4.3: Map of the census undercount by state, 1980, 1991, 2000 (in %). Source: PES of 1980, 1991 and 2000 censuses

as MG, despite a moderate increase in the 2000 Census, reaching 5.4%. The states of RJ, Espírito Santo (ES), and to a lesser extent SP, with undercount rates of 7.0%, 8.2%, 6.5%, respectively, are examples of developed states with high census undercount in the 2000 Census.

The states of the North Region, a less developed region, have had high census undercount. Results for this region in 1980 and 1991 were calculate only for the region as a whole. The map shows the regional average for all states. The 2000 PES shows that, in fact, this region is far from homogeneous in terms of census coverage. It had states with both high undercount, such as AC (8.1%) and AP (7.7%) and low undercount, such as Rondônia (RO) (3.6%).

The relationship between the socioeconomic levels of the states and census undercount is nonetheless unclear. Several states of the Northeast Region, also extremely poor, have had low census undercount in all the censuses under analysis. The case of Piauí (PI) is noticeable, as the state has been one of the poorest in the country and managed to have low undercount in all three censuses. Another poor state from the Northeast region that had low undercount rate in the 1980 and 2000 censuses is PB. The report of the 2000 PES recognizes the high quality of the census work in this state in the collection, supervision stages, as well as in the different phases of the survey. Bahia (BA) is a state in the Northeast with consistently high census undercount.

Another type of omission calculated by the PES refers to the omission within the private households enumerated in both the census and the PES, which is independent from the previous one. These rates have been calculated only for the 1991 and 2000 censuses.

Results show that although omission rates of households enumerated in the PES and missed by the census 4.2 is higher in 2000 than in 1991, there was an improvement in the omission rate in households enumerated in both surveys between these two years, from 4.0%

in 1991 to 2.6% in 2000. The overall undercount rate, considering both people living in households missing from the census and people missed in households counted in the census, reduced slightly from 8.3% in 1991 to 7.9% in 2000.

Figure 4.4 shows the map of the overall undercount, which includes both people living in households missing from the census and people missing from households counted in both surveys, for the 1991 and 2000 censuses by state. In general, the spatial pattern of the overall undercount (Figure 4.4) is similar to that of people living in households missed by the census (Figure 4.3). The main difference is in the state of PI in 1991, which shows a high overall undercount, due to high omission of people living in households counted by both the census and the PES.

The overall census undercount reduced for most states from 1991 to 2000. Among the few states that had an increase are those in the Southeast Region. The state of SP went from from 5.6% in 1991 to 8.0% in 2000, RJ from 7.4% to 10.4% and ES from 7.9% to 9.2% (Figure 4.4).

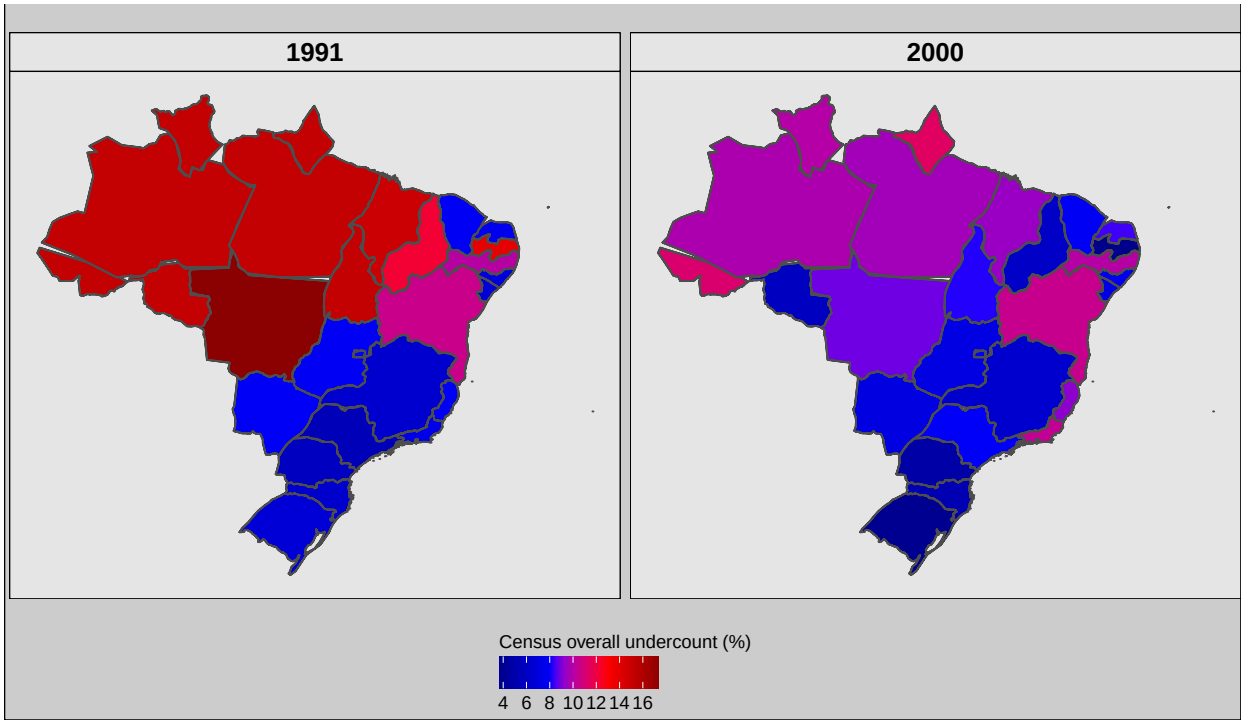

Figure 4.4: Map of the census overall undercount by state, 1980, 1991, 2000 (in %). Source: PES of 1980, 1991 and 2000 censuses

The PES of the 2000 Census was the first one to calculate the erroneous inclusion rates so that it is possible to calculate gross census coverage error and net census coverage error. Gross error refers to the total number of errors made in the census, while net error refers to the net effect of these errors on the resultant census coverage measure. Gross error is the sum

of the erroneous inclusion and omission rates, while the net error is the difference between omissions and duplications. When omissions exceed duplications, it is said that there is net census undercount (UNSD, 2010; US Bureau of the Census, 1985).

The erroneous inclusion rate for the 2000 Census in Brazil was 2.34%, which in combination with overall omission rate of 7.87%, lead to a net omission rate of 5.52% and gross omission rate of 10.21%.

Maps of the same indicators for states are shown in Figure 4.5. There is a weak correlation between overall undercount rate and erroneous inclusion rate. The three Southern states (RS, SC, PR), PB and RO had low rates for both indicators, whereas states like AC, Amazonas (AM), MA, PB, RJ, Mato Grosso (MT) had both high undercount and high erroneous inclusion. These result in low gross undercount for the first group of states and high gross undercount rate for the second group.

The map of the net undercount shows that the rate is always positive, indicating that net undercount is greater than erroneous inclusions for all states. The group of states with low gross undercount (RS, SC, PR, PB and RO) also had low net undercount. Other states that had low net undercount are those with relatively low census undercount, but significant erroneous inclusions (PI, MG, Goiás (GO)).

As noted above, the results of the 2010 PES have not been published, which hampers the ability to perform a direct evaluation of this census, although indirect evaluation through DA showed above indicates that both the 2000 and the 2010 censuses had similar coverage (Table 4.2).

In the absence of the results for the PES, a new procedure adopted in the 2010 Censuses may be used to assess the overall quality of this census by state. The housing units surveyed in the Brazilian censuses are classified according to the status of their residents in the day of the interview. A housing unit is considered occupied if a person or group of persons are living in it at the time of the interview or if the occupants are only temporarily absent. The interview cannot be carried out in all occupied housing units for several reasons, e.g. if occupants are absent during the census or they refuse to respond to the interview. In those cases, IBGE performed a count imputation in the 2010 Census to fill in the housing unit status and the main characteristics of their residents (IBGE, 2013a)

This imputation corresponded to only 1.5% of the population in the 2010 Census, but varied between 0.3% and 3.2% between states, as shown by the map in Figure 4.6. Even though this is not a precise measure of census undercount, this can be used a a proxy of the quality of the census in different states.

The three Southern states (RS, SC, PR), which have shown low census undercount since 1980 (Figure 4.3), also had low proportion of count imputation in the 2010 Census. RJ, ES, DF and to a lesser extent SP are developed states that had high proportion of count imputation. These states also had high census undercount in the 2000 Census (Figure 4.5). The states of the Northeast region, except MA and Pernambuco (PE), had relatively low count imputation. The states in the regions North and Midwest had a more heterogeneous pattern, with several states with high imputation and states with low proportion of imputation.

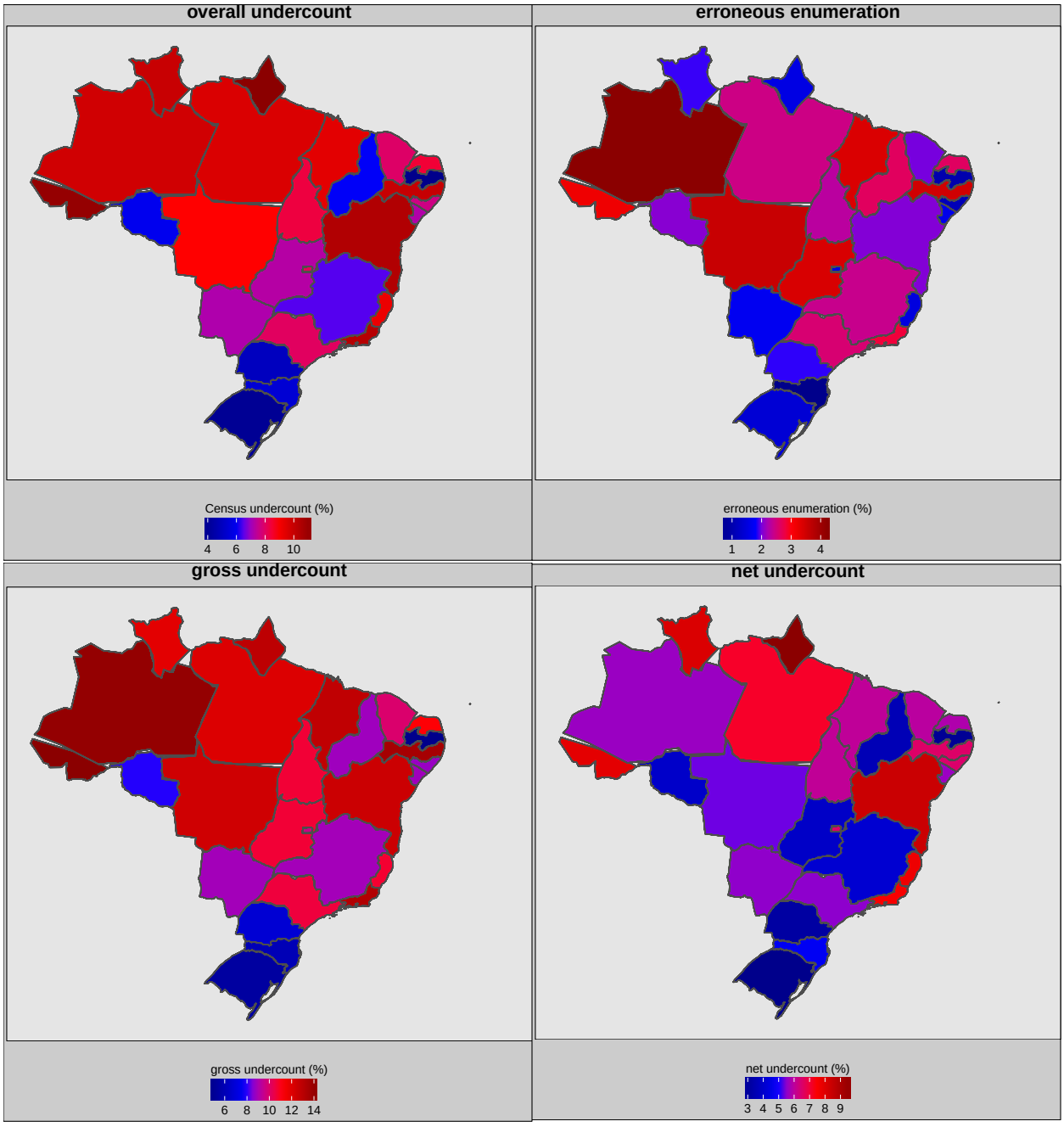

Figure 4.5: Map of the census overall undercount, erroneous enumeration, gross census error and net census error by state, 2000 (in %). Source: PES of the 2000 Census

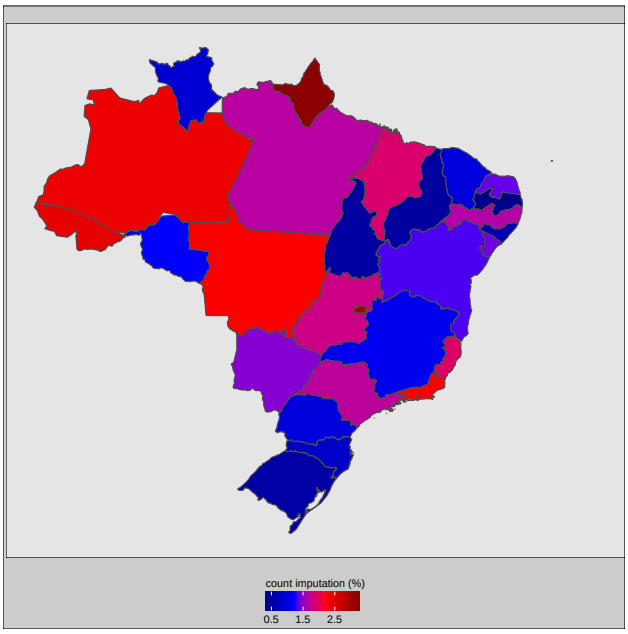

Figure 4.6: Map of the proportion of the population filled in as count imputation (in %). Source: IBGE, 2010 Census

The PES reports of the 1991 and 2000 censuses (L. C. S. Oliveira, de Freitas, et al., 2003; L. C. S. Oliveira, Indá, et al., 1996) have also published undercount rates for Brazilian states for the age groups: 0-1; 1-4; 5-14; 15-59; 60+. This type of indicator refers to the omission of people living in households in which the family group is the same in both the census and the PES. As discussed above, this type of omission accounted for a much higher proportion of the overall census undercount in 1991 than in 2000.

Children under age 1 are by large the most omitted group, and the age groups of people over 60 and aged 5-14 have the lower undercount rates. The omission of young children is consistent with the national and international experience, despite efforts to improve the coverage of this group. The low undercount rates for the elderly contradicts the belief that old people are also more likely to be omitted in the censuses. The pattern by age of census coverage in households enumerated in the PES and missed by the census is unknown. One may assume that the composition of the population living in households missed by the census and enumerated by the PES is the same as that captured by the census. It is possible, however, that these households are disproportionally composed of certain population groups, such as newly formed households, which would lead to a disproportionally higher omission of young adults.

DA carried out by IBGE at the national level in Brazil (IBGE, 2013b, 2018) confirms the well known pattern of census undercount, with high omission of children of both sexes, undercount of the working age population, particularly men, in addition to a net over-enumeration of the elderly, possibly associated to problems related to age misstatement in

this group. Although in principle this seems consistent, part of the estimated census undercount may result from failure to estimate demographic parameters (fertility, migration and mortality) correctly, which is the main limitation of DA, particularly in contexts with defective administrative data. The 2018 revision of the population projections (IBGE, 2018), for example, re-estimated fertility rates downwards, resulting in a lower estimate of undercount rates of children, compared to that calculated previously IBGE, 2013b.

Census evaluation using DA for subnational level in Brazil is more complex, since, in addition to the lack of good vital statistics to estimate fertility and mortality, there is also high uncertainty in migration estimates. Contrarily to international migration, which has been relatively low in Brazil in the past decades, internal migration accounts for an important portion of population growth in Brazilian states. The next section 4.2 presents some demographic indicators that can be used to indicate inconsistencies in demographic trends at the subnational level.

In principle, direct and indirect methods of census evaluation are not competitive. Even when both techniques produce discrepant results, they should be used in combination in an attempt to find the sources of the discrepancies. The PES could provide additional information to incorporate to the analysis of the population dynamics. On the other hand, DA requires consistency between population estimates and the demographic dynamics, and could be used to ratify PES results or demonstrate the need for adjusting their figures. Moreover, the PES can provide geographically disaggregated information, which is more complex to obtain by indirect evaluation, in particular because of the distortions which may occur with open populations regarding mobility, or the occurrence of local events without an almost unimportant national expression (Chackiel, 2009).

Quality of age reporting in Brazil and states from 1980 to 2010

In addition to census coverage, evaluation of the quality of the reporting of age is an important part of census evaluation.

The form of collection of the information about age in Brazil has been essentially the same since the 1940 Census: respondents are first asked about their date of birth; if they fail to respond the complete date, they are asked about the completed age at the reference date of the census [7].

Digit preference and avoidance

Important improvements in the reporting of age have been observed in Brazilian censuses from 1940 to 1980, reaching reasonably accurate levels of digit preference in 1980 (IBGE, 2012). These improvements were result of advances in the way the information is collected, in addition to the decreasing difficulty of people to provide their age precisely.

[7] The 1960 Census only asked about the age of the respondent and the 2000 Census asked both questions, regardless the response of the first one.

Demographers have assessed age heaping by using visualization techniques and by calculating specific indices. One powerful and simple visualization technique is to plot population pyramids by single year of age.

Figure 4.7 shows the population pyramids by single years of age for the four Brazilian censuses between 1980 and 2010, indicating important demographic changes, such as population growth and aging. In addition to changes in population size and age structure, Figure 4.7 also shows patterns of age heaping. The preference for ages ending with 5 and, specially 0, is clear in the 1980 census. This problem improved significantly in the 1991 Census, although part of this improvement may be due to the year the census was carried out, which ends in 1. There seems to be preference for both ages and year of birth ending in 5 and, specially in 0. The problem of digit preference seems to have exacerbated in 2000 and remained in 2010 (Figure 4.7).

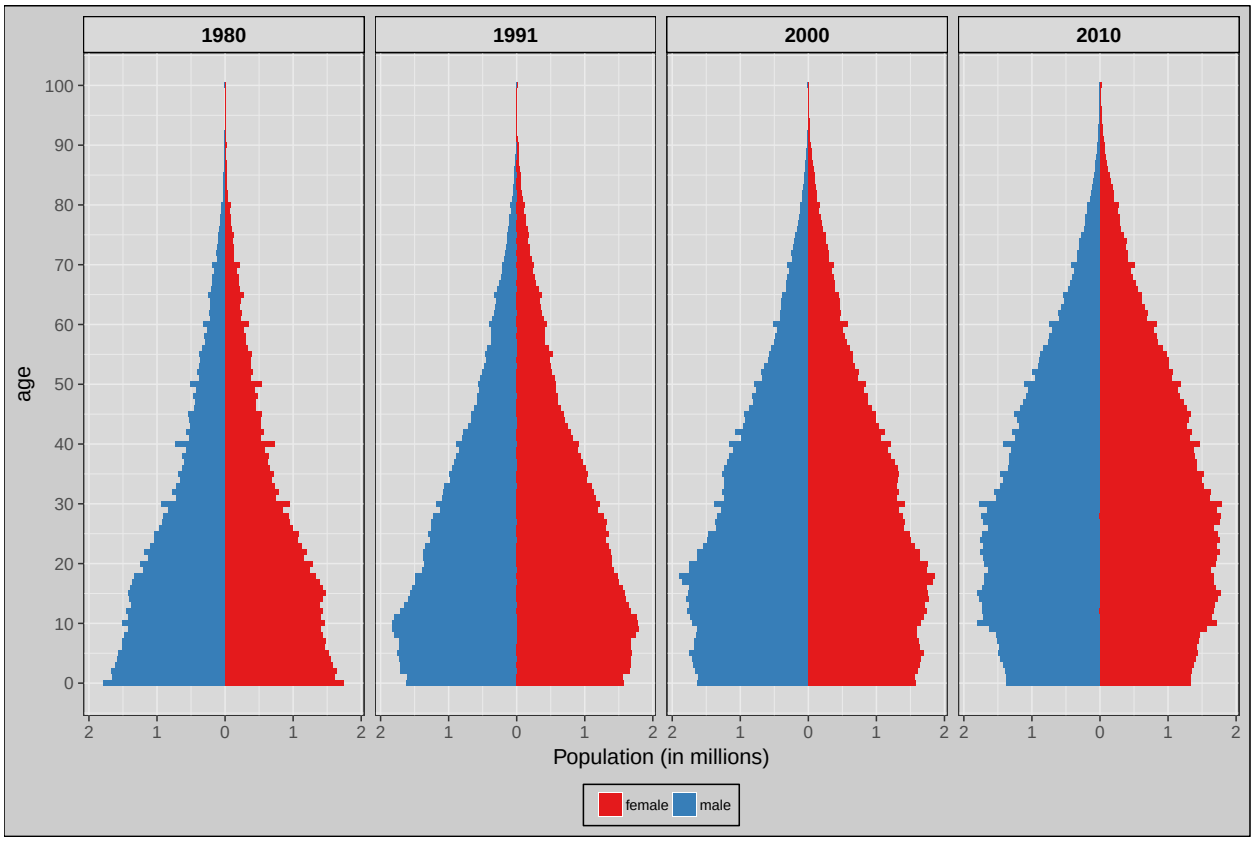

Figure 4.7: Population pyramids for Brazil, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

Section B.1 in the Appendix B shows the population pyramids for all states from 1980 to 2010. A few states, particularly in the less developed regions, had serious problems of digit preference in the 1980 Census, e.g. MA (Figure B.3), PB (Figure B.4), AL (Figure B.5) and

BA (Figure B.6). More developed states, such as SP (Figure B.7), SC, RS (Figure B.8) and DF (Figure B.9) had almost no preference for terminal digits in 1980. Age heaping seems to have improved for the less developed states, although the problem persists in 2010, and remained at low levels in the more developed states. An exception is RJ (Figure B.7), where age reporting seems to have worsen in the two most recent censuses.

In addition to visualization techniques, demographers have long used special indices to measure digital preference. The two most used indices are the Whipple Index and the Myers Index (Bryan and Heuser, 2004).

The Whipple Index measures concentration in specific terminal digits, such as 0 and 5, and is often calculated for ages between 25 and 60. The index for ages 0 and 5 is given by:

$$\mathrm{WI}_{0,5} = \frac{5 \sum\limits_{x \in N} K_x}{\sum\limits_{x \in D} K_x}, \tag{4.3}$$

where $N = \{25, 30, \cdots, 55, 60\}$, $D = \{23, 24, \cdots, 61, 62\}$ and K_x is the number of people at age x.

This index ranges from 0 to 500, where values below 100 represents avoidance for digits 0 or 5, 100 indicates no preference for these digits and 500 means that only 0 or 5 are reported. The Whipple Index is an accurate index, especially in distinguishing among relatively low degrees of heaping, but it fails to capture forms of heaping other than multiples of five (A'Hearn, Baten, and Crayen, 2006).

The Myers Index is conceptually similar to the Whipple Index, but it considers concentration at ages ending in each of the digits from 0 to 9. The Myers's Blended Method takes into account the shape of the population age structure. This methods provides an index of preference or avoidance for each terminal digit by comparing the proportion of the total population reporting ages with a given terminal digit to 10%, under the assumption that all digits are equally likely (Hobbs, 2004).

Spoorenberg, (2007) proposes an extension of the Whipple Index that summarizes all age preference and avoidance, which is given by:

$$\mathrm{SWI}_{tot} = \sum_{i=0}^{9} \left| \mathrm{WI}_i \right|, \tag{4.4}$$

where WI_i is the variation of the Whipple Index for terminal digit i proposed by Noumbissi, (1992). For $i = 5$, for example, WI_i is given by:

$$\mathrm{WI}_5 = \left(\frac{5 \sum\limits_{x \in N} K_x}{\sum\limits_{x \in D} {}_5K_x} \right) - 1, \tag{4.5}$$

where $N = \{25, 35, 45, 55\}$, $D = \{23, 33, 43, 53\}$ and ${}_5K_x$ is the population between ages x and $x + 5$.

When WI_i is greater than 1, there is preference for digit i, whereas when WI_i is lower than 1, there is avoidance of digit i.

Spoorenberg's Total modified Whipple Index (SWI_{tot}) is the preferred index used in this dissertation, since it gives similar pattern, though it is simpler than the Myers blended index (Spoorenberg, 2007). Section B.2 in the Appendix B shows results for the Whipple Index and Myers' Index, since these are the two most commonly used indices of digit preference and may be used for international comparison.

Figure 4.8 shows the avoidance or preference for digits from 0 to 9 for the 1980, 1991, 2000 and 2010 Brazilian censuses. There was significantly higher digit preference and avoidance in 1980 than in the other three censuses under analysis. As expected, the digit with stronger preference is 0, followed by 5 and, to a lesser extent, 2. There is a weak preference for 1 in 1991, which contrasts with the strong avoidance of this digit in the other years. This indicates preference for year of birth ending in 0, in addition to ages ending in 0.

The avoidance of 1 is stronger than that of 9. The avoidance of these two digits probably indicates transfer from ages ending in 1 and 9 to ages ending in 0. The transfer from 9 is more problematic because it results in changes of five-year age groups, which is the most common grouping in demographic analysis. When ages ending in 0 are chosen more often than those ending in 5, which is the case for the 1980, 2000 and 2010 censuses, there tends to be a surplus of persons reported at age groups such as 30-34, 40-44 and 50-54 (Ewbank, 1981). Bias in the age structure as a results of digit preference can also occur at old ages. The pyramidal shape of the population age structure convert even random age heaping to net overcount among these age groups (Preston and Elo, 1999).

Table 4.4 shows the Spoorenberg's Total modified Whipple Index (SWI_{tot}) by sex for the Brazilian censuses from 1980 to 2010. The results confirm the visual analysis of the population pyramids, showing an important decline in the digit preference from 1980 to 1991, partially due to the fact that the 1991 Census ends in 1. Compared to 1991, there is stronger digit preference among males in 2000 and 2010. The difference by sex was higher in the 2010 Census, in which the SWI_{tot} was 0.211 for males and only 0.152 for females. Despite long term improvements in the quality of reported age since 1940, it is not clear that digit preference has improved in the last two censuses.

These trends are consistent with the results of the Whipple Index and Myers' Index (section B.2 in the Appendix B).

Figure 4.9 shows the maps of SWI_{tot} for both sexes combined for states from 1980 to 2010. Again, it is clear that the great improvement in terms of age reporting took place between 1980 and 1991. The regional pattern is similar in 1991, 2000 and 2010. Contrary to census undercount, there is a clearer relationship between digit preference and socioeconomic characteristics of the states. States of the South, Southeast and Midwest have had lower digit preference, whereas the Northern states have had stronger digit preference.

Most of the states in the North and Northeast regions improved the reporting of age between 2000 and 2010. The state with the most significant deterioration in the digit preference

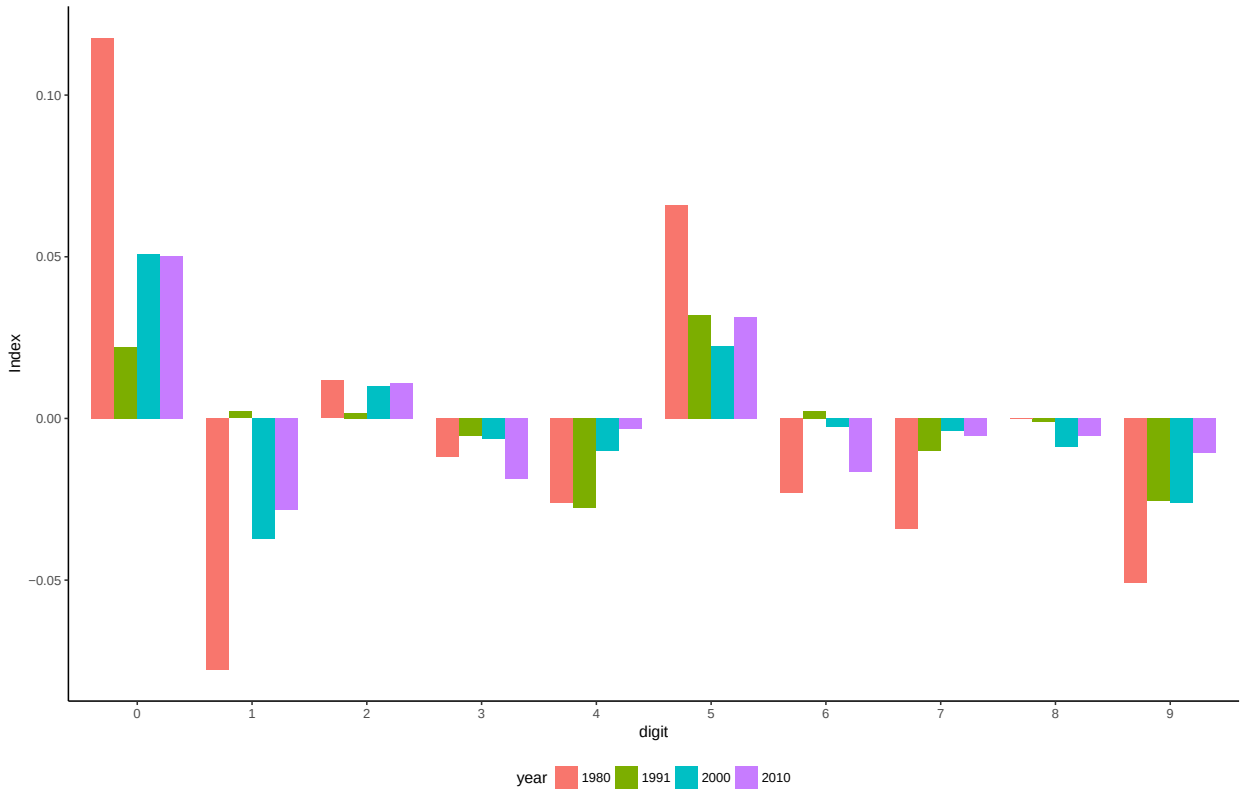

Figure 4.8: Modified Whipple's Index for terminal digit i, WI_i, for the 1980, 1991, 2000 and 2010 censuses. Source: Source: IBGE: 1980-2010 censuses

Table 4.4: Spoorenberg's Total modified Whipple Index by sex, Brazil, 1980-2010

Year	Sex	
	male	female
1980	0.424	0.419
1991	0.127	0.137
2000	0.190	0.167
2010	0.211	0.152

Source: IBGE: 1980-2010 censuses

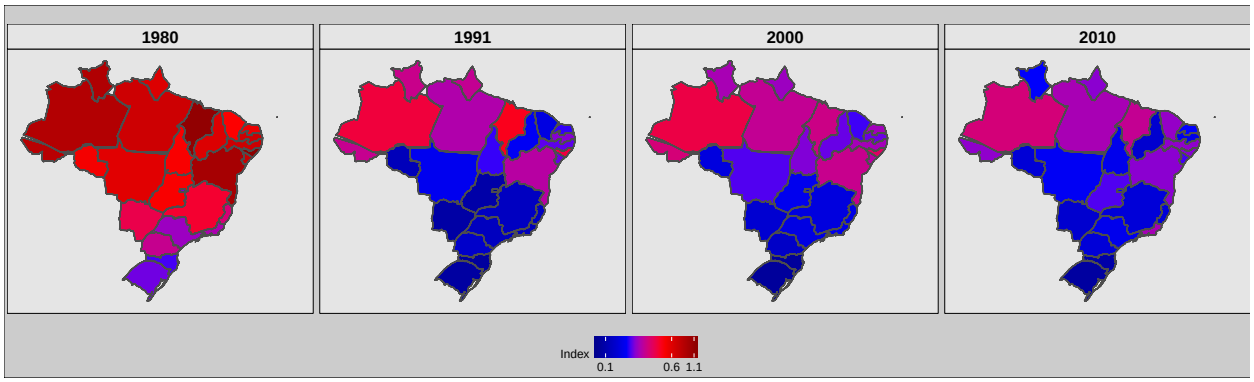

Figure 4.9: Map of Spoorenberg's Total modified Whipple Index, Brazil, 1980-2010. Source: IBGE: 1980-2010 censuses

in that decade was RJ, which had an increase in SWI_{tot} from 0.16 to 0.25.

Sex Ratios

A simple indicator that allows the identification of sex-specific age misstatement is the SR by age, which is given by the ratio between the male and the female populations by age group. It can also indicate differential census coverage by sex. The main limitation of this measure is that the SR in a population is also result of its demographic features, such as the SR at birth and differentials by sex in migration and mortality. However, since some of these patterns are known, data quality checks can be performed.

Figure 4.10 shows the SR by five-year age groups for the four censuses under analysis. The SR for 2010 resemble the expected pattern for this indicator, with values higher than 1 for children, due to the SR at birth around 1.05, followed by a decline due to higher mortality of male population at all ages. The steep decline between ages 10 and 35 can be justified demographically by a much higher male mortality at these ages or by a significantly higher migration among men compared to women. This pattern can also indicate data quality issues, such as differential census undercount, in which omission of males is higher than that of females at these age, a feature seen in many contexts. It can also indicate issues of age misstatement, such as female age understatement.

The pattern for the 2000 Census is similar to that observed in the 2010 Census, but with a steeper decline below age 35. The possible explanations for the shape of the SR by age are also similar for both censuses.

SR for the 1980 and 1991 censuses show a more irregular shape, with some peaks (e.g. at age 40) and valleys (e.g. at age 35). The excess female between ages 15 and 35 and relatively more male in the forties and fifties, more pronounced in 1980, is typical of young female age understatement. This is consistent with the known Latin American pattern of age misstatement, which shows a general surplus of women reported at young ages, around 20-29 (Ewbank, 1981).

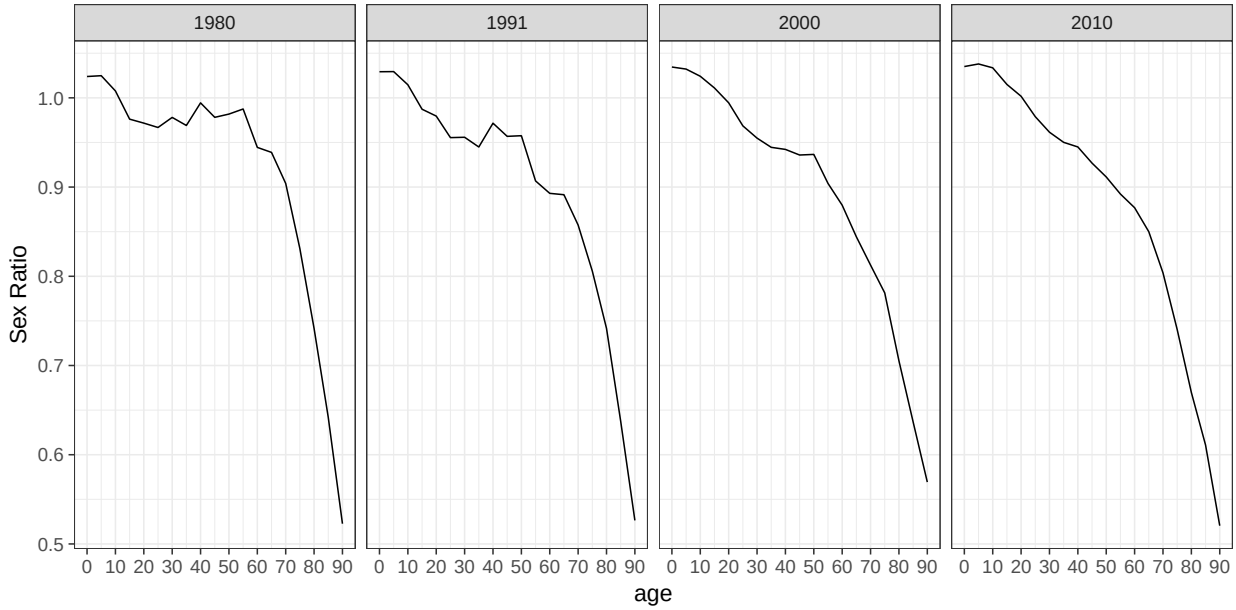

Figure 4.10: Sex Ratios by age, Brazil, Brazilian Censuses of 1980, 1991, 2000 and 2010. Source: IBGE: 1980-2010 censuses

In addition to age misstatement, there have been other competing explanations for the pattern of the SR in the 1980, 1991 and 2000 censuses. They have been explained, for example, by selective international migration (Carvalho, 1996; Carvalho and Campos, 2006) or by differential census undercount (IBGE, 2013b) under the assumption of zero net migration.

It is also worth noting that the SR in 1991 is significantly lower than that in 1980. Carvalho, (1996) and Carvalho and Campos, (2006) claim that this might be due to selective migration, in which males migrate more than females. Alternatively, this can be due to improvements in age misstatement, following the historical trends of previous censuses (IBGE, 2013b).

It is unlikely that a single factor is able to explain the SR observed in the censuses. It may be better explained by a combination of all the reasons listed above. However, it is extremely difficult to disentangle the contribution of each one. There remains a need for methods that combine all these explanation in a comprehensive model. A proposition along these lines is described in Chapter 3, with application to Brazil, which is shown in Chapter 5.

Section B.3 in the Appendix B shows the SR for all states from 1980 to 2010. It is more difficult to identify data quality issues for states, because demographic dynamics for subnational levels are more influenced by internal migration, which tends to have an significant impact in the SR.

Cohort Survival Ratios

Another simple method that is often used to check the quality of censuses is the analysis of the CSR. This procedure consists of comparing populations of the same cohort in two consecutive censuses. In addition to patterns of errors, these ratios also capture purely demographic changes, which makes disentangling what is census error from what is demographic change the main difficulty of interpreting the CSR. However, the knowledge of the structure of migration and mortality rates can help identifying patterns of census error.

To calculate these ratios for Brazil and states, populations enumerated in the censuses by age and sex were interpolated or extrapolated to July 1st of the years 1980, 1990, 2000 and 2010, resulting in exact 10 year intervals. This was done by using the intercensal exponential growth rate of the respective sex and age group:

$$r = \frac{logK(t_2) - logK(t_1)}{t_2 - t_1} \tag{4.6}$$

where $K(t_2)$ is the population at time t_2 and $K(t_1)$ is the population at time t_1.

Figure 4.11 shows the CSR for males and females by intercensal period. In the absence of migration, CSR should be necessarily less than or equal to one and monotonically decreasing for most of the age groups, as a result of the increase in mortality with age. CSR should be lower for males than for females, also as a result of differential mortality, in favor of women.

The first noticeable data quality issue is indicated by CSR greater than one for children aged 0-4 and, to a lesser extent, 5-9 at time t_1. This suggests an undercount of children below age 10 for both sexes in all censuses. CSR for these ages is slightly lower in the period 1980/1990, probably due to higher infant mortality.

Another recurrent pattern in recent Brazilian censuses is a valley between ages 15 and 24 (10-19 at time t_1 and 20-29 at time t_2), followed by a peak between ages 25 and 34 (20-29 at time t_1 and 30-39 at time t_2), depending on the year, which includes values greater than one for females aged 30-34 in 2005. These could be explained by emigration at ages around 15-24 and immigration (possibly return migration) at ages 25-34. Alternatively, this pattern can result from census undercount among young adults.

Another hump, around age 50 in the middle of the intercensal period, more pronounced for females, reinforces the idea of age understatement of women at these ages, as it is difficult to think about plausible demographic events that would explain this pattern.

Figure 4.11 and Figure B.15 (Appendix B) indicate certain consistency in the CSR at old ages, reflecting the expected mortality differentials, in that the curves for females are higher than those for males and they increase over time.

This apparent consistency may be hiding patterns of age overstatement, which is common in many contexts (Coale and Caselli, 1990; Preston and Elo, 1999), including in Latin America (Del Popolo, 2000; Romero and Freitez, 2008). Only the DA, which seeks for the consistency between consecutive censuses and the plausible mortality and migration estimates for those ages in the intercensal periods, can indicate whether the CSR are within a reasonable range.

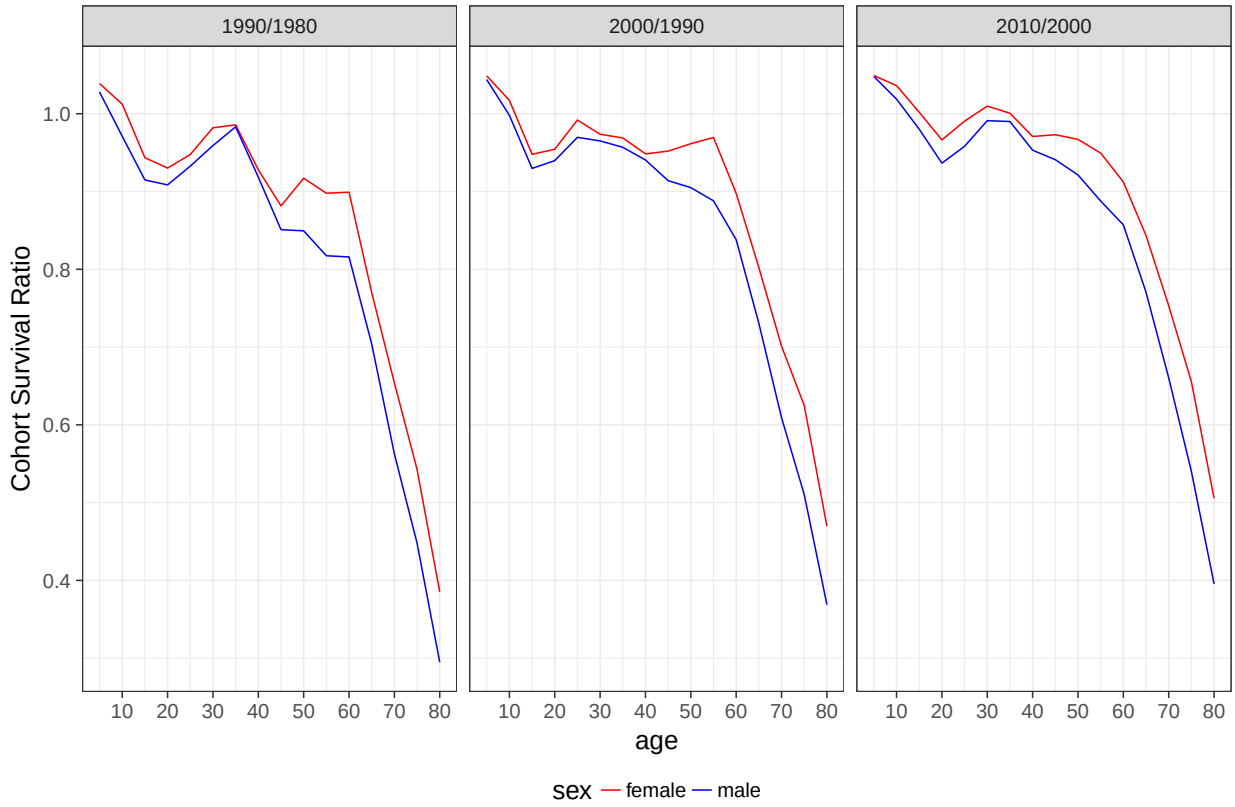

Figure 4.11: Cohort Survival Ratios for the intercensal periods 1980/1990, 1990/2000 and 2000/2010 by sex and age group at the middle of the intercensal period, Brazil. Source: IBGE: 1980-2010 censuses

Section B.4 in the Appendix B shows the same information for Brazil with a different presentation, with the graph broken down by sex and all the periods on the same plot, as well as the CSR for all states from 1980 to 2010. It is also difficult to identify pattern of census errors for states based only on the CSR, due to the high migration rates.

Discussion and conclusions

This section presented a comprehensive evaluation of the last four Brazilian censuses, conducted from 1980 to 2010. This is an important exercise as censuses are the main source of demographic data and are used for innumerable purposes.

The analysis in this section innovates by exploring for the first time the results of the PES by age, sex and geography, in combination with demographic techniques, to be used further to produce more reliable and consistent population estimates.

The discussion about the different techniques for census evaluation, particularly PES and DA, and the examination of their results evidence their strengths and limitations. The

PES and the procedures to derive indicators of census coverage from them are extremely complex, and have many potential sources of error. However, they produce useful results that would otherwise be particularly difficult to obtain, such as census coverage rates for subnational levels. On the other hand, the PES may produce results that are inconsistent with the population dynamics of the region under analysis, which can be resolved by DA and then consistent demographic estimation can be produced. Furthermore, the PES rarely provide evaluation on content errors, whereas DA can indicate some of these problems.

The results of both direct and indirect census evaluations in Brazil have shown some consistent patterns of census coverage, such as higher undercount of children compared to other population groups. On the other hand, the PES have indicated higher census undercount than DA carried out by IBGE. Time trends are also inconsistent.

The PES results for Brazilian censuses have shown a net census undercount, which does not mean that censuses are free of overcount problems, but that undercount tends to exceed overcount.

The reasons that lead to census errors are complex, and it is extremely difficult to derive conclusions only from the PES and DA results, but some results are consistent with the expected patterns.

Errors may derive from factors directly related to the census activities, such as planning, training, human and material resources and questionnaire design. These would explain, for example, a higher census undercount in the 1991 Census indicated by the DA. This is known to be a problematic census, that had to be postponed due to political and institutional issues. Another example is the low undercount rates in PB in the 2000 Census, attributed to a high quality census work in this state.

Census coverage may also be related to the difficulties of conducting a census in certain areas, which would explain higher census undercount in rural areas, and in states of the North Region, which are large and have many areas of extremely difficult access.

Census errors, particularly age misstatement, can be partially controlled by checking mechanisms, better training and questionnaire design, but are also likely to be related to factors extrinsic to the census. The educational level of the population enumerated, for example, may influence the knowledge and perception of the population about their age, which is consistent with the results that show higher digit preference in the least educated states. Increase in digit preference in highly educated states, such as that seen in RJ in 2010, can indicate deterioration of the overall quality of the census in that state.

Even though the census is an activity with rigorous standard procedures, it is relatively decentralized operationally. The IBGE state level agencies have a certain autonomy to coordinate the processes of the preparation and conduction of the census in their states. This can be one reason to explain different levels of census undercount for neighbors and relatively similar states.

Results from census evaluation provide extremely useful information that could be used for many purposes. The basic premise of census evaluation, which is to improve future censuses, for instance, should be put into practice. Results from previous evaluation should be used for planning the next censuses. Special attention should be devoted to states that

have shown higher undercount rates. The experience of the states with low undercount rates should be studied as best practice to be potentially replicated in other states.

Results of census undercount by age group should be also used in planning future censuses. For example, a reminder about old people is common in census questionnaires, based on the supposition that this group is disproportionately undercounted. However, results from direct and indirect evaluations have often contradicted this premise. On the contrary, old people often appear to be over-counted due to age misstatement.

In cases where the undercount of children typically occurs due to an omission of respondents in properly counted households, additional control measures can be applied. Warnings highlighting this issue in the enumerator and supervisor training and manuals is a simple measure that should be included. Another measure to address this problem is a change in the questionnaire to make the enumerator ask explicitly for the number of children living in the household.

Furthermore, the strategy of census evaluation should be part of the census process, and clearly defined how PES and DA will be used.

Direct and indirect methods of census evaluation are complementary and should be used in combination, trying to make use of the advantages of each technique and resolving inconsistency between results from the different methods. Uncertainty in results of both the PES and demographic information can be incorporated, and results will depend on both the uncertainty in individual pieces of the demographic balancing equation and the consistency among them.

In many cases, it is unclear whether the pattern observed in demographic indicators are result of data quality issues or consequence of purely demographic changes. Thus, the incorporation of uncertainty measures can be used to derive the more or less likely scenarios. Chapter 5 details the procedures used to transform the census indicators of quality and coverage discussed in this section into probability distributions to be used in the models described in Chapter 3.

4.3 Fertility Estimates for Brazil and States from 1980 to 2010

CRVS systems are the natural data source for fertility estimation, but they suffer from coverage and quality problems, especially in developing countries. This might be due to lack of incentives to register a birth, which is aggravated when the child dies shortly after birth. Some of the births would be registered late, which can also affect fertility estimates (Moultrie, 2013b).

These limitations have stimulated the development of demographic methods for fertility estimations, often based on censuses and surveys.

One of the earliest indirect methods for fertility estimation is the own-children method, which consists of a reverse-survival technique that uses the population of children in a census to estimate fertility in the recent past (Cho, Retherford, and Choe, 1986; Grabill and Cho, 1965). The main limitation of this method is related to the accuracy of the children population in the censuses. As discussed in section 4.2, this group is one of the most likely to be omitted.

Fertility can be also estimated directly through data containing birth histories collected in surveys such as the DHS. Another approach, perhaps the most used in developing countries, combines the recent and cumulated lifetime fertility measures routinely collected in censuses and surveys. These methods reconcile information from recent fertility, which is often underestimated, with the total parity by age group, which is thought to be more reliable. There are several procedures that use these idea, generally called P/F ratio methods (Moultrie, 2013c; UN, 1983). The most used of these methods is the Brass P/F ratio method (Brass, 1964; Moultrie and Dorrington, 2008; UN, 1983), which will be discussed in detail below.

In Brazil, demographers have estimated fertility rates by using indirect demographic techniques, primarily the Brass P/F ratio method, for decades. Even though the limitations of this method have been widely known, the lack of an alternative data sources has led to a general consensus that these estimates reasonably describe the overall levels and trends in fertility (Berquó and Cavenaghi, 2014; Borges and L. Silva, 2015; Carvalho, 1982). More recently, with the continuous decline of fertility levels and the rapid change in the age schedule, in addition to a greater availability of alternative data sources due to the improvement of vital registration systems, scholars have challenged the results of these techniques for the Brazilian context (Carvalho, Gonçalves, and L. Silva, 2018; Castanheira and Kohler, 2015).

This issue remains unsolved and there has been significant disagreement about the levels of fertility for the past decades, particularly for subnational levels. Fertility estimates using different methods and data sources have led to different results. Indirect demographic methods have several limitation, but CRVS systems in Brazil are also limited. Despite substantial improvements oven the last years, there are still a large proportion of births that are not registered, particularly in the less developed regions, which undermines their use without any adjustments. Furthermore, population estimates require long time series for periods when

administrative records were wisely limited and sometimes inexistent.

This section discusses different possibilities for estimating fertility and completeness of birth counts, to be further used in combination in the integrated model for population estimates.

Evaluation of registered births

Brazil has currently two administrative record systems that collect information on births: the CR and the VS.

There have been some efforts to establish public registration of births since the beginning of the 20th Century, but only in the mid-1970s the statistics of the CR started to be produced in a more structured basis, by the IBGE. IBGE have collected these data from all registry offices in the country.

Despite improvements over time, problems such as under-registration and late registration as still common, particularly in less developed regions. An important measure that has promoted the registration of births is the legislation from 1997, which establishes the universal and free access to birth registration. Poverty and social exclusion have been linked to the underregistration of births. Until 1997, there was a cost for requesting the death certificate. The implementation of the gratuity in 1997 and further policies implemented in the 2000s have reduced the underregistration of births, but this problem remains as a result of the difficulties to access the civil registration offices, particularly in those regions where the offices are far away (A. T. R. Oliveira, 2018).

More recently, other measures have been implemented to ensure the birth registration as a basic and fundamental human right.

Contrary to the slow progress in improving the quality of VS in the developing world (Mikkelsen et al., 2015), including in some Latin American countries (Guzmán et al., 2006), Brazil has made considerable improvements on this matter. However, CRVS systems in Brazil are still incomplete in several regions in the country.

To overcome the limitations in the CR system and to respond to specific demands about health issues, the Ministry of Health created, in the 1990s, a VS system, named Live Births Information System (SINASC). SINASC is a system that compiles information about live births occurred in hospitals and other health facilities. The implementation of the system was gradual and started in the capitals of the states (Jorge, Laurenti, and Gotlieb, 2007).

In the first years after the implementation, the number of registered births in the SINASC was lower than that reported to the CR. More recently, the SINASC has had higher coverage. In 2000, the number of births reported to the SINASC was 7.3% lower than that of the CR (including late registration), with a great regional variability. In the more developed states, the numbers in both systems were close, whereas in MA, the number of births in the SINASC was almost 30% lower. The 2000s was a decade of convergence in the coverage of both systems and the number of births reported to both systems was pretty similar in 2010 for Brazil and almost all states (Borges and L. Silva, 2015).

There are essentially two ways to evaluate the completeness of registered births.

First, it possible to calculate fertility rates from independent sources such as censuses and surveys and then estimate the number of births to be compared to the registered births. These methods will be discussed below (Section 4.3).

Alternatively, direct techniques that use the registered births can be applied. One of these techniques is the capture-recapture, which uses information from overlapping data collected from different sources to estimate the completeness. Because it requires matching individual records, this method is expensive and time-consuming, and has also methodological limitations, similar to those discussed in the Section 4.2 for the PES (Hook and Regal, 1995; Sekar and Deming, 1949).

Capture-recapture studies have been conducted recently to combine both CRVS systems in Brazil. Trindade, L. F. L. Costa, and A. T. R. Oliveira, (2018) apply the capture-recapture method to estimate the undercount of both the CR and the VS systems. The estimation process is done after matching the births registered in both systems and then calculated using the capture-recapture method described in the section 4.2 under the assumption of independence between both sources.

Figure 4.12 shows the under-registration of births by state for the year 2015 for both CR and VS systems. The results show that there are still problems with the births registration in both systems and regional inequalities persist. The map also shows that the under-registration of births in the CR is higher than in VS, particularly for the states in the North region. This is probably due to the difficulties to access the registry offices in this region due to the long distances to registry offices. The underregistration of births in the CR in the state of RJ is surprisingly high.

These results offer some insights about the the recent pattern of completeness of registered births in Brazil. However, since the study was only conducted for 2015, the results provide limited use for population estimates for Brazil and states for the period 1980-2010, the main focus of this dissertation.

Moreover, one of the main assumptions of the capture-recapture technique is independence between between both sources. This assumption is questionable in this context, indicating that the true underregistration could be even higher for both data sources.

Another study that aims to estimate the underregistration of births in the SINASC is the "proactive search" survey, carried out in 129 municipalities of the Amazon and Northeast regions in 2010. This survey searched for unregistered vital events with a reference date in 2008 and found unregistered vital events in hospitals and other health facilities, as well as in non-official sources, such as illegal cemeteries. (Szwarcwald, de Frias, et al., 2014; Szwarcwald, Morais Neto, et al., 2010). The sample stratification and the method used for adjustment was developed to allow for generalization for correctness of vital statistics for all municipalities for the period 2000–2010.

This generalization to all municipalities based on a limited sample and, more importantly, to other time periods, is based on strong assumptions about regularities in the adjustment factors by strata and time, which probably does not hold in practice. Furthermore, contrary to the methodological idea of the capture-recapture technique, this procedure assumes that

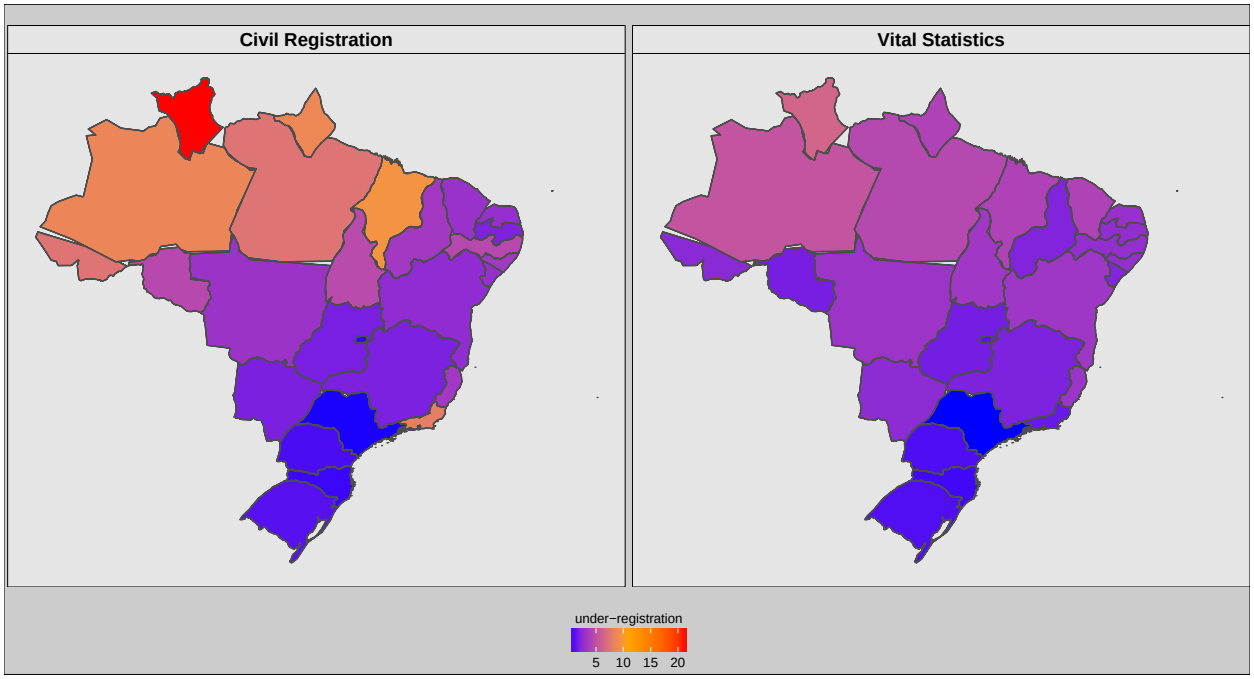

Figure 4.12: Map of the under-registration of births data source (CR and VS) and state, 2015 (in %). Source: Trindade, L. F. L. Costa, and A. T. R. Oliveira, (2018)

all births that occurred and were not recorded were captured by the survey, which could lead to a sub-enumeration of the completeness measure.

Comparison between the number of births corrected by the adjustment factors estimated based on the "proactive search" and the births registered in the CR indicate that the "proactive search" adjustment factors are underestimated, at least for 2000, in many states where the registered births in the CR is higher than the adjusted (Borges and L. Silva, 2015).

Despite the limitations of the method, these results are useful as they are the only source of information of the completeness os births informed to the SINASC for the period 2000-2010. Figure 4.13 shows the under-registration of births by state for the years 2000 and 2010, indicating important regional inequalities, similar to those found by the capture-recapture technique described above. The Southern states have higher completeness, whereas the states in the North and Northeast regions are those with higher underregistration of births.

The map also confirms the significant improvement in the completeness of registered birth in Brazil in recent decades, particularly for the states in the North and Northeast regions. In 2000, according to the "proactive search" survey, more than 30% of the births in MA were not registered. This proportion reduced to 11% in 2010. Several other states in the North and Northeast regions had around 10% of under-registration of births. All the states in the South region, those in the Southeast (except MG), DF and Mato Grosso do Sul (MS) have had almost complete VS systems since 2000 4.13.

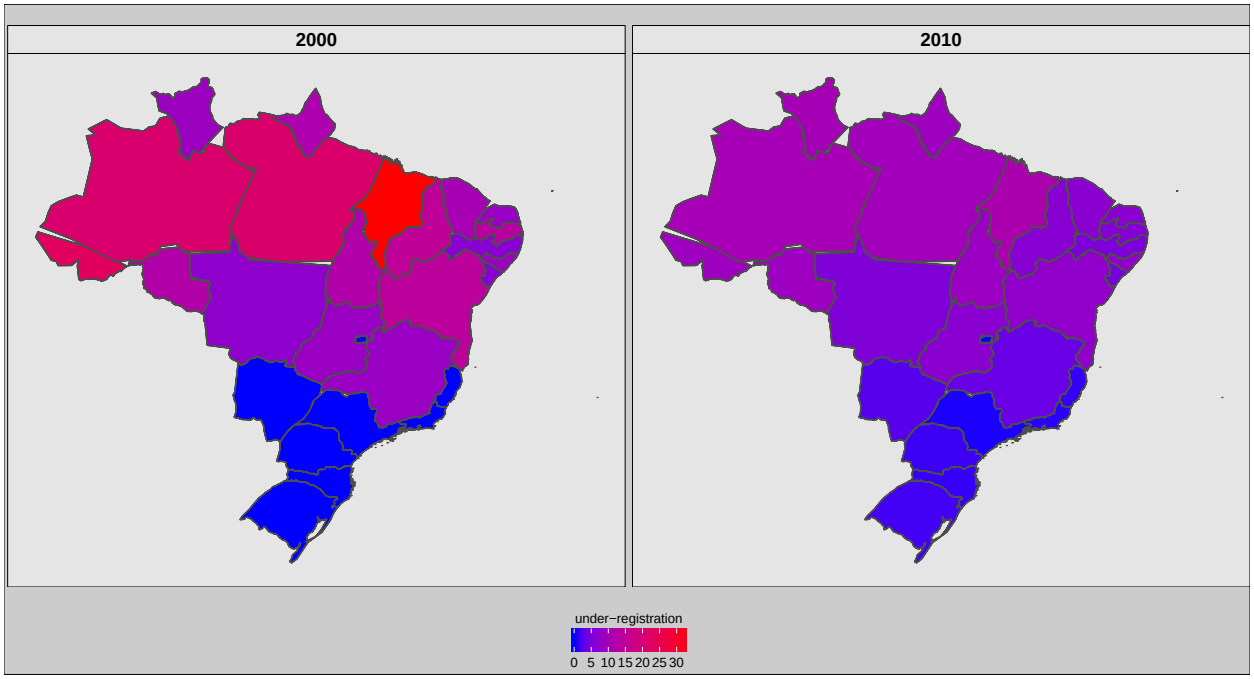

Figure 4.13: Map of the under-registration of births by state, 2000 and 2010 (in %). Source: RIPSA, (2013)

No direct estimates for the completeness of registered births are available for years prior to 2000. For this period, the under-registration of births needs to be estimated indirectly through the calculation of fertility rates. Methods for fertility estimation are discussed in the next section (Section 4.3).

Brass P/F ratio method and sensitivity analysis

As discussed above, the most used indirect technique to estimate fertility in contexts of defective vital registration systems is the P/F ratio method (Moultrie and Dorrington, 2008). This method was first proposed by William Brass in the 1960s, initially for application to the African populations (Brass, 1964; Brass and Coale, 1968), and remains useful for estimating fertility in many countries.

The basic idea of the method is that, under certain assumptions, the number of children ever born, or parity (P_i), at an exact age i equals the sum of the period age specific fertility rates (F_i) to that age. Any difference in these two measures would be attributed to data quality problems, often an underreport of current fertility. The ratio for the age group 20-24, for example, given by $P_{20\text{-}24}/F_{20\text{-}24}$, could then be used to adjust the fertility rates for all ages groups under the assumption that current fertility is underreported by a constant factor (Brass, 1964; UN, 1983).

The information used to calculate P_i and F_i comes from questions asked in censuses and surveys about children ever born and children born in the 12 months prior to the census.

Information on children ever born is often collected by questions such as: *"How many children born alive have you had?"*. This question requires no information on ages and dates so that there are no dating errors. On the other hand, there might be some imprecision in the number of children reported. It has been observed that children tend to be omitted, particularly by women aged 35 and over. Information for very young women are also thought to be more subject to reporting problems (UN, 1983).

Current fertility is often collected in censuses by asking questions such as: *"Have you given birth to any children in the past 12 months?"*. This information is thought to be underreported, presumably because of a misperception of the reference period (Brass and Coale, 1968; UN, 1983). The way this question is asked has changed to allow for a more precise measure of the number of children born in the 12 months before the survey by asking about the date the last child was born: *"What is the date of birth of the last children born alive you had?"*. Brazilian censuses implemented this change in 1980. In fact, the estimated adjustment factor for the reported current fertility declined from 1.31 to 1.12 between 1970 and 1980, indicating an important improvement on this information. The adjustment factors remained similar for the next two censuses (Carvalho, Gonçalves, and L. Silva, 2018).

Filtering the number of children born in the last 12 months improves the reporting of current fertility, as this avoids the error in the reference period, but this information remains underreported. One possible reason to this omission is related to the same phenomenon that leads to undercount of children in the censuses, that is, hard to count areas or population groups have a disproportional higher number of children. This may be related to the finding that newly formed households, which are more likely to have young children are more likely to be omitted in the census (see section 4.2).

In summary, in the P/F ratio Method, the age pattern of current ASFR obtained from surveys and censuses is accepted and the fertility level is adjusted by the average parity (number of children ever born) reported by young women. The idea of the method is to combine the measurements which are likely to be most reliable given the different nature of the deficiencies in the two sets of information (Brass, 1964). Since the nature of errors tends to be different, this procedure may produce plausible estimates even when both sets of data are subject to errors (Brass and Coale, 1968).

The method has thus two main assumptions in terms of the quality of the reported data: i) current fertility rates are underreported by a factor that is constant with age; ii) the reported number of children ever born is reliable.

Furthermore, there are three main conditions that need to be true for the method to be valid: i) constant fertility over time; ii) fertility of surviving woman is the same as fertility of those who died; iii) fertility of immigrant women is the same as non-immigrant fertility.

The P/F ratios are often calculated for all age groups between 15 and 49 because the pattern of the ratios by age may also reveal data quality problems and fertility trends. For adjustment purposes, the age group 20-24 is by far the most used and recommended, due to data quality issues for women in the extremes of the age range, in addition to attempts

to minimize the effects of fertility change (Brass, 1964; Moultrie and Dorrington, 2008; UN, 1983). Thus, this section concentrates on the analysis of the P/F ratio technique for women aged 20 to 24 at the date of the survey, which will be called P_2F_2[8], but all the results can be easily extended to other age groups.

To avoid confusion with the definition of age groups, let i be a constant representing the age to which the calculations are done and x be the actual age.

Formally, let

$$P_i(t) = \int_0^i f_x(t-i+x)\,dx, \text{ be the cumulative fertility up to age } i \text{ at time } t \tag{4.7}$$

$$F'_i(t) = \int_0^i f'_x(t)\,dx, \text{ be the cumulative observed period fertility up to age } i \text{ at time } t \tag{4.8}$$

$$F_i(t) = \int_0^i f_x(t)\,dx, \text{ be the cumulative true period fertility up to age } i \text{ at time } t \tag{4.9}$$

where $f_x(t)$ is the true ASFR at age x for year t. In this formulation of the method, the parity reported by woman at age i in census t, $P_i(t)$, is taken as the sum of the true historical fertility rates for these woman, as the core assumption of the method is that this information is reliable. The current ASFR at age x, given by the reported fertility rates in the 12 months prior to the census t, $f'_x(t)$, relates to the true ASFR by a factor $\mathrm{PF}_i(t)$ as follows:

$$f_x(t) = f'_x(t) \times \mathrm{PF}_i(t) \tag{4.10}$$

$$F_i(t) = F'_i(t) \times \mathrm{PF}_i(t) \tag{4.11}$$

When $f'_x(t)$ is underestimated, which is often the case, $\mathrm{PF}_i(t)$ is greater than one.

For example, $P_{25}(2000) = \int_0^{25} f_x(1975+x)\,dx$ is the parity, or retrospective fertility, of women aged 25 in 2000 and $F'_{25}(2000) = \int_0^{25} f'_x(2000)\,dx$ is the cumulative observed period fertility up to age 25 in 2000.

If the above-mentioned conditions are met, e.g., fertility below age i has been constant over time, $f_x(t-i+x) = f_x(t)$. Thus, $P_x(t) = F_x(t)$ and the P/F ratio $\frac{P_x(t)}{F_x(t)} = 1$. If the current fertility rates are omitted by a constant factor, as in equation 4.11, the P/F ratio will be greater than 1 and will indicate the adjustment factor for the current fertility:

$$\frac{P_i(t)}{F'_i(t)} = \mathrm{PF}_i(t) \tag{4.12}$$

[8]The index 2 has been used to refer to the second age group of women at reproductive ages, that is 20-24 (Brass, 1964; UN, 1983).

In practice, $P_i(t)$ is given by the average parity reported in the census at time t by women aged i:

$$P_i(t) = \frac{\text{CEB}_i(t)}{K_i(t)} \tag{4.13}$$

where $\text{CEB}_i(t)$ is the number of children ever born to women aged i at time t and $K_i(t)$ is the number of women from the age enumerated in the same census.

Similarly, $F'_i(t)$, in the discrete form of Equation 4.9, is calculated as a sum of the ASFR, $f'_x(t)$, up to age i reported in the census at time t:

$$F'_x(t) = \sum_{x=0}^{i} f'_x(t) \tag{4.14}$$

$$f'_x(t) = \frac{\text{B}_x(t)}{K_x(t)} \tag{4.15}$$

where $\text{B}_c(t)$ is the number of births the 12 months before the census at time t to women aged x and $K_x(t)$ is the number of women as previously defined.

When the implementation of the method is done by five-year age groups, which is the most common approach, and the P_2F_2 factor is used, the information on current fertility at age 20-24 needs an adjustment to be compatible with the average parity at the same age, as $\text{CEB}_{20\text{-}24}(t)$ refers to the cumulated fertility experience of women in the age group 20-24 at time t, including those in the beginning of the age group, e.g. at age 20. Thus, the process of cumulating current fertility considers the fertility of the entire group 15-19 and only part of the fertility observed in the age group 20-24:

$$f_{20\text{-}24}(t) = 5f_{15\text{-}19}(t) + k_{20\text{-}24}f_{20\text{-}24}(t) \tag{4.16}$$

where $k_{20\text{-}24}$ is the multiplying factor for deriving the parity from ASFR, which also considers a half-year displacement backward in time as woman had their children, on average, about six months before the reported age in the census. There are many ways to estimate these factors. Brass, (1964) propose a model that relates the multiplying factor ($k_{20\text{-}24}$) to the ratio of the ASFR between the first two groups: $\frac{f_{15\text{-}19}(t)}{f_{20\text{-}24}(t)}$. The bigger the ratio, the younger the fertility age schedule and the bigger the adjustment factor.

Sensitivity Analysis

This subsection develops an analytical framework to evaluate how results of the application of the method are biased when one or more conditions of the method are not met,

proposing adjustment factors to correct for these biases and to incorporate uncertainty in the estimates. Sensitivity analysis of indirect demographic methods, such as the P/F ratio, provides insights about the limitations of the methods and how sensitive they are to the violations of their assumptions and assist the estimates of the measures of uncertainty in demographic parameters in the next chapter.

Sensitivity analysis is this context is intended to evaluate how the results of a certain method would change when the assumptions are not satisfied. This procedure is useful to identify the most important assumptions required by the methods, indicating those that deserve special attention and those that are likely to have only minor impact in a certain context.

This analysis is also useful to propose adjustments in the methods if information on the factors affecting the results is available. In the specific context of this dissertation, this will be also useful to allow more precise prior distributions for the measures of uncertainty in demographic parameters.

Moultrie and Dorrington, (2008) conducted a sensitivity analysis of the P/F ratio method, evaluating the impact of changes in fertility and mortality on the resulting adjustment factors and fertility estimates. The authors propose the use of simulations to overcome the analytical complexity of the basic equations of the model. The simulations were carried out to mimic a typical demographic transition change, particularly in the African experience.

The results show that the errors in the P/F ratio for the age group 20-24 are relatively small in the scenarios of changes in the fertility levels and age distributions. For most of the time, the errors are of the order of 5% or less, reaching a maximum of about 10%. The results under these hypothesis tends to overestimate fertility. The authors claim that in the context of the generally poor data in which these methods are normally applied, errors of this magnitude are not a major cause for concern.

Simulations were also carried out to test the sensitivity of the method to differential fertility between between survivors and non-survivors. This is operationalized by testing differential fertility between HIV-infected and HIV-uninfected women, which indicates that this has a trivial impact on the methods, even in an environment with a simulated highly generalized epidemic. That effect serves to attenuate the overestimation of the adjustment factors due to fertility changes, but errors are of the order of magnitude of only -0.5%.

The study of Moultrie and Dorrington, (2008) offers important insights about the possible biases in the P/F ratio methods, and the order of magnitude of these errors. However, since they are based on simulations, the results are conditioned to the specific scenarios considered by the authors. Furthermore, they offer no possibilities for adjustments of the original proposition of the method.

In order to extend this analysis to other contexts and allow for adjustments in the original method, this section proposes an analytical sensitivity framework taking into consideration its main conditions.

Sections below develop a sensitivity framework to the hypothesis of differential fertility between survivors and non-survivors (4.3); differential fertility between migrants and non-migrants (4.3); and fertility change (4.3).

Mortality differential

The average parity calculated with the information of children ever born collected in censuses and surveys obviously refers to the fertility experience of the survivors of a particular cohort. If female mortality is low or there is no significant difference in fertility between survivors and non-survivors, then this information is a good proxy of cohort fertility up to that age.

It is generally assumed that the effect of mortality on the average number of children ever born is negligible, mainly because mortality is generally low for young women (UN, 1983). In fact, this effect seems to be negligible even in extreme cases. As previously discussed, Moultrie and Dorrington, (2008) report only trivial impact of differential fertility between HIV-infected and HIV-uninfected women on the method even in contexts of high HIV prevalence.

The formulation below offers some insights about the reasons mortality may have only a minor impact on biasing the results of the method. The results presented here are similar to those shown by Feehan and Borges (2018) for the sensitivity framework of the sibling survival method.

The average parity of women at the moment of the interview obviously reflects only those who survive, and is given by[9]:

$$P^s = \frac{\text{CEB}^s}{K^s} \tag{4.17}$$

The same quantity can be defined for women who died, although this is unobservable, reflecting the average parity they would have had if they had survived to the date of the interview:

$$P^d = \frac{\text{CEB}^d}{K^d} \tag{4.18}$$

Now, let P^{ds} represent the aggregate average parity:

$$P^{ds} = \frac{\text{CEB}^s + \text{CEB}^d}{K^s + K^d} \tag{4.19}$$

Let

$$\pi^d = \frac{\text{CEB}^d}{\text{CEB}^d + \text{CEB}^s}, \text{be the proportion of births from women who died} \tag{4.20}$$

$$\pi^s = \frac{\text{CEB}^s}{\text{CEB}^d + \text{CEB}^s}, \text{be the proportion of births from women who survived} \tag{4.21}$$

[9]the age group index is omitted in this subsection and all results refers to women aged 20-24

Suppose P^d and P^s differ by a factor R^{ds}, for $R^{ds} > 0$:

$$R^{ds} = \frac{P^d}{P^s} \tag{4.22}$$

Based on these definitions, Feehan and Borges (2018) show that the ratio between the unobservable average parity of survivors and non-survivors (P^{ds}) and the average parity for the survivors P^s can be expressed as:

$$\frac{P^{ds}}{P^s} = \frac{R^{ds}}{\pi^d + R^{ds}(1 - \pi^d)}. \tag{4.23}$$

Let PF^s be the adjustment factor of the age group 20-24 calculated by the application of the P/F ratio method and PF^{ds} be the "true" adjustment factor that would have been observed if all the women had survived to the date of the interview. Since the denominator of the P/F ratio is the same, the only difference between PF^s and PF^{ds} is in the numerator (the average parity): P^s and P^{ds}.

Thus, the ratio between PF^{ds} and PF^s, which signals the bias in the P/F ratio method due to the unmet condition of equal fertility for women who died or survived, is given by:

$$\frac{\text{PF}^{ds}}{\text{PF}^s} = \frac{R^{ds}}{\pi^d + R^{ds}(1 - \pi^d)}. \tag{4.24}$$

When $R^{ds} = 1$ (parity of survivors and non-survivors is equal) *or* when $\pi^d = 0$ (proportion of births from women who died is zero), $\text{PF}^{ds} = \text{PF}^s$, meaning that there is no bias in the result of the method.

This result shows that two conditions need to be simultaneously true to the unmet condition of independence between fertility and mortality have an impact on the results: i) fertility of women who do not survive to the interview differs significantly from those who survive; and ii) the proportion of women who die between the beginning of their reproductive lives and the age 20-24 is considerable.

Since these are exceptionally rare conditions in real populations, this issue should not be a major cause for concern of researchers when using this method. In any case, researchers can easily use equation 4.24 to asses the biases caused by fertility differentials in vital status with their own data.

In Brazil, even in the state with the highest mortality in 1980, the probability of a woman die between the age groups 15-19 and 20-24 was only about 1%. Even if a strong relationship between fertility and vital status is assumed (e.g. $R^{ds} = 2$), the bias will be very small (0.5%):

$$\frac{\text{PF}^{ds}}{\text{PF}^s} = \frac{2}{0.01 + 2 \times 0.99} = 1.005. \tag{4.25}$$

Migration differential

The number of children ever born collected in the censuses, which is further compared to the current fertility in the P/F ratio method, refers to the fertility experience of all people enumerated in a certain geographic area, including the immigrants. Similarly to what occurs with fertility differentials by vital status, the P/F ratio may be biased if there is significant difference in fertility by migration status.

To illustrate this effect, imagine that if immigrants come from a region with high adolescent fertility, they will have higher parity than the cumulated period fertility in the region of destination, and the P/F ratio will then be overestimated.

The average parity of women at the moment of the interview reflects the fertility experience of both immigrant and non-immigrant women, and is given by:

$$P^{mn} = \frac{\text{CEB}^{mn}}{K^{mn}} \tag{4.26}$$

The same quantify can be defined for non-immigrant and immigrant women separately:

$$P^{n} = \frac{\text{CEB}^{n}}{K^{n}} \tag{4.27}$$

$$P^{m} = \frac{\text{CEB}^{m}}{K^{m}} \tag{4.28}$$

where CEB represents the number of children ever born and K the population of women. The indices n, m and mn represent the non-immigrants, the immigrants and both groups combined, respectively.

Let

$$\pi^{n} = \frac{\text{CEB}^{n}}{\text{CEB}^{n} + \text{CEB}^{m}}, \text{be the proportion of births from non-immigrant women} \tag{4.29}$$

$$\pi^{m} = \frac{\text{CEB}^{m}}{\text{CEB}^{m} + \text{CEB}^{m}}, \text{be the proportion of births from immigrant women} \tag{4.30}$$

Suppose P^n and P^m differ by a factor R^{mn}:

$$R^{mn} = \frac{P^{m}}{P^{n}} \tag{4.31}$$

Following the same approach used in the previous section, let PF^{mn} be the adjustment factor of the age group 20-24 calculated by the application of the Brass method and PF^{n} be the adjustment factor of the non-immigrants, which will be more consistent with the recent fertility experience of the enumerated population in a certain region. Again, since the denominator of the P/F ratio is the same, the only difference between PF^{n} and PF^{mn} is in the numerator (the average parity): P^n and P^{mn}.

Thus, the ratio between PF^n and PF^{mn}, which indicates the bias in the P/F ratio method due to the unmet condition of equal fertility according to the migration status, is given by:

$$\frac{PF^n}{PF^{mn}} = \left(\frac{\pi^m + R^{mn}(1 - \pi^m)}{R^{mn}} \right). \tag{4.32}$$

When $R^{nm} = 1$ or when $\pi^m = 0$, $PF^n = PF^{mn}$, meaning that there is no bias in the result of the method.

Similarly to what happens with mortality, this result shows that two conditions need to be simultaneously true to the unmet condition of independence between fertility and migration have an impact on the results: i) fertility of immigrant women differs significantly from the non-immigrants; and ii) the proportion of births from immigrant women is considerable.

Contrary to the mortality analysis, in which the proportion of woman who die at the beginning of the reproductive period is low, the proportion of immigrant at these ages, which in turn have different fertility rates, may be significant in many contexts, including subnational levels in Brazil.

It is not possible to derive from Brazilian censuses the number of births each woman had in the regions of origin or the region of destination. In order to approximate the quantities R^{mn} and π^m, the immigrants are defined as the woman aged 20-24 who migrated less than 3 years before the census date. This groups would reflect the average fertility experience of woman who had their children outside the region of enumeration.

Some "immigrant" woman could have had their children recently in the region of destination, as well as the "non-immigrant" could have have their children more than 3 years ago in another region. However, the average parity and number of births provides a good approximating for illustrative purposes.

Figure 4.14 shows that the bias caused by the hypothesis of migration is higher than that caused by differential fertility by vital status, although it is still relatively low. The P/F ratio method overestimates fertility by slightly less than 2% in states such as DF and ES. In DF, this is mostly due to high proportion of immigrant women, whereas in ES the bias is mostly caused by the high fertility differential between immigrants and non-immigrants. The method underestimates fertility in a few states, such as RR.

Fertility change

The last condition assessed in the sensitivity analysis is that of constant fertility. This is thought to be the strongest condition required by the method (Moultrie and Dorrington, 2008; UN, 1983).

In general, the condition of the method is that ASFR have remained constant over time. In fact, if the age group 20-24 is used for calculating the adjustment factor, the requirement is that fertility for the two first age groups has remained constant in the recent past (see Equation 4.16). It is common that fertility for the young women remains roughly constant in the beginning of the fertility transition period, despite rapid decline in old woman fertility.

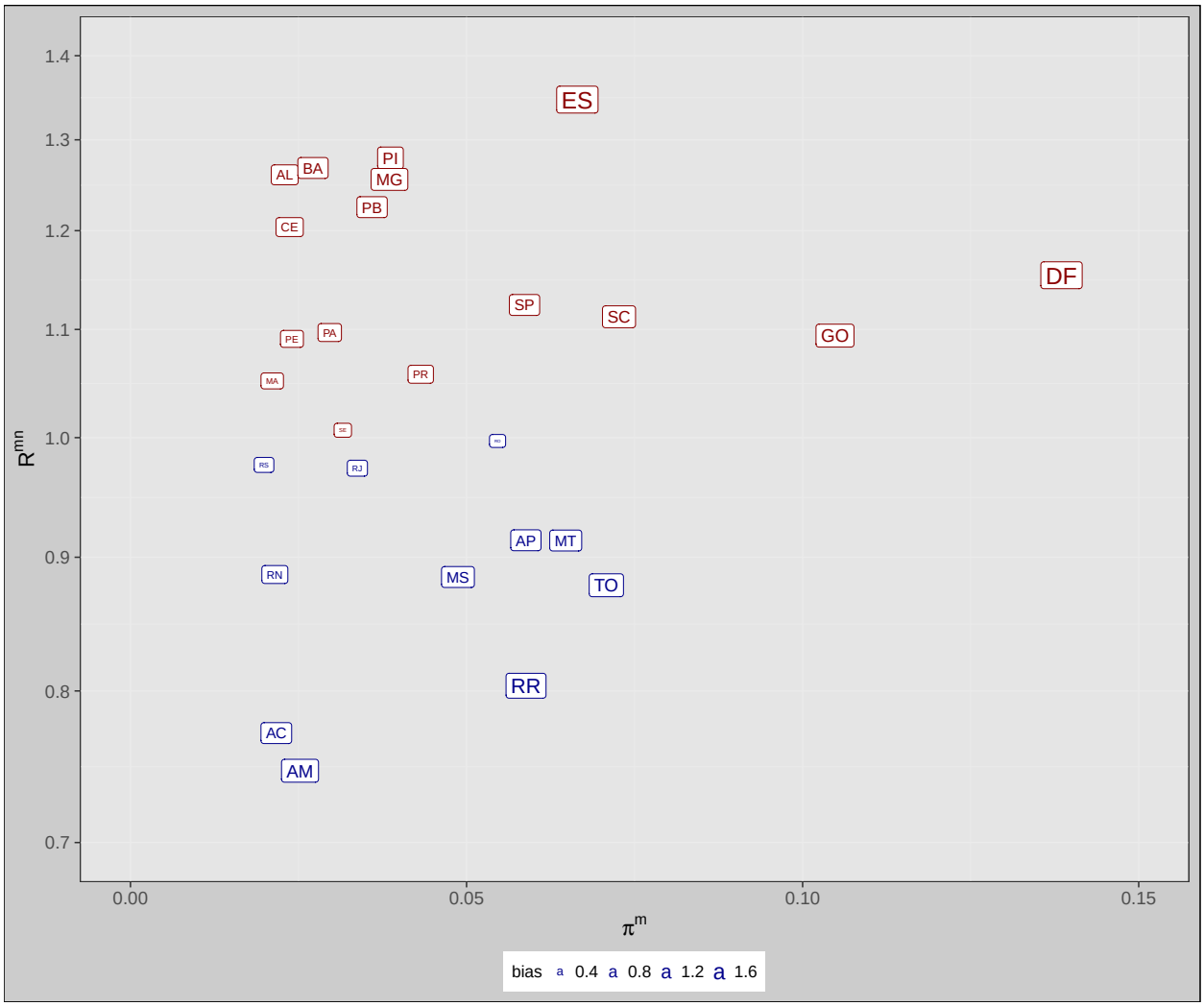

Figure 4.14: Sensitivity of the P/F ratio method to migration, Brazilian states, 2010. Source: IBGE, 2010 Census

In Brazil, as shown by Figure 4.15, adolescent fertility rates (15-19) increased in the periods 1980-1991 and 1991-2000, despite the rapid overall fertility decline. Fertility for the age group 20-24 reduced in the same period, but at a slower pace than the observed for older ages. Thus, the almost constant fertility for the two younger age groups (increase in $ASFR_{15\text{-}19}$ offset by a moderate decline in $ASFR_{20\text{-}24}$) between 1980 and 2000, allowed the use of the P/F ratio method without marked biases (Carvalho, Gonçalves, and L. Silva, 2018). On the other hand, $ASFR$ for these two age groups reduced more than 30% between 2000 and 2010, imposing a strong limitation to the use of the P/F ratio method for this period.

This issue has been a matter of debate in the past years. Castanheira and Kohler, (2015) identified the problem of the basic assumption violation for Brazil in 2010 and recommend that the method be possibly discontinued. Cavenaghi and Alves, (2016) also recognize the limitations, but argue that the method still presents reasonable results. Carvalho, Gonçalves, and L. Silva, (2018) discuss several alternatives. The final author's recommendation is the use of the adjustment factors for the year 2000 for 2010, with fertility estimated for the period of reference at about 2.5 years prior to the census.

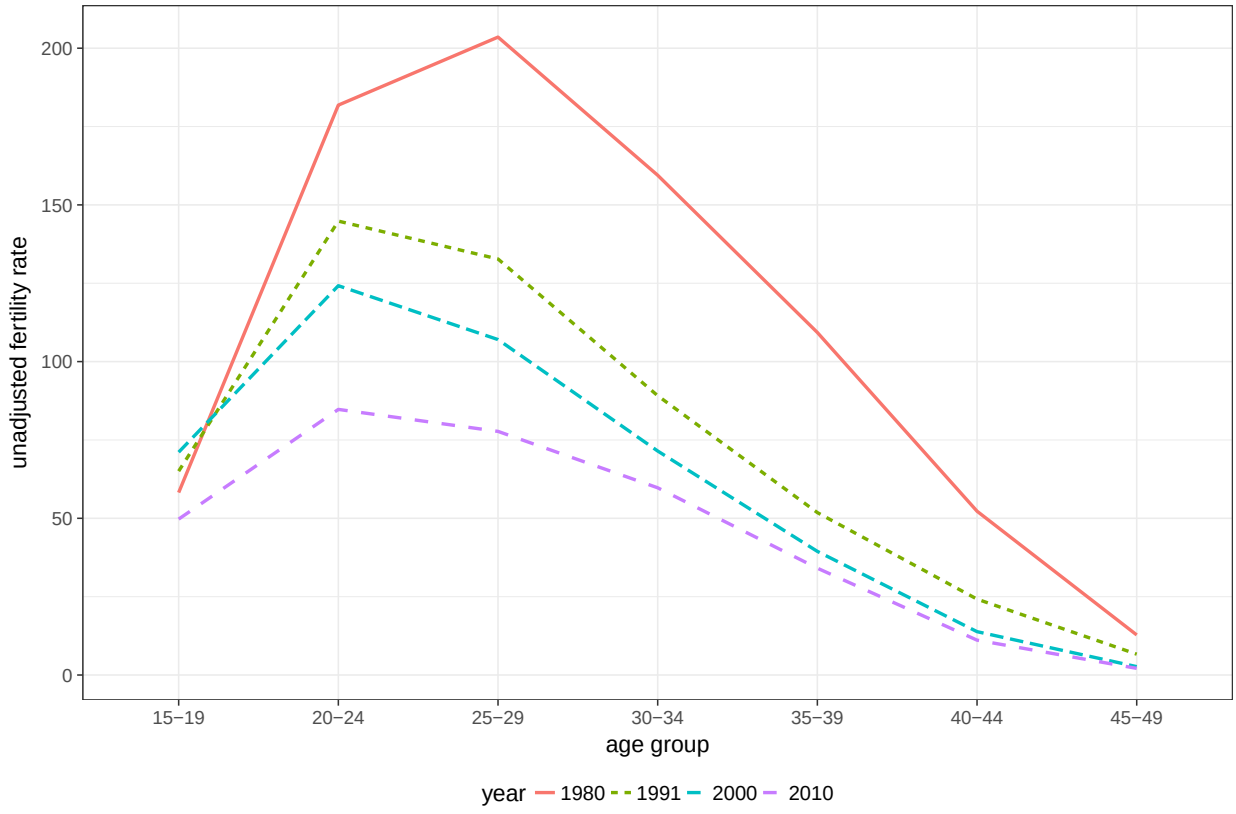

Figure 4.15: Map of the TFR by state, for the years 1980, 1991, 2000 and 2010. Source: IBGE, Censuses of 1980, 1991, 2000 and 2010

The P/F ratio estimated by the Brass original method (PF^B) can be decomposed into two factors: i) the adjustment factor to correct the reported current fertility rates ($\widehat{\text{PF}}$), which is what the method is trying to estimate; ii) a factor that estimates the effect of fertility change in the P/F ratio resulting from comparing cohort and period fertility (PF^{fc}). If fertility is constant over time, this second factor is 1.

$$\underbrace{\text{PF}^B}_{\substack{\text{P/F ratio} \\ \text{estimated by Brass} \\ \text{original method}}} = \underbrace{\widehat{\text{PF}}}_{\substack{\text{underreport of} \\ \text{current fertility} \\ \text{effect}}} \times \underbrace{\text{PF}^{fc}}_{\substack{\text{fertility change} \\ \text{effect}}} \tag{4.33}$$

As previously discussed, the P/F ratio method assume constant fertility up to age i used for the calculation of the adjustment factor, so that $f_x(t)$ is constant over time and the P/F ratio is given by:

$$\text{PF}_i(t) = \frac{P_i(t)}{F_i'(t)} = \frac{\int_0^i f_x(t)\,dx}{\int_0^i f_x'(t)\,dx} \tag{4.34}$$

The second factor of Equation 4.33 calculates the effect of fertility change in the estimate of the P/F ratio method and is given by the ratio of the observed parity and the true period fertility rates:

$$\text{PF}_i(t) = \frac{P_i(t)}{F_i(t)} = \frac{\int_0^i f_x(t-i+x)\,dx}{\int_0^i f_x(t)\,dx} \tag{4.35}$$

Following a similar approach used by Schmertmann, Cavenaghi, et al., (2013) to produce retrospective fertility estimates, let $f_x(t-i+x)$ be defined in terms of current fertility rates $f_x(t)$ and multipliers $\rho_x(t)$ that relate current (period) and retrospective (cohort) fertility rates, as follows:

$$\rho_x(t) = \frac{f_x(t-i+x)}{f_x(t)} \tag{4.36}$$

$$\text{PF}_i(t) = \frac{P_i(t)}{F_i'(t)} = \frac{\int_0^i \rho_x(t) f_x(t)\,dx}{\int_0^i f_x(t)\,dx} \tag{4.37}$$

This ratio is the average ratio of past to present fertility rates, weighted by current fertility rates. This factor can be calculated if the past and present ratios are available.

However, this is rarely available in contexts where this method needs to be applied. Thus, further approximations are necessary to transform equation 4.37 into parameters that can be estimated more easily.

Equation 4.37 can be approximated further by the ratio between past to present fertility at the mean age of childbearing up to age i, μ_i at time $t-(i-\mu_i)$, where $(i-\mu_i)$ indicates the number of years prior to the census the experience of the cohort i refer to:

$$\mathrm{PF}_i(t) = \rho_{\mu_i}(t-i+\mu_i) \tag{4.38}$$

For example, assuming that $(i-\mu_i) = 2.5$ for the cohort aged 20-24 in 2000, $\mathrm{PF}_{20-24}(2000) = \rho_{17-21}(2000-2.5)$, meaning that it refers to the ratio between fertility rates of the cohort 17-21 in 1997.5.

A calibrated spline estimation procedure that interpolates detailed fertility schedules from age-group data (Schmertmann, 2014) is used to calculate $(i-\mu_i)$, the average time since previous births. The ratio between fertility rates of the cohort, $\rho_{\mu_i}(t-i+\mu_i)$, is estimated based on the growth rate of the reported fertility rates between the census under analysis and the preceding census.

Section 4.3 shows the results of the application of this adjustment to Brazilian data from 1991 to 2010.

Omission of fertility among women under age 15

Finally, the application of the P/F ratio method often uses the traditional age groups of woman at reproductive ages, from 15 to 49. However, if fertility below age 15 is relatively high, this may bias the results.

The fertility of women below age 15 are reported in the cumulated parity P_i, but this will not be taken into account if this group is not included in the calculations for the current fertility F_i. Thus, the adjustment factor will be overestimated.

For example, the fertility rates below age 15 reported in the 2010 Census in Brazil in a few states represents around 2% of the fertility rates up to age 24. This means that the P_{20-24}/F_{20-24} factor would be overestimated by 2%.

Fertility estimates for Brazil and states from 1980 to 2010 using the P/F ratio method

As sections below indicate, the only condition that can lead to important biases in the P/F ratio method in the Brazilian context is that of fertility change. Thus, the complete sensitivity framework derived in this section can be used in other contexts for to adjustment the final results of the methods. In this study, however, adjustments are made considering only this condition.

Figure 4.16 shows the comparison between adjusted an unadjusted TFR estimate by using the P/F ratio method. The figure shows that the only year when the adjusted TFR

estimated by the method described above differ significantly from the original proposition is 2010. The adjusted rate consistently reduces fertility estimates for 2010. Is a few other cases, this occurs for other years as well, such as RR in 2000. For 1980, since there is no information on past fertility rates, both estimates are equal.

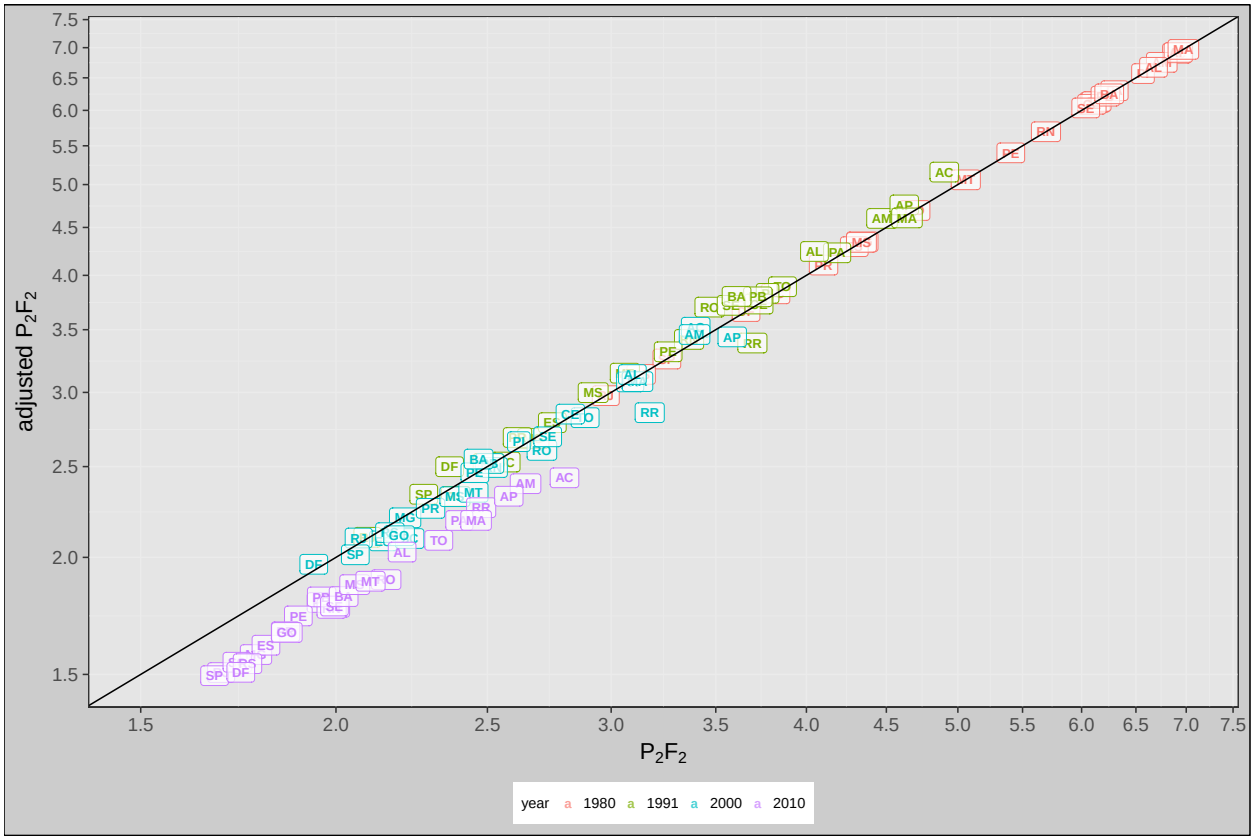

Figure 4.16: Comparison between adjusted an unadjusted TFR, Brazilian states, 1980, 1991, 2000 and 2010. Source: IBGE, Censuses of 1980, 1991, 2000 and 2010

4.17 shows the TFR by state, for the years 1980, 1991, 2000 and 2010 calculated by the P/F ratio method, using the adjustment ratio for the age group 20-24 (P2/F2).

The map indicates that fertility has declined steadily from 1980 to 2010. The TFR for Brazil, calculated by the P/F ratio method declined from 4.36 in 1980, 2.88 in 1991, 2.35 in 2000 and 1.71 in 2010. The regional differentials remain, despite the generalized fertility decline in all states.

Conclusion

Brazil has made impressive improvements in the quality of its CRVS systems in regard to birth counts since the 1990s. These improvements impose a greater use of these data to con-

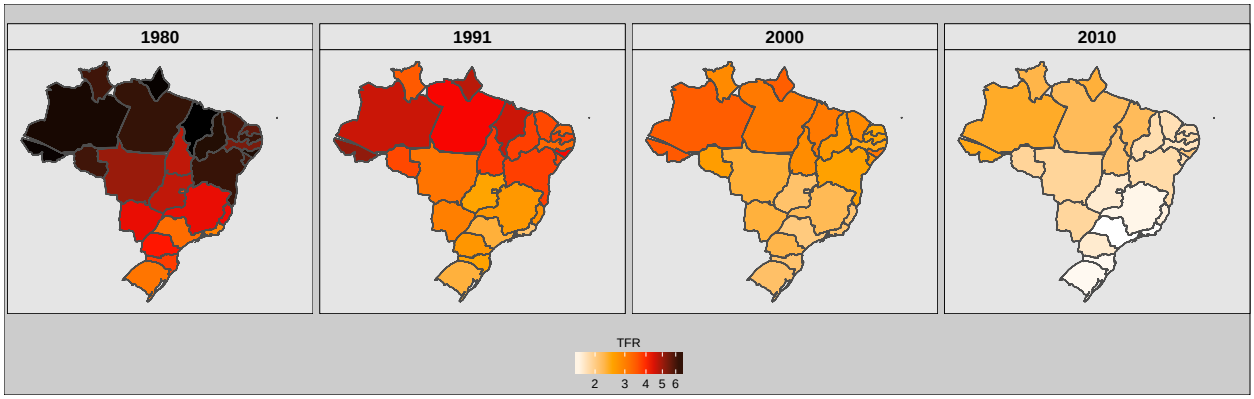

Figure 4.17: Map of the TFR by state, for the years 1980, 1991, 2000 and 2010. Source: IBGE, Censuses of 1980, 1991, 2000 and 2010

duct fertility estimations, which have been done traditionally by using indirect demographic techniques.

On the other hand, indirect demographic techniques remain relevant for many purposes. First, this is essentially the only way to estimate fertility for periods when CRVS were still incomplete. Second, despite improvements, administrative data still have problems in less developed regions, even for recent periods. Third, indirect demographic techniques, such as the P/F ratio method, have the advantage of using information from the same data source to estimate fertility. The use of CRVS also requires an estimated population of women at reproductive ages, which normally come from censuses. If the completeness of the census is different from the completeness of registered births, then fertility estimate will be biased. Finally, the fact that these questions are in censuses and surveys, which tend to have rich questionnaires, fertility for different population groups, e.g. education, income, migration and marital status, can be estimated.

This section presented an evaluation of different methods and data sources to estimate fertility. Based on a sensitivity analysis it also proposes an adjustment for estimating fertility in the context of adolescent fertility decline.

4.4 Mortality Estimates for Brazil and States from 1980 to 2010

The study of mortality is relevant for many purposes. First, death rates at all ages form indicators that are relevant for understanding population health. Mortality rates are also fundamental part of the demographic dynamics, playing an important role in shaping the population age structure, particularly at old ages.

In Brazil, mortality estimates have been also relevant for targeting population groups and prioritizing actions and investments necessary to face and detain infant mortality (M. Costa et al., 2001) and non-communicable chronic diseases in adults (Malta et al., 2011). Furthermore, official estimates of life expectancy, published annually by IBGE, have also been used as one of the parameters to define social security benefits, through an adjustable formula (*"fator previdenciário"*). Inequalities in the life expectancy between different social groups and regions have been subject of much debate on the inequalities of the Brazilian pension system and the proposed pension system reform, as well (A. T. R. Oliveira, 2017; Ribeiro and Fígoli, 2008).

Likewise fertility, CRVS should be the main data source for mortality estimation. However, there are also coverage and quality problems with them in developing countries. A relatively small number of this countries have high quality registration of deaths and census, imposing difficulties to mortality estimation in these contexts (Timæus, Dorrington, and Hill, 2013).

There are numerous proposals to overcome these limitations. Indirect demographic techniques based on questions in censuses and surveys and those that make use of the coherence between demographic parameters have prevailed.

Demographers have used methods for mortality estimation that either estimate mortality rates independently from registered deaths or that estimate the completeness of registered deaths, which are further corrected so that mortality rates can be calculated.

Methods for calculating mortality rates without making use of registered deaths often treat infant and adult mortality separately.

Indirect methods for estimating infant mortality are normally based on reports of mother about their children and their surviving status. They can be based on information about the *full birth history* of woman, where they are asked about live birth, the date of the birth, and, if the child has died, their age at death, or only aggregate information on children ever born and surviving children, as a *summary birth history*. Full birth histories are usually collected in detailed household surveys, such as the DHS, whereas summary birth histories are oftentimes also included in censuses as they require a reduced number of questions (Hill, You, Inoue, et al., 2012; R. Silva, 2012)

Methods for estimating adult mortality without using registered deaths vary significantly in terms of methodology and data required. The essence of these methods is to derive mortality measures based on questions to survey and census respondents about close relatives, household member or other network relationships, such as friends. The leading approach on

this matter is the sibling survival method, which asks respondents questions about their sibling, including their vital status (Gakidou and King, 2006; Masquelier, 2012; Trussell and Rodríguez, 1990 Nov-Dec). Adult mortality can be also estimated based on information about orphanhood and widowhood (UN, 1983). Questions on recent deaths in the household have been also used to estimate mortality at all ages (Queiroz and Sawyer, 2012).

There has been considerable disagreement about mortality estimates in Brazil, particularly for subnational levels. Mortality estimates using different methods and data sources have led to inconsistent results. Schmertmann and Gonzaga, (2018), for example, have recently produced mortality estimates that differ enormously from the official estimates published by IBGE (IBGE, 2013c). Schmertmann and Gonzaga, (2018) find, in general, higher life expectancy than IBGE in the more developed states and lower life expectancy in the less developed states. They also find higher variability in the less developed states, and much narrower prediction interval for the more developed states. The main methodological procedures of both sources are discussed below, as well as their main limitations and possible sources of the divergences.

Official estimates of infant mortality have also been challenged for the last two decades. Inconsistencies in estimates produced by IBGE of infant mortality have been reported (Szwarcwald, 2008; Szwarcwald, Leal, et al., 2002; Szwarcwald, Morais Neto, et al., 2010), that uses indirect demographic techniques, such as the surviving children Brass method.

This section discusses different possibilities for estimating mortality and completeness of death counts in Brazil, to be further used in combination in the integrated model for population estimates.

Infant and child mortality

Infant and child mortality in Brazil have been historically estimated by using indirect demographic methods, since the quality of information on infant deaths and live births were poor until very recently.

Vital statistics have improved significantly in the last two decades and official estimates of infant mortality, and particularly the indirect demographic methods used, have been challenged. (Szwarcwald, 2008; Szwarcwald, Leal, et al., 2002; Szwarcwald, Morais Neto, et al., 2010).

If reliable CRVS systems are available, infant mortality is calculated simply by taking the ratio between the infant deaths in a year and the live births in that same year.

$$\text{IM(t)} = \frac{B(t)}{D_0(t)} \tag{4.39}$$

where IM(t) is the infant mortality rate for year t, $B(t)$ is number of live births in year t and $D_0(t)$ is the number of infant deaths in the same year.

As discussed in the previous section 4.3, Brazil has currently two administrative record systems that collects information of birth counts: the CR and the VS, which have improved,

but still have problems such as under-registration and late registration, particularly in less developed regions.

Similarly, two systems are available for collecting death counts data: CR and the VS, called SIM. These also have coverage and quality problems, and since completeness of registered infant deaths is usually lower than completeness of registered live births, infant mortality calculated directly in contexts of lack of comprehensive CRVS systems is usually biased downwards.

The "proactive search" survey, discussed in the previous section (4.3), was carried out to search for unregistered vital events with the main objective of calculating a measure of infant mortality alternative to indirect demographic methods. The following section briefly discusses the main findings of the study, as published by (RIPSA, 2013) [10].

Proactive Search estimates of underregistration of infant deaths and infant mortality

The main idea of the estimation procedure used by the "proactive search" study is to adjust live births and infant death counts and then calculate infant mortality based on equation 4.39.

Figure 4.18 shows the under-registration of infant deaths (below age 1) by state for the years 2000 and 2010. The map shows a remarkable improvement in the completeness of the SIM for infant deaths. The underregistration of infant deaths of the SIM reduced from 26% to 15% in ten years.

The map also shows the great regional disparity in the country. Whereas several states already had near complete report of infant deaths in 2000, in states like MA only 1/3 of infant deaths were reported. In PI, less than 50% of the deaths were reported in the same year. Regional inequalities persist in 2010, and 35% of infant deaths were still not registered in MA, but differences have reduced.

When Figure 4.18 is confronted with the map that show the underregistration of live births (Figure 4.13), it becomes clear that, in fact, infant deaths are more underreported than births.

These estimates lead to the infant mortality rates shown in Figure 4.19. According to these estimates, Infant Mortality Rate (IMR) declined from 26‰ to 16‰, an impressive decline of almost 40% in only 10 years.

As expected, the poorest states, mostly in the North and Northeast regions, have had the highest infant mortality rates. In 2000, all the states in the Northeast had IMR higher than 30‰, whereas in the more developed states IMR was close to 15‰. MG and RJ were the states with the highest IMR in the South and Southeast regions for both years. Ceará (CE), PE and Rio Grande do Norte (RN), all from the Northeast region, are examples of states that have reduced IMR more than the national average and thus changed position in the ranking by state of this indicator.

[10] For a brief methodological description of the study, refer to section 4.3

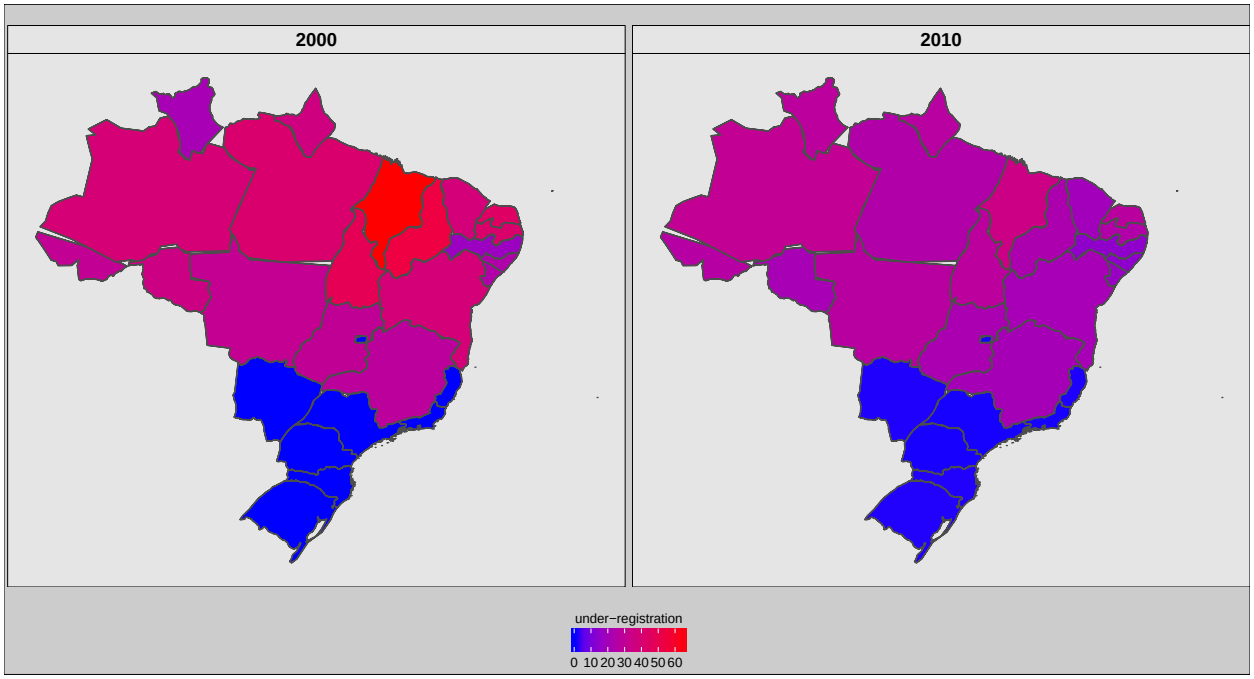

Figure 4.18: Map of the under-registration of infant deaths by state estimated by the Proactive Search survey, 2000 and 2010 (in %). Source: RIPSA, (2013)

Brass method for estimating child mortality

In contexts of lacking vital statistics on births and infant deaths, infant and child mortality have been estimated indirectly. Child mortality is often estimated separately from mortality at other ages. First, child mortality has gotten special attention since it is recognized as an important summary indicator of the overall socioeconomic and health status of the population. Second, data and methods used to estimate infant mortality tend to be different from those used to estimated adult mortality.

The Brass method is the leading approach for indirect infant and child mortality estimation. The basic idea of the method is to estimate infant mortality based on the proportion of surviving children among the children ever born by age of the woman. The information reported by young woman would reflect the infant mortality for a more recent period, whereas the older mothers would report the average mortality experience of their children up to a longer period. (Brass, 1964; Brass and Coale, 1968; UN, 1983).

The information required for the use of the method is part of the *summary birth history*, which contains data on children ever born and surviving children. Information on children ever born is often collected by questions such as: *"How many children born alive have you had?"*. This question is usually followed by question on the number or children alive or dead at the reference date of the census, such as *"How many children born alive have died before the census date?"* or *"How many of the born alive children are still alive?"*. Brazilian

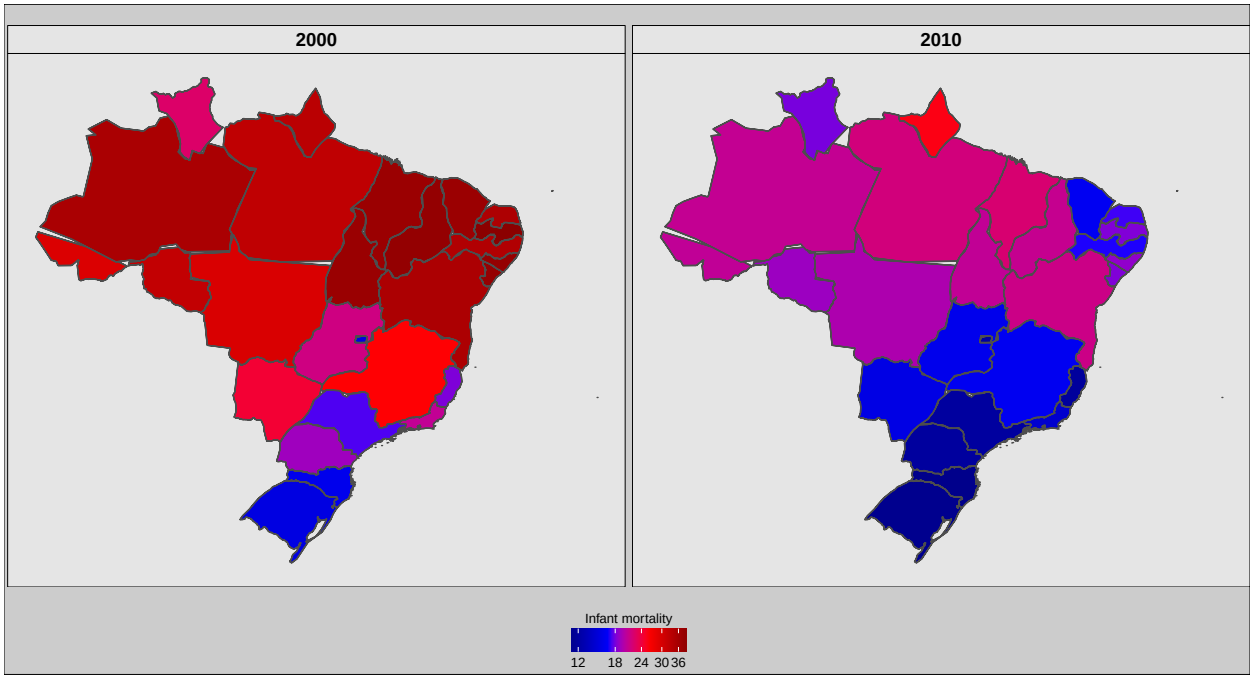

Figure 4.19: Map of infant mortality by state estimated by the Proactive Search survey, 2000 and 2010 (per thousands). Source: RIPSA, (2013)

censuses have had these question since 1940. There have been a few changes in the way this information is collected, mostly to follow recommendations on how to reduce error, for example by asking specifically separately about the number of children by sex or by asking specifically about those who still live in the household or elsewhere.

Information for very young women is thought to be more subject to reporting problems. Furthermore, children of young mothers experience higher mortality. Since the proportion of births from this cohort is usually small, this group is often disregarded in the analysis (UN, 1983).

With this information, the proportions of surviving children by age of the mother at different ages provide mortality estimates between birth and several ages. Brass, (1964) found that the proportion of children dead and the probability of dying up to a certain age is strictly related to the fertility schedule.

For example, the proportion of children dead for the age group 15-19, corresponds approximately to the probability of dying between birth and age 1, ${}_1q_0$. The following six age groups correspond to the probably of dying from birth up to the ages 2, 3, 5, 10, 15, 20 respectively. Since these are only approximations, adjustment factors based on the parity of the first age groups are used (Brass, 1964; Hill, 2013; UN, 1983).

In practice, $P_i(t)$ is given by the average parity reported in the census at time t by women aged i:

$$P_i(t) = \frac{\text{CEB}_i}{K_i(t)} \tag{4.40}$$

where CEB_i is the number of children ever born to women aged i and K_i is the number of women from the age enumerated in the same census.

The proportion of children dead, $D_i(t)$, is given by:

$$D_i = \frac{\text{CD}_i}{\text{CEB}_i} \tag{4.41}$$

where CD_i is the number of children dead born to women aged i and CEB_i is the number of children ever born to women from the same age group in the same census.

The method has thus two main assumptions: fertility and childhood mortality have remained constant in the recent past. If fertility has been changing, the ratios of average parities obtained from the census will not represent the experience of any cohort of women and will not provide a good idex of the distribution in time of the births to the women of each age group (UN, 1983).

Figure 4.20 shows the IMR estimated for Brazil from 1980 to 2010 by using the Brass method. One of the main advantages of this technique is that it can combine information for different censuses to produce trends. This also allows for checking the consistency between estimates from different censuses. The figure shows that estimates from different censuses tend to be consistent. It also shows steep decline in infant mortality over the period under analysis.

Figures 4.21 and 4.22 show the estimated IMR by using the Brass method for the five-year periods between 1980 and 2010. The two sets of maps are in different scales to facilitate visualization of regional differentials.

The results confirm the rapid IMR decline in all Brazilian states since the 1980s. The results of this methods seem to indicate a greater convergence in infant mortality than the indicated by the "proactive search" results. It also indicate lower levels of infant mortality for the recent past.

The main assumptions of the method (constant fertility and mortality) are violated for the entire period under analysis. Future work should conduct a sensitivity analysis to evaluate to what extend these violations may be biasing the results.

Adult Mortality

The estimation of adult mortality can be performed by estimating mortality indicators directly based on questions in censuses and surveys, or by adjusting the death counts and then calculating mortality.

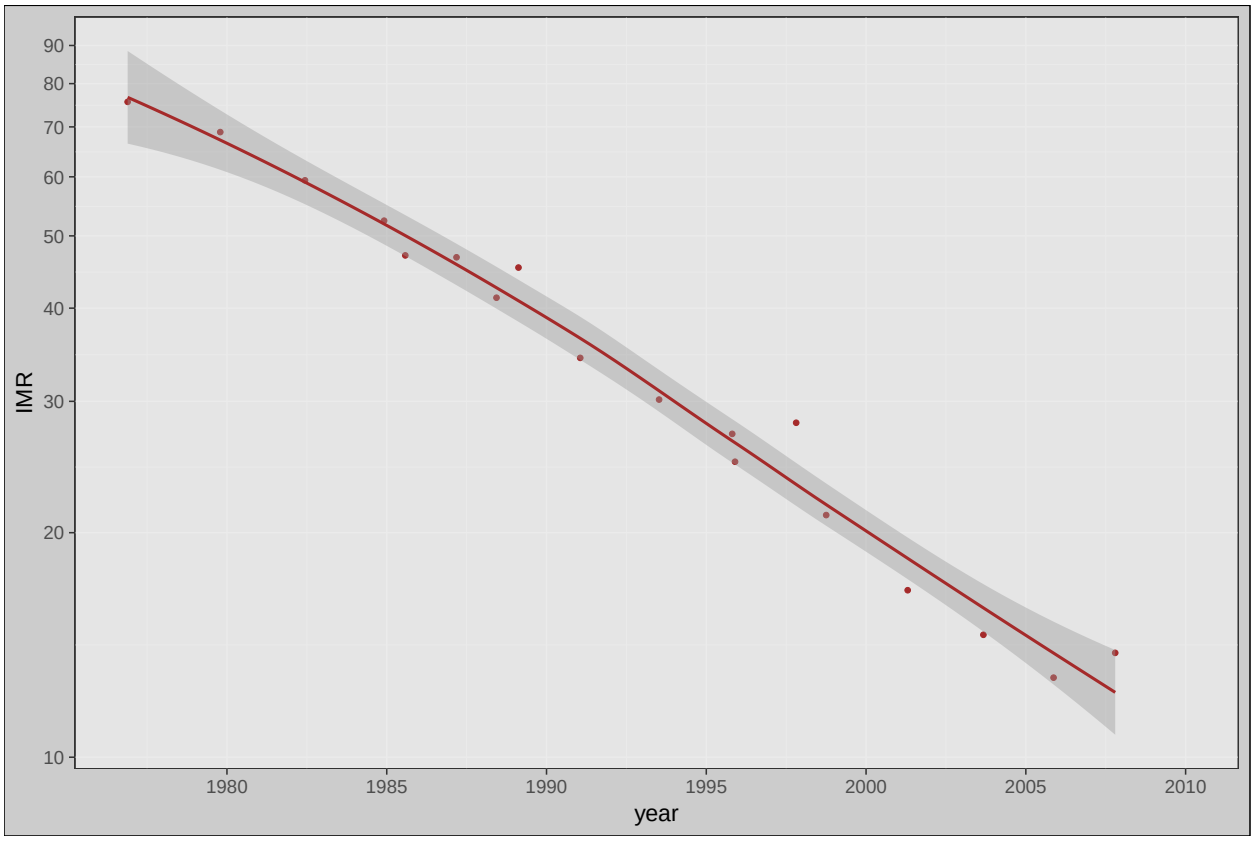

Figure 4.20: Infant mortality rate by year estimated by Brass method, (per thousand). Source: 1991, 2000 and 2010 censuses

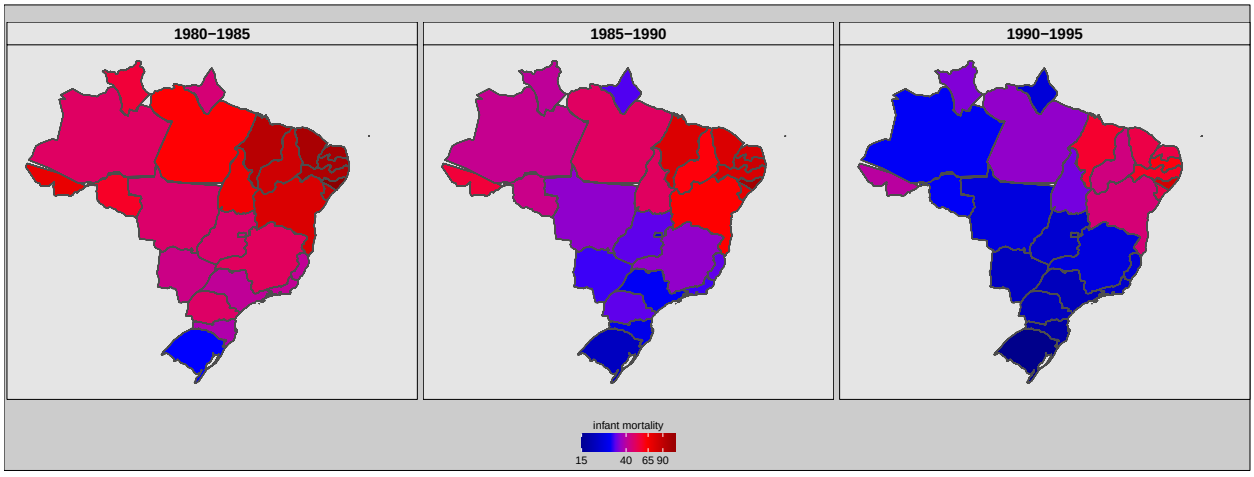

Figure 4.21: Map of infant mortality by state estimated by Brass method, 1980-1995 (per thousands). Source: 1980, 1991, 2000 and 2010 censuses

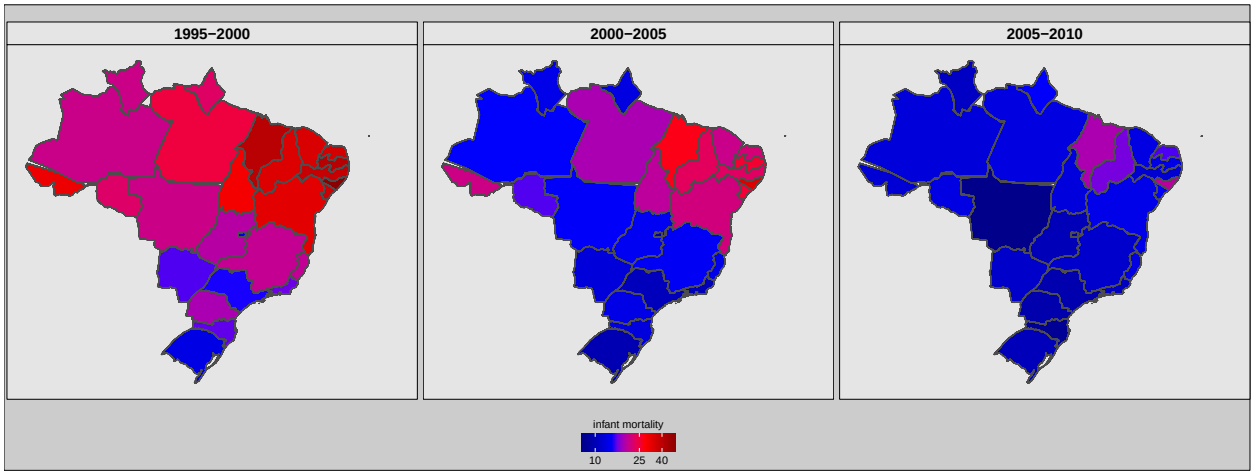

Figure 4.22: Map of infant mortality by state estimated by Brass method, 1995-2010 (per thousands). Source: 1980, 1991, 2000 and 2010 censuses

Evaluation of registered deaths

Registered death counts are subject to coverage and quality problems. Coverage refers to underregistration and late registration, whereas quality problems refer to errors in the information of registered deaths, mostly related to the reporting of age, such as unknown age and age misstatement.

Figure 4.23 shows the proportion of unknown age in the deaths informed to the SIM by year and state, from 1979 to 2016. The proportion of unknown ages in Brazil was less than 1% up to 1989. This proportion increased in 1990 and remained relatively high until 1998. It has declined consistently since then.

The trend for the states has followed the same pattern. The proportion of deaths of unknown ages seem to have no correlation with the socioeconomic status of the state. For example, RJ is the state with the highest proportion of unknown in the last decade. The DF is another example of a rich state with high proportion. On the other hand, AP, AL, Sergipe (SE), AC are poor states in the North and Northeast regions that have low proportion of unknown reported ages.

Figure 4.24 shows the pyramids of the registered death counts (excluding deaths at age 0) by sex for the years 1980, 1990, 2000 and 2010. The median age at death has increased, with more deaths concentrated at older ages, as a result of population aging and mortality decline. It is clear some age heaping for ages ending in 0 and 5 for the years 1980 and 1990. Age heaping reduced in 2000, appearing mostly at ages 60 and 70. In 2010, this problem nearly disappears.

Section D.1 in the Appendix D shows the population pyramids of the death counts for all states for the years 1980, 1990, 2000 and 2010.

Age heaping in death counts for Brazil and the states can be summarized by measures

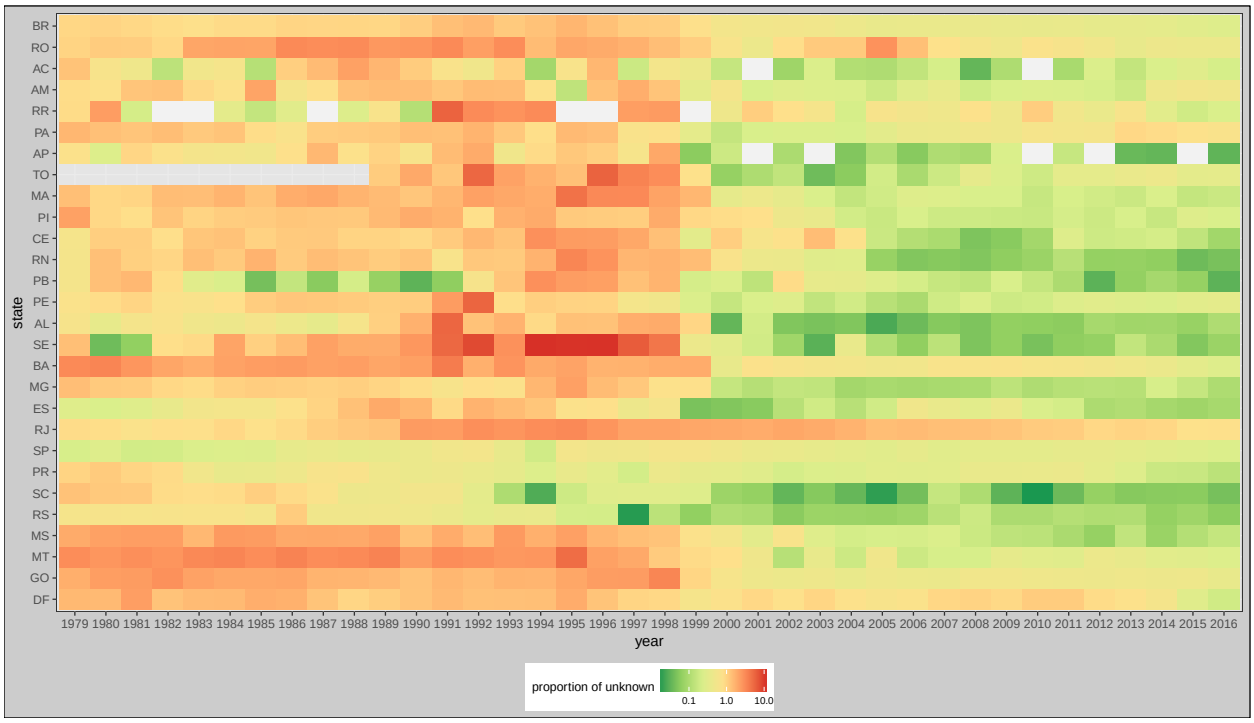

Figure 4.23: Heatmap of the proportion of unknown age in the registered deaths by year and state, 1979-2016 (in %). Source: SIM

such as the Spoorenberg Index, described in Section 4.2. Figure 4.25 shows the trend of the Spoorenberg Index for the entire series of the SIM. The great improvement in the age reporting of deaths took place between 1980 and mid 1990s. It has improved since then, but the index already at low levels. There has been more pronounced attraction for terminal digits among females than males.

The Spoorenberg Index by state is shown in Figure 4.26, which confirms the higher indices for females. The figure also shows the overall improvement for all states. Furthermore, the graph shows that the age heaping problem has been concentrated in states of the North and Northeast regions.

Another way to assess the quality of the VS system (SIM) is by comparing the number of registered deaths with those informed to the CR system. Figure 4.27 shows the ratio between the total registered deaths in the CR and the VS, by year, sex and state. Most of the states in the South and Southeast regions have had, since the beginning, very similar figures in both systems. For the other states, the CR reported more deaths until 2000 and since then the VS have had higher coverage.

Figure 4.28 shows the same ratio for Brazil, disaggregated by age. This analysis is relevant to assess the hypothesis often adopted that the underregistration is nearly constant after a certain age. The figure shows that at least one of the systems violates the assumption of

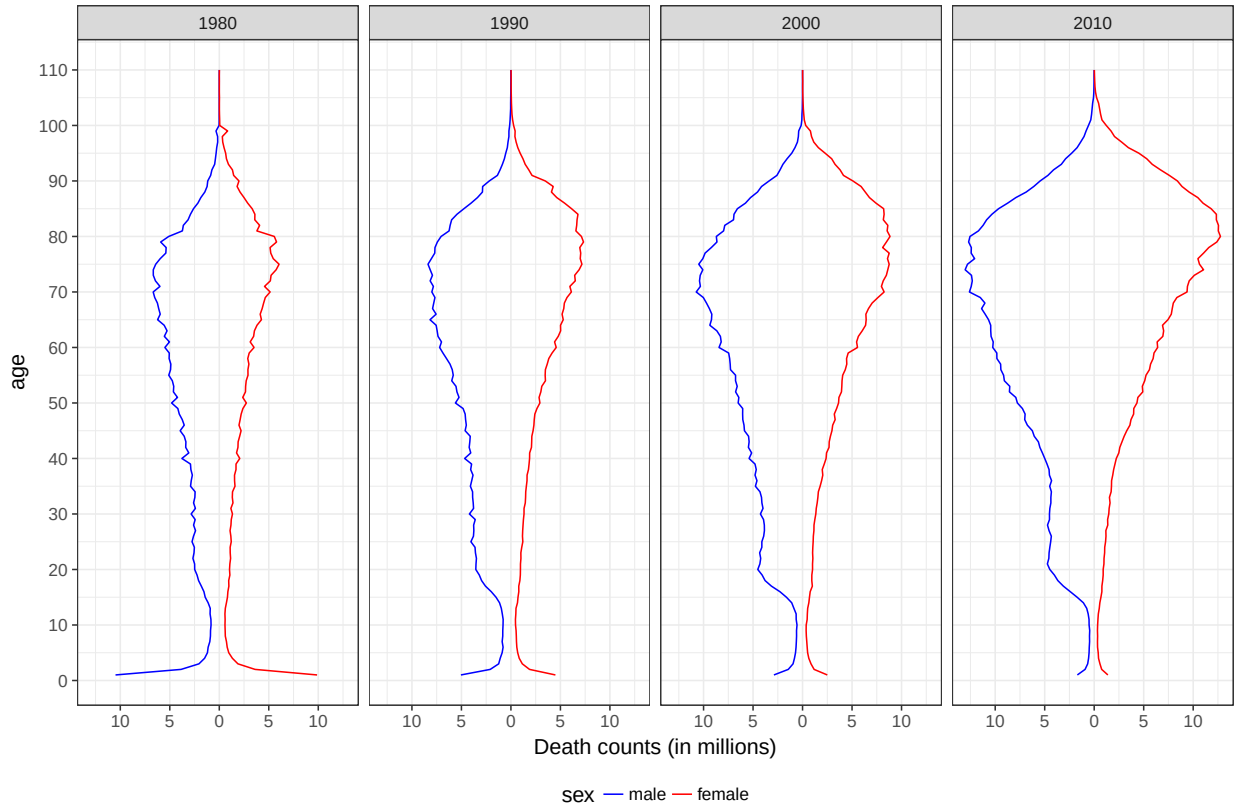

Figure 4.24: Pyramid of the registered death counts by year and sex, 1980, 1990, 2000 and 2010. Source: SIM

constant completeness by age. The CR has higher coverage at old ages, whereas the VS captures more deaths at younger ages, particularly children.

In addition to this descriptive analysis, there are other two ways to evaluate the completeness of registered deaths: direct and indirect.

The same studies that estimated completeness of birth counts, did it for death counts as well, by using the same methodology (see section 4.3).

Figure 4.29 shows the under-registration of deaths by state according to both the (CR and VS) for 2015. These were calculated by using the capture-recapture study described in section 4.3 Trindade, L. F. L. Costa, and A. T. R. Oliveira, (2018). The results show that there are still limitations in both systems and regional inequalities in the registration of births persist. The under-registration of deaths in the CR is higher than that of the VS, particularly for the states in the North region, and a few in the Northeast, such as MA and PI. The explanation for that is the same as for the births, that is there difficulties to access the registry offices in these states.

The "proactive search" survey also estimates the underregistration of adult deaths in the SIM, which are shown in Figure 4.30. Again, the less develop states have the highest

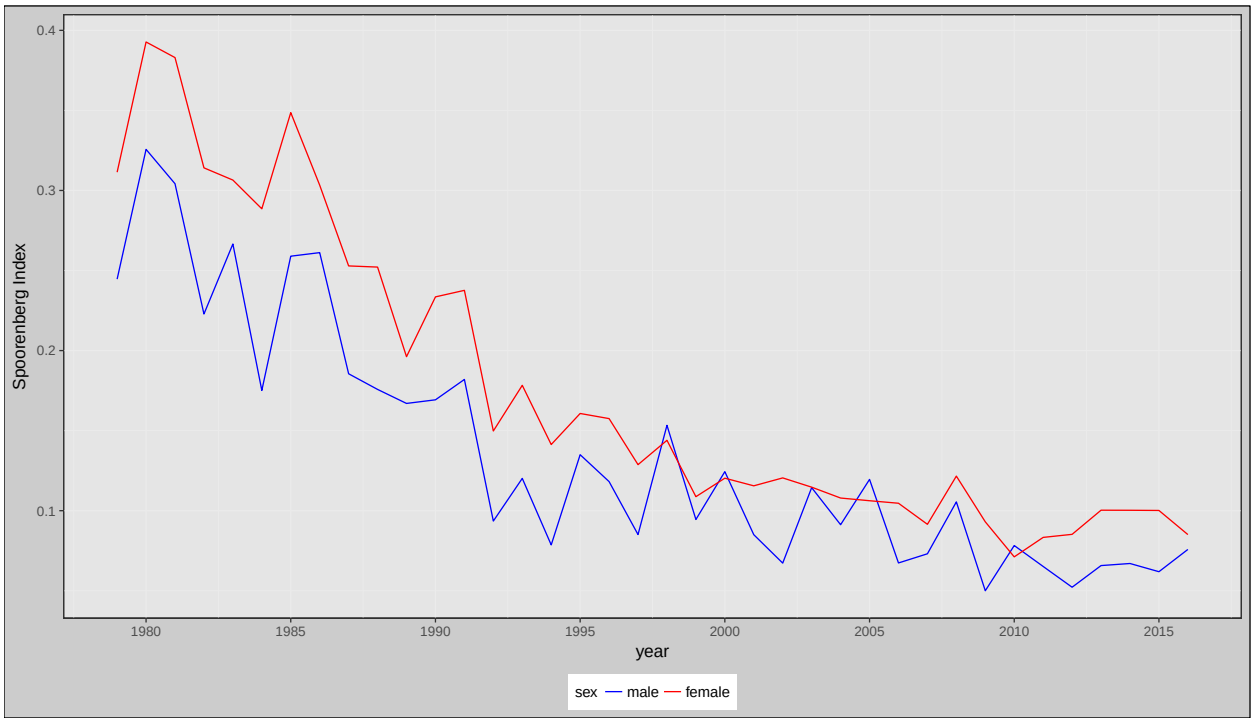

Figure 4.25: Spoorenberg Index for the registered deaths, by year and sex, 1979-2016. Source: SIM

underregistration of deaths. The almost complete coverage of registered deaths shown in the maps for the states of the South, Southeast (except MG), MS and DF is partially an artifact of the method used. Previous evaluation studies indicates that these states has almost complete coverage, which were assumed to be true in the publication of RIPSA, (2013). Results of the capture-recapture study (4.29) shows that these are indeed among the states with highest quality of registered deaths, but there is a more subtle difference between them and the other states.

Death Distribution Methods for estimating completeness of registered deaths

DDM are one of the most used demographic techniques to estimate adult mortality, through the estimation of the under-registration of deaths relative to census undercount. These methods were built on the demographic knowledge about the relationships between the distribution of deaths, the population age structure and the rates of change in populations.

This set of methods can be divided into two major approaches: i) the Growth Balance methods and ii) the SEG methods.

Figure 4.26: Spoorenberg Index for the registered deaths, by year, sex and state, 1979-2016. Source: SIM

General Growth Balance (GGB) methods

The Growth Balance method was first proposed by Brass, (1975) based on the relationships observed in closed and stable populations. When this is the case, the population growth rate is constant for all age groups ($r = r_x = r_{x+}$) and the completeness of registered deaths, δ, which is assumed to be constant by age, is given by the slope of the following relationship:

$$\frac{K_x(t)}{K_{x+}(t)} = r + \frac{1}{\delta}\frac{D^{obs}_{x+}(t)}{K_{x+}(t)} \tag{4.42}$$

where $K_x(t)$ is the observed population in the age group x at time t, $K_x+(t)$ is the population at time t aged x years old and over and $D^{obs}_{x+}(t)$ is the observed death counts of people aged x and over. Notice that this approach assumes no error in the census counts: $K_x(t) = K^{obs}_x(t)$.

Hill, (1987) and Preston and Hill, (1980) propose a variation of this method that makes no assumption of stability and is based on the intercensal comparison of two successive age groups. The main difference between the two approaches is that Hill, (1987) uses the the growth rates of the same age groups (r^{obs}_x) in order to avoid instability due to age misreporting errors existing in the censuses, while the Preston and Hill, (1980) approach, compares the same cohort in two different periods of time (r^{obs}_c).

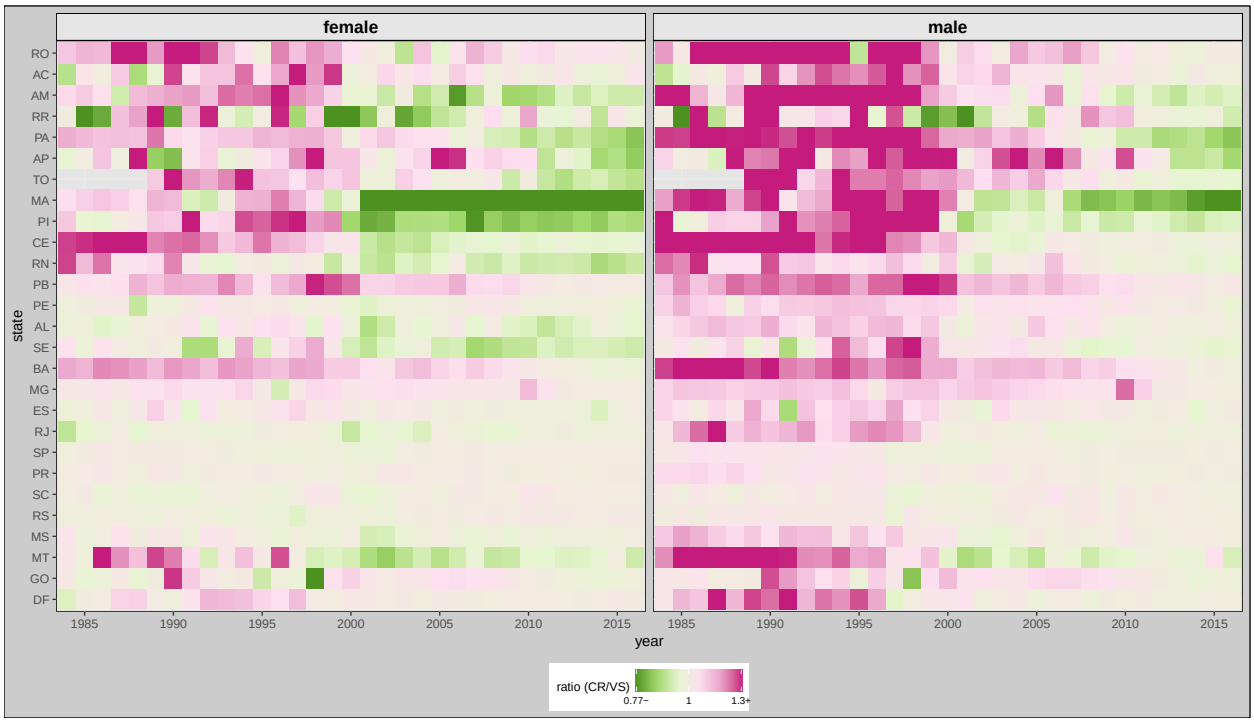

Figure 4.27: Ratio between registered deaths in the CR and the VS, by year, sex and state, 1984-2016. Source: CR and SIM

A by-product of these techniques, in addition to a measure of completeness of registered deaths, is an estimate of the relative coverage of two consecutive censuses $\left(\frac{\alpha(t_2)}{\alpha(t_1)}\right)$, which is assumed to be constant by age.

$$\frac{K_x^{obs}(t)}{K_{x+}^{obs}(t)} - r_x^{obs} = \kappa + \frac{1}{\delta}\frac{D_{x+}^{obs}(t)}{K_{x+}^{obs}(t)} \tag{4.43}$$

where κ is the "error" in the calculated growth rate, represented by a change in census coverage between two consecutive censuses. More specifically, this equation can be rewritten as:

$$\frac{K_x^{obs}(t)}{K_{x+}^{obs}(t)} - r_x^{obs} = \frac{1}{t} log\frac{\alpha(t_2)}{\alpha(t_1)} + \frac{1}{\delta\sqrt{\alpha(t_2)\alpha(t_1)}}\frac{D_{x+}^{obs}(t)}{K_{x+}^{obs}(t)} \tag{4.44}$$

SEG methods

The second group of DDM are represented by the SEG approaches. They make use of the observation that the number of people of a given age alive at a point in time must be equal to the number of people of that cohort who die from that point in time onward (Timæus,

Figure 4.28: Ratio between registered deaths in the CR and the VS, by year, sex and age, 1984-2016. Source: CR and SIM

Dorrington, and Hill, 2013). In a stationary population, as in a life table, this means that $l_x = \sum_{a=x}^{w} d_a$.

Preston, Coale, et al., (1980) proposed a method to estimate the completeness of registered based on this relationship for stable populations, with a constant growth rate, r:

$$K_x = \frac{1}{\delta} \int_{a=x}^{w} D_a^{obs} e^{r(x-a)} da \tag{4.45}$$

Bennett and Horiuchi, (1981) extended this method for non-stability contexts by accommodating the idea of differential growth rates by age, r_x:

$$K_x = \frac{1}{\delta} \int_{a=x}^{w} D_a^{obs} e^{\int_x^a r(u)du} da \tag{4.46}$$

This method can also be adjusted by differential census coverage. Bennett and Horiuchi, (1981) propose an iterative adjustment until until the slope of the curve of the completeness by age is zero. This issue can also be addressed by combining the differential census coverage calculated by the GGB method, adjust the census data and then apply the SEG method.

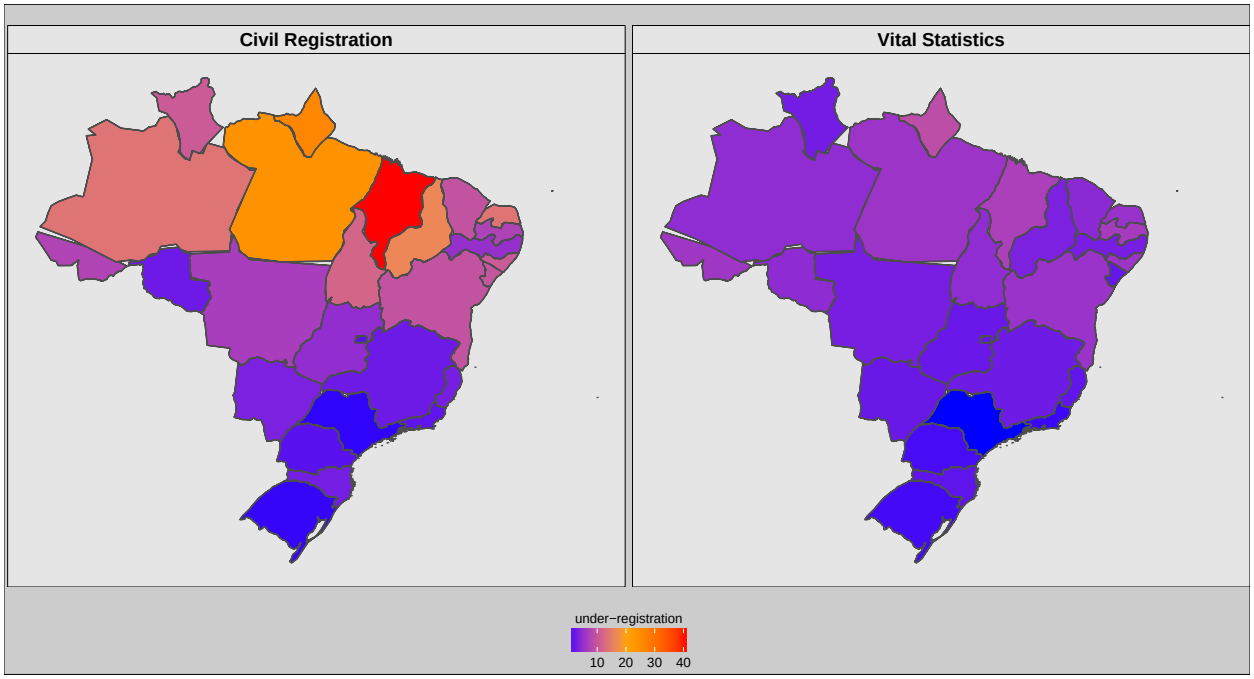

Figure 4.29: Map of the under-registration of deaths data source (CR and VS) and state, 2015 (in %). Source: Trindade, L. F. L. Costa, and A. T. R. Oliveira, (2018)

One of the most discussed issues in the practical applications of the methods is the choice of the age groups to be used. The section below will discuss some findings about this choice.

Sensitivity Analysis of DDM

DDM assume that the population is closed, the completeness of censuses and death registrations are constant by age and there is no age misreporting. In practice, these methods are more sensitive to some assumptions than to the others (Hill, You, and Choi, 2009).

Although widely recognized in the literature, only a handful of papers have tried to measure the impacts of the violation of the assumption in the results Dorrington and Timæus, (2008), Hill, You, and Choi, (2009), and Murray et al., (2010).

Hill, You, and Choi, (2009) find that the methods perform well in the presence of typical patterns of age misreporting. The GGB approach is more sensitive to this kind of error. Another important finding is that all methods are extremely sensitive to migration. The GGB and the combined GGB-SEG underestimate coverage in the presence of immigration, whereas the SEG overestimates coverage. They do the opposite in populations affected by emigration (Hill, You, and Choi, 2009).

The studies that have conducted sensitivity analysis of the DDM recommend different age trims and combinations of the methods. Dorrington and Timæus, (2008) recommend the use of the SEG with the delta adjustment for difference in census coverage. (Hill, You,

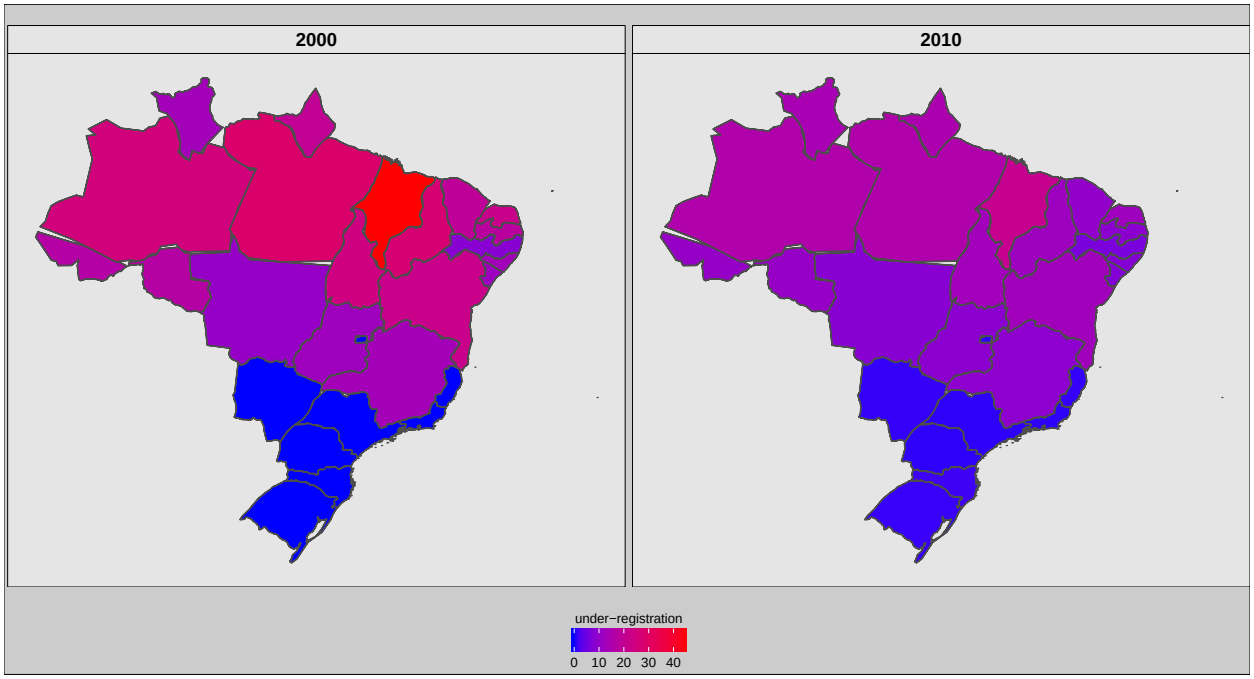

Figure 4.30: Map of the under-registration of deaths by state, 2000 and 2010 (in %). Source: (RIPSA, 2013)

and Choi, 2009) indicates that the best strategy based on the error patterns they analyzed is to combine the GGB and SEG methods, by first adjusting for the estimated change in census coverage from GGB, and then applying SEG. They also recommend the use of the age range 5+ to 65+ for fitting purposes, but they also indicate that applying either the GGB or SEG to the age range 30+ to 65+ reduces the effect of migration and the biases in the results. Murray et al., (2010) propose a different age trim for each method (SEG 55–80, GGB 40–70, and GGBSEG 50–70) and recommend the use of the median result across the three methods. The authors also the indicate that uncertainty around relative completeness of registration is likely to be at least $\pm 20\%$.

These papers calculate the sensitivity of mortality estimated using these methods to deviations from their main assumptions by using simulations. These studies have produced important results about the performance of the methods under violations of different assumptions, but they have not provided a framework to be used more broadly to identify the impact of different sources of biases. There are still no suitable ways to evaluate the sensitivity of these methods to assumptions other than those selected in these studies. Future work should develop a more comprehensive framework for evaluating the sensitivity of the methods to varies assumptions, which could be then used for adjusting the results.

Results of completeness of registered deaths for Brazil and states using DDM

Figure 4.31 shows the map of the completeness of registered deaths by state, sex and year, for both the CR and VS systems. These are estimated by the combined GGB-SEG methods. The under-registration is higher in the states of the Northeast region. The results have some inconsistencies, as well. For example, there are several states with more than 100% of completeness.

This could indicate that census undercount is higher than underregistration of deaths. In the more developed states in the South and Southeast regions, where CRVS are nearly complete, and there are problems with the censuses (as shown in Section 4.2), adjustment factor greater than one are expected. However, factors greater than one for states in the North region for the period 1990-2000, for example, are more likely to be indicating violations of the main assumptions of the methods.

DDM provide relevant results for the evaluation of completeness of registered deaths, particularly for those years with no alternative information for that purpose. However, they should be interpreted and used carefully. For subnational levels in Brazil, all assumptions required by the methods are likely to be violated. Section 4.5 shows that population of Brazilian states are far from being closed. Section 4.2 shows that Brazilian census have coverage problems, which vary by year, state, age. Finally, the results presented below (Section 4.4) indicate that completeness of registered deaths is probably not constant by age.

Mortality estimation from deaths in the household

The question about the existence of a recent death in the household has been included in at least 70 censuses in all regions of the the world, such as Latin America (e.g. Brazil, Bolivia and Paraguay), Sub-Saharan Africa (e.g. Nigeria, South Africa, Rwanda) and Southeast Asia (e.g. Vietnam, Indonesia).

This information has some advantages over others used to estimate mortality, which are mostly related to the fact that it often collected in censuses, which normally does not have implicit sampling errors. Furthermore, numerator and denominator come from the same source and mortality inequality can be estimated according to different characteristics of the household also collected in the censuses, for instance income, rural/urban classification and access to water supply and sanitation. This is also the only information collected in censuses and surveys that provides mortality estimates for all ages (Queiroz and Sawyer, 2012; UNSD, 2004).

Despite the availability and the potentialities, these data has not been much used, although some successful applications have been performed in China (Banister and Hill, 2004), Brazil (Queiroz and Sawyer, 2012) and Sub-Saharan African countries: Cameroon (Bangha, 2010), Lesotho and Botswana (K. Thomas and Hill, 2007).

The lack of a more extensive use of the household deaths is partially due to the limitations of the data (Hill, Choi, and Timæus, 2005). Potential sources of bias in this information are

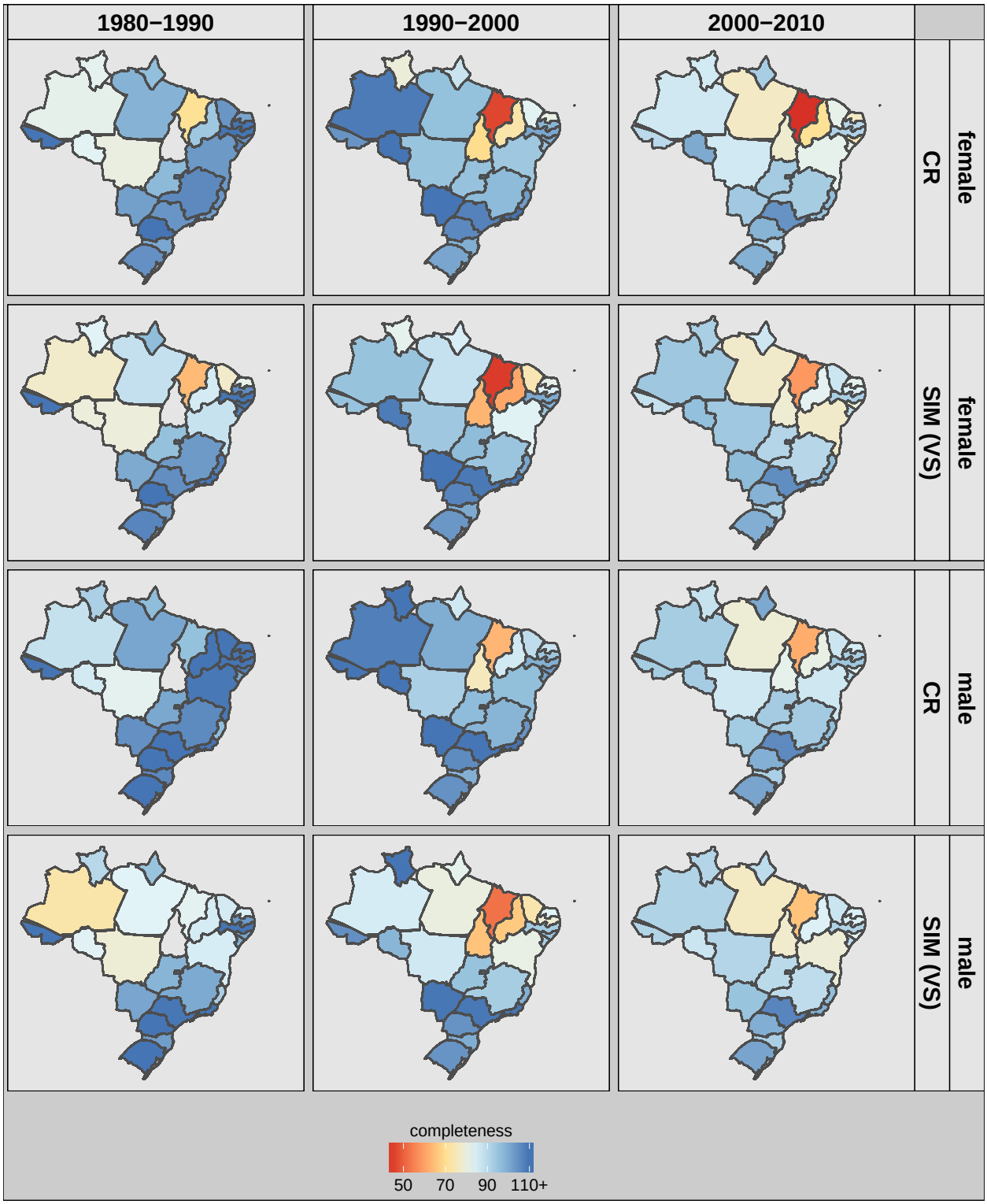

Figure 4.31: Map of the completeness of deaths by state, sex, year and data source (in %)

related to the confusion about how to report deaths in the case of the disintegration of the household due to a death or when people are seen as belonging to more than one household (Dorrington, Timaeus, and Gregson, 2006). More importantly, under-reporting inevitably results from households with no survivors to report the death. This is particularly clear in one-person households, where the death, by definition, cannot be reported. This bias is likely to be differential by age, being higher at older ages, in which one-person households are more prevalent (Dorrington, Timaeus, and Gregson, 2006; UNSD, 2008). As a result of this main limitation, Queiroz and Sawyer, (2012) claim that household deaths reported in the 2010 Census in Brazil represent a good picture of the age structure of mortality, except for the older age groups.

This section discusses the results of mortality estimates based on the question in the household in the 2010 Census in Brazil. Census interviewers asked about the occurrence of deaths of residents in the previous 12 months before the census, by asking the question: *"From August/2009 to July/2010, did any person who used lo live in this household pass away?"*. Sex and age of the deceased was also asked by the census interviewer. The question was asked to 51 million non-institutional households in Brazil and captured more than 1 million deaths.

First, this section explores the results of the reported deaths, by comparing them with the deaths informed to the VS system(SIM). Second, an adjustment is proposed to correct for the main source of bias of this information, the impossibility to capture deaths from single-person households. Finally, mortality indicators are calculated based on the adjustment proposed.

The adjustment method proposed in this section builds on the existing literature about corrections for selection biases in the sibling survival method, such as the resulting from the fact that some sibships are not observed (Gakidou and King, 2006; Masquelier, 2012), in addition to the recent works on the network reporting framework for estimating adult mortality (Feehan, Mahy, and Salganik, 2017).

The underlying idea of proposed adjustment is to reconcile the numerator (reported deaths) with the denominator (population exposed to the risk of dying).

The first adjustment, perhaps obvious, but often neglected, is to exclude from the denominator the population living in households where the question about household deaths was not asked, for instance people living in institutions.

The second adjustment intends to correct for the bias of under-reporting as a result of households with no survivors to report the death. The lack of this information on death in single-households clearly bias mortality estimation downward, since part of the population at risk is still being reported, although the deaths are not.

One way of adjusting for this bias would be through the estimation of the deaths that occurred in households in which all members died. Assuming that the joint probability of dying of all household members within a year is significantly low for households with two or more members, the proposed adjustment focuses on the estimation of the deaths that occurred in one-person households only.

Alternatively, instead of trying to estimate the deaths in households with no survivors, this section proposes an adjustment in the denominator, which is the exposed-to-risk population. A more consistent denominator for the mortality rate would be the one which removes the population exposed to the risk in one person-households estimated a year before the census. Note that the new population estimate excludes not all on-person households in the census, but only those with no deaths reported. The main assumption in this procedure is that mortality does not vary significantly by household size.

The naïve mortality estimation for the year before the census by age and sex group (M_x), which has been usually calculated, is given by the total number of reported deaths (D_x) divided by the population in the same group counted in the census (K_x):

$$M_x = \frac{D_x}{K_x} \tag{4.47}$$

The adjusted mortality estimation is given by:

$$M'_x = \frac{D_x}{K'_x} = \frac{D_x}{K_x + 0.5D_x - K_x^{oph}} \tag{4.48}$$

where K'_x equals the population enumerated in the census (K_x) plus half of the reported deaths ($0.5D_x$), minus the population at time in the census living in one-person households with no reported deaths in the 12 months period prior to the census (K_x^{oph}).

To test the accuracy of this adjustment, the next section presents the results for the 2010 Census in Brazil, which has relatively reliable mortality estimations using different data sources, which are used for comparison. In addition to compare to the official mortality estimations for the country as a whole, the adjusted mortality estimations can be also compared to the mortality calculated using vital registration in states where this information is known as of good quality.

To test the accuracy of this method, unadjusted and adjusted mortality rates are compared with official mortality estimates. Figure 4.32 shows this comparison with official estimates published by IBGE, which are calculated using vital registration data and adjusted for undercount of deaths using indirect demographic techniques (IBGE, 2013c).

The comparison shows that mortality rates estimated using the adjustment method proposed in this section are much closer from the official estimates than the unadjusted rates. For females, mortality rates practically coincide with official estimates for all age groups above 25 years old, whereas the unadjusted estimates diverge from the "true" value as age increases. It should be noticed that, since the official estimates correct the registered deaths, this information is also subject to errors. It is worth mentioning that household deaths information appears to overestimate mortality among children and young adults from 5 to 25 years old.

Figure 4.33 shows the estimated adult mortality (${}_{45}q_{15}$) calculated based on the question on household deaths for the 2010 Census, adjusted by the procedure described above. The

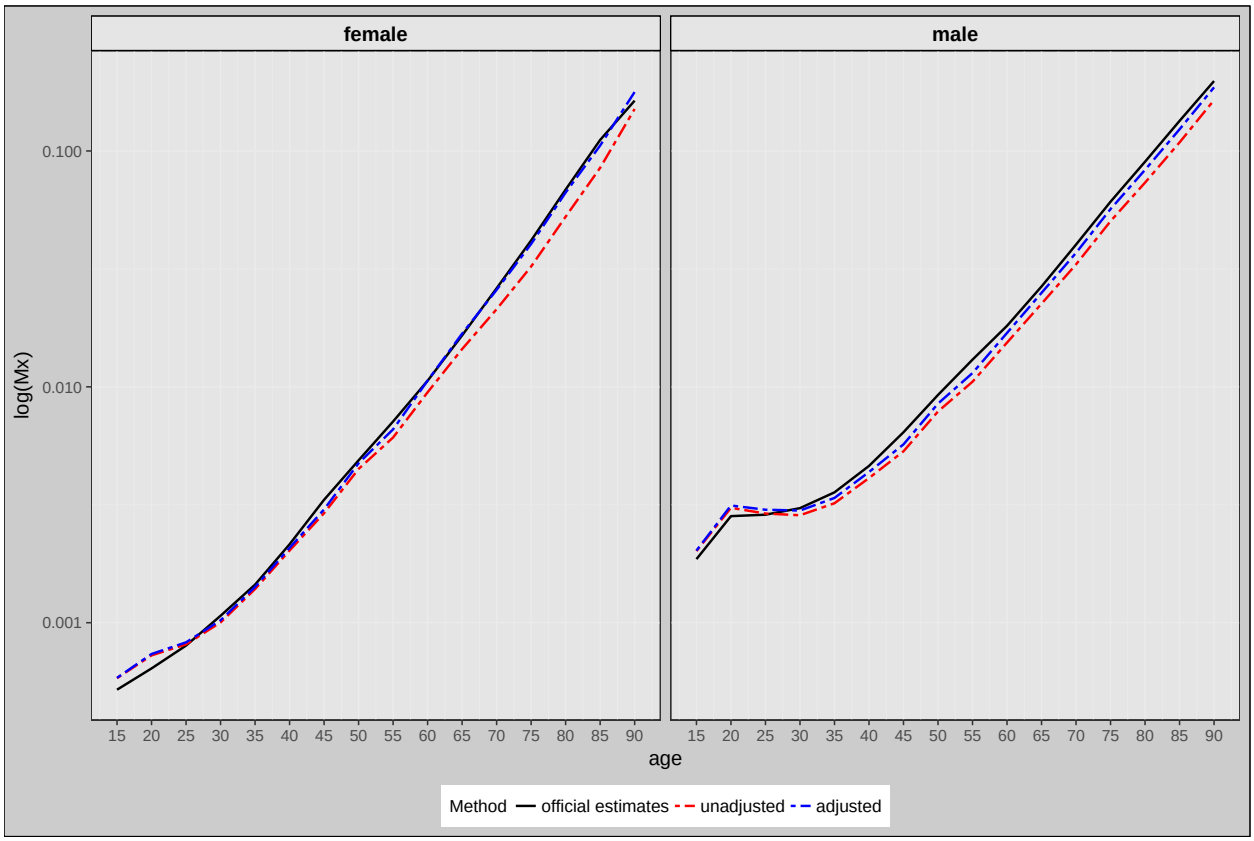

Figure 4.32: Mortality rates by age sex and estimation method compared with official estimates. Source: 2010 Census

figure shows low mortality for developed states, such as DF, SP, SC, but also indicate that a few states in the North and Northeast regions may also have low adult mortality, such as PI, CE and RN. The state of AL has high mortality for both male and female populations. RJ seems to have high adult mortality, particularly among females.

This section has proposed a method to adjust for bias on recent deaths in the household reported in censuses. Application for Brazil shows that the adjustment greatly improves mortality estimations, particularly at old ages, resolving one of the main limitations of this information. The application of the proposed method to estimate mortality in Brazil show figures extremely consistent with other information available, showing no need to further adjustments. For other contexts, however, adjusted mortality rates might need additional corrections, which could be accomplished by the existing methods for assessing and adjusting completeness of deaths counts from vital statistics. In this case, since these methods normally assume constant adjustment factor over age, the method proposed in this paper will still be useful as it corrects for the underreported biases by age.

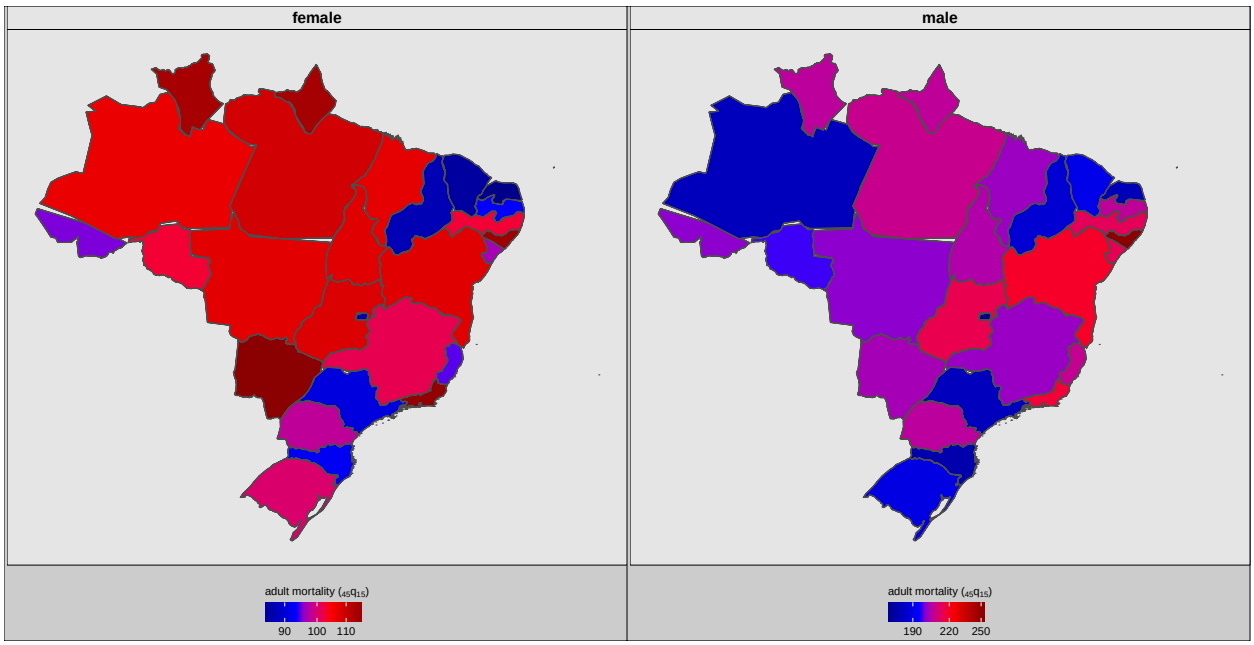

Figure 4.33: Adult mortality ($_{45}q_{15}$) by sex and state, 2010 (in ‰). Source: 2010 Census

Estimation of adult mortality from maternal orphanhood

Orphanhood methods estimate mortality based on questions about the vital status of the father or mother of the respondents. This section concentrates on the a analysis of maternal orphanhood, since this is the only information collected in Brazilian censuses. This information is present in censuses with questions such as *"Is your mother alive?"*. This question is simple, easy to answer and is may be less subject to errors than other questions used with a similar purposes, as those about siblings. This makes it easier to include such a question on a questionnaire of censuses rather than in surveys.

The 1980 and 1991 censuses in Brazil had this question, but the 2000 Census did not includ it. This question is present again in 2010. The main purpose of this question was to reconstruct families cohabiting the same holsehold. However, it offers a great opportunity for estimating mortality as well. These data have never been used to estimate mortality. This will be particularly useful to shed some light on the regional inequalities in mortality and compare with other mortality estimates.

The principle of this method is similar to that of the Brass method for estimating child mortality. In this case, the proportion of mother dead is a measure of mortality between the age at which the mother gave birth and the age of the respondent at the date of the survey. The latter tends to be a very precise measure, whereas the former can be approximated by the average of the age difference between mother and child in the population [11]. Results

[11] Note that this is different from the mean age at childbearing

for different age groups give a measure of mortality in different periods in time, allowing for time trend estimates of adult mortality. The conversion of distinct measures of survivorship into a single mortality indicator and the further time allocation depends on the choice of a life table.

The results of the application of this method is an estimate of female adult mortality, such as ${}_{45}q_{15}$. This information can be combined with infant mortality estimates, such as that discussed in section 4.4 using relational models.

The limitations of this method are closely related to the limitations more extensively discussed in the literature for the sibling survival method. Only adults who have living children are reported, and adults with more children are over-represented. However, the results of the method will be unbiased if mortality does not vary by the number of living children.

One of the advantages of this method is that it requires no assumption on closed population, which may be relevant to the context of estimating subnational mortality in Brazil. However, the interpretation of these results should be taken with care, particularly for states with high migration, as the parents may not live in the same place as their children.

The average age difference between the mother and the child, called here mean age of maternity, is an indicator of the average age at which mothers begin their exposure to the risk of dying. This is given by the average of the age of woman giving birth [12]:

$$M = \frac{\sum_{x=15}^{45} {}_5B_x(x+2)}{\sum_{x=15}^{45} {}_5B_x} \tag{4.49}$$

where ${}_5B_x$ is the number of births in the 12 months prior to the census reported by women of age x and $(x+2)$ represents the midpoint of the age group, with a 6 month shift to take into account that mothers were, on average, six months younger when they had their children than the reported age in the census.

The value of M is used in combination with the proportion of surviving mothers by age group to produce estimates of the conditional probabilities of survival between age 25 and $25+n$, where n is the upper limit of the respondent age group, ${}_np_{25}$. The regression coefficients used are those proposed by Timæus (1992) and presented in Timæus, (2013b).

Once the probabilities ${}_np_{25}$ are estimated, the adult mortality measure of interest, in this case ${}_{45}q_{15}$, which represents the probability of dying between ages 15 and 60, are calculated by finding a correspondent model life table. In this study, the Coale and Demeny West life table is used (Coale and Demeny, 1966).

Finally, the estimates are located in time, since the probabilities for different age groups refer to estimates of distinct points in time.

[12] This measure is not weighted by the age structure of the population, as the mean age of childbearing.

Results for both the 1991 and the 2010 censuses show a rapid decline in mortality in the period close to the censuses. This is probably indicating an underestimation of survival of women reported by children aged 5 to 15. This information is problematic, as this group is probably not reporting the maternity survival themselves, which, among other causes, could bias the results. Thus, the trend of mortality change was adjusted excluding this first three groups. The adult mortality calculated fo Brazil by using this procedure is consistent with official mortality estimates for the same years.

4.34 shows the estimated ${}_{45}q_{15}$ by combining information from the 1991 and 2010 censuses by state and year. The map show a reduction in mortality over time. The regional differentials are consistent with those indicated by the previous section. The results show relatively low mortality in states in the South and Southeast regions, as well as the three states in the Northeast: RN, CE and PI.

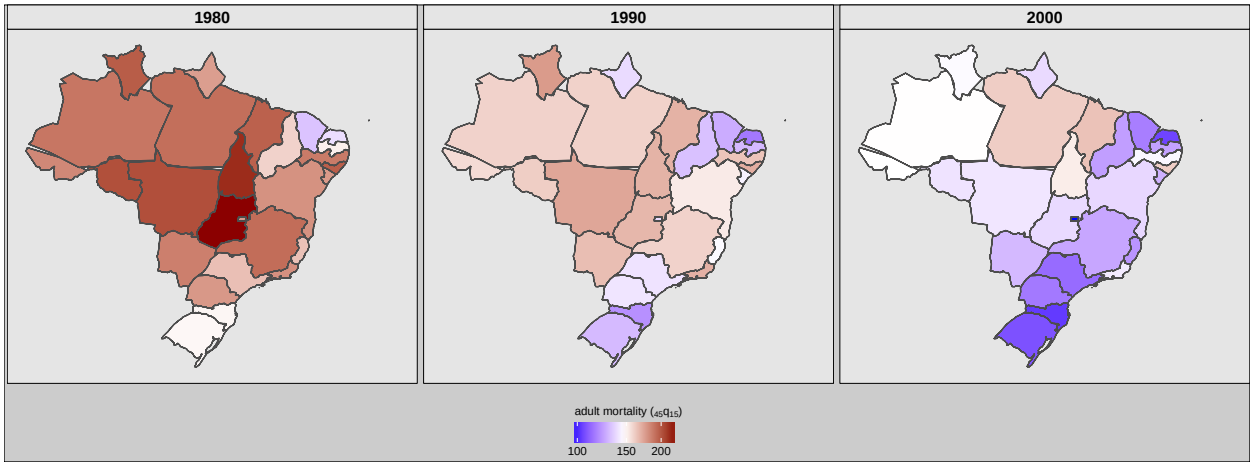

Figure 4.34: Map of adult mortality (${}_{45}q_{15}$) by year and state (in ‰). Source: 1991 and 2010 Demographic censuses

Conclusion

This section presented a comprehensive evaluation of available methods to estimate mortality and completeness of registered deaths in Brazil and states. Several results were produced and these will be used in combination in the next chapter (5) to produce mortality estimates with uncertainty measures.

4.5 Migration Estimates for Brazil and States from 1980 to 2010

Migration is probably the most difficult piece to measure in the balancing equation, due to the lack of administrative data.

Indirect demographic techniques have been used to estimate international migration, for instance by using intercensal residual methods or by including surveys and census questions about relatives or household members living abroad (Hill and Dorrington, 2013).

Information about the migration status of the population is also collected regularly in censuses, which identify the non-native population. This information, combined with the country of origin, allows an estimation of the stock of immigrants and emigrants. These estimations are done for Latin American countries, for instance, in the context of the project entitled "Investigation of International Migration in Latin America" (IMILA). This project collects information on immigrants in several Latin American censuses, and also captures similar information from the United States and Canada censuses (Pizarro and Villa, 2005). Questions about the place of residence one or five years before the census date are also collected in some countries and provide an estimate of recent immigration flow.

Internal migration flows are better captured by censuses than international migration, because it is possible to estimate both in-migration and out-migration using the same data source [13].

Brazilian censuses have a rich history of questions about migration, such as place of birth, duration of residence in the location where the respondent was enumerated, place of previous residence and place of residence five years preceding the census.

The latter question has been asked in Brazilian censuses since 1991 and is relevant for estimating quinquennial migration flows, which is further used for population projections. This question was also included in the 1996 and 2007 population counts. The advantage of the 1996 Count over the censuses is that all the question were asked for the entire population, whereas in the the census this question has been in the long form and the results are subject to sampling errors. However, as discussed in Section 4.2, the 1996 Count had serious problems of coverage, and it is likely that they were differential by state and population groups. The 2007 Count, in addition to coverage problems, was only a partial count of the population, which prevents analysts from calculating migration flows, due to the lack of information on in-migration for certain states.

Note on migration rates

Demographers have calculated demographic rates for spatial and temporal comparisons. Even though absolute numbers may be useful for some purposes, rates that take into account the population exposed to the risk are often more meaningful for these comparisons. Since

[13] The terms "immigration" and "emigration" are used here for international migration, whereas the terms "in-migration" and "out-migration" are used for internal interstate migration (Haupt, Kane, and Haub, 2011)

demographic events tend to differ significantly, for instance, by age and sex, rates are also calculated for these different population groups.

As opposed to mortality rate, which has a clear definition (number of deaths over the number of people at risk of dying), migration rates are more complicated and can be ambiguous. First, there are several definitions of migration, for instance depending on the period under analysis. Second, population used in the denominator are not precisely the population "at risk" of migrating, particularly for immigration and in-migration rates.

Migration rates will always involve a number of migrants in the numerator and a population in the denominator. The definition of the rates will be given on a case-by-case basis according to the available information for both pieces.

International Migration in Brazil from 1980 to 2010

Brazil had been historically a receiving country, with the number of immigrants exceeding the number of emigrants. There is evidence that this has changed in the 1980s, when Brazil shifted from positive to negative net migration (Carvalho, 1996; A. T. R. Oliveira, 1996). Emigration was initially to the US and then to certain European countries, such as Portugal, the UK, Spain, Italy, and Japan. Despite the predominance of emigration flows, immigration from neighboring countries such as Bolivia and Paraguay was also observed starting in the 1980s (A. T. R. Oliveira, 2015).

The net international migration flows were still negative in the 1990s, although at lower levels than in the preceding decade (Carvalho and Campos, 2006). This tendency continued in the next decade, and net migration seems to have been close to zero in the 2000s (Campos, 2011).

A large portion of Brazilians living abroad is undocumented. The proportion of Brazilian undocumented immigrants in the US, for instance, is estimated in 60% (Soares and Fazito, 2008), which complicates the estimation of the number of Brazilians living abroad.

The estimate of the number of Brazilians living abroad varies depending on the data source. The Ministry of Foreign Affairs estimates that this figure is slightly more than 3 million (MRE, 2016), whereas the International Organization for Migration (IOM) claims that there are between 1 and 3 million Brazilians living abroad. In the 2010 Census, $491,645$ people were reported living in another country.

The estimated stock of immigrants living in Brazil is more precise and is based on the direct question in the censuses about the place of birth of the resident population, although there might be some imprecision, mainly due to an underestimation of the undocumented immigrants whom are harder to count and/or are more likely to hide their migration status. The 1991, 2000 and 2010 censuses reported respectively $767,781$, $683,830$ and $592,569$ immigrants living in Brazil. The immigrants represented only 0.3% of the Brazilian population in 2010 and this population was largely concentrated at old ages, as a result of historical migration flows, e.g. from Portugal and Japan.

Indirect estimates of net international migration in Brazil

The main estimates of net international migration in Brazil come from indirect demographic techniques that use the intercensal residual method. The idea of this method is to compare two consecutive censuses through the CSR, discount the effects of mortality and then attribute the difference to international migration. Scenarios can be created given different mortality estimates for the intercensal period, as well as differential census coverage.

Carvalho, (1996) estimates the net migration in the the 1980s by using the CSR and two scenarios of mortality. In one scenario, the net migration in the decade was slightly more than 1 million people (302 thousands females and 741 thousands males). In the second scenario, net migration increased to 2.54 million people (1.19 million females and 1.35 million males). This publication was innovative and the results caused some surprise because that was the first time negative net migration was reported in Brazil.

Figure 4.35 shows the international net migration rates by age for males and females estimated under the two mortality hypothesis for the period 1980-1990. Carvalho, (1996) defines net migration rates $nm_{x,s}^{1980;1990}$ as follows:

$$nm_{x,s}^{1980;1990} = \frac{\mathrm{NM}_{x,s}^{1980;1990}}{K_{x,s}^{1990}} \times 1000 \tag{4.50}$$

where $\mathrm{NM}_{x,s}^{1980;1990}$ is the international net migration (difference between the number of immigrants and the number of emigrants) for age group x and sex s between the years 1980 and 1990; K_x^{1990} is the population in the age group x and sex s estimated for 1990.

International net migration rates ($nm_{x,s}^{1980;1990}$) are negative for most age groups and are lower for males than for females.

These estimates have some limitations, most of them recognized by the author. The broad range of the estimated net migration (1 to 2.5 million) shows that figures resulting from the application of this technique are extremely sensitive to the choice of the intercensal mortality. Figure 4.35 shows that most of the difference in the rates between the two hypothesis appear at old ages. The figure also shows that estimates by age are inconsistent with the known age pattern of migration, which indicates that large part of what is attributed to international migration, may be actually an artifact of census undercount, age misstatement, and/or poorly estimated intercensal mortality.

A. T. R. Oliveira, (1996) used a similar method, but focus the analysis in the population aged 20-44. The authors estimate the decennial net migration flow of this age group in -1.3 million. Although this is an expressive figure in absolute number, this decennial migration represents only a small portion of the overall Brazilian population, which was about 150 million people in 1991.

Carvalho and Campos, (2006), following a similar methodology, estimated the net migration for the period between the 1991 and 2000 censuses in around 550,000 (294,000 males and 256,000 females) for those aged 10 years old or more in 2000. The main difference between the studies of Carvalho, (1996) and Carvalho and Campos, (2006) is that the former assumes

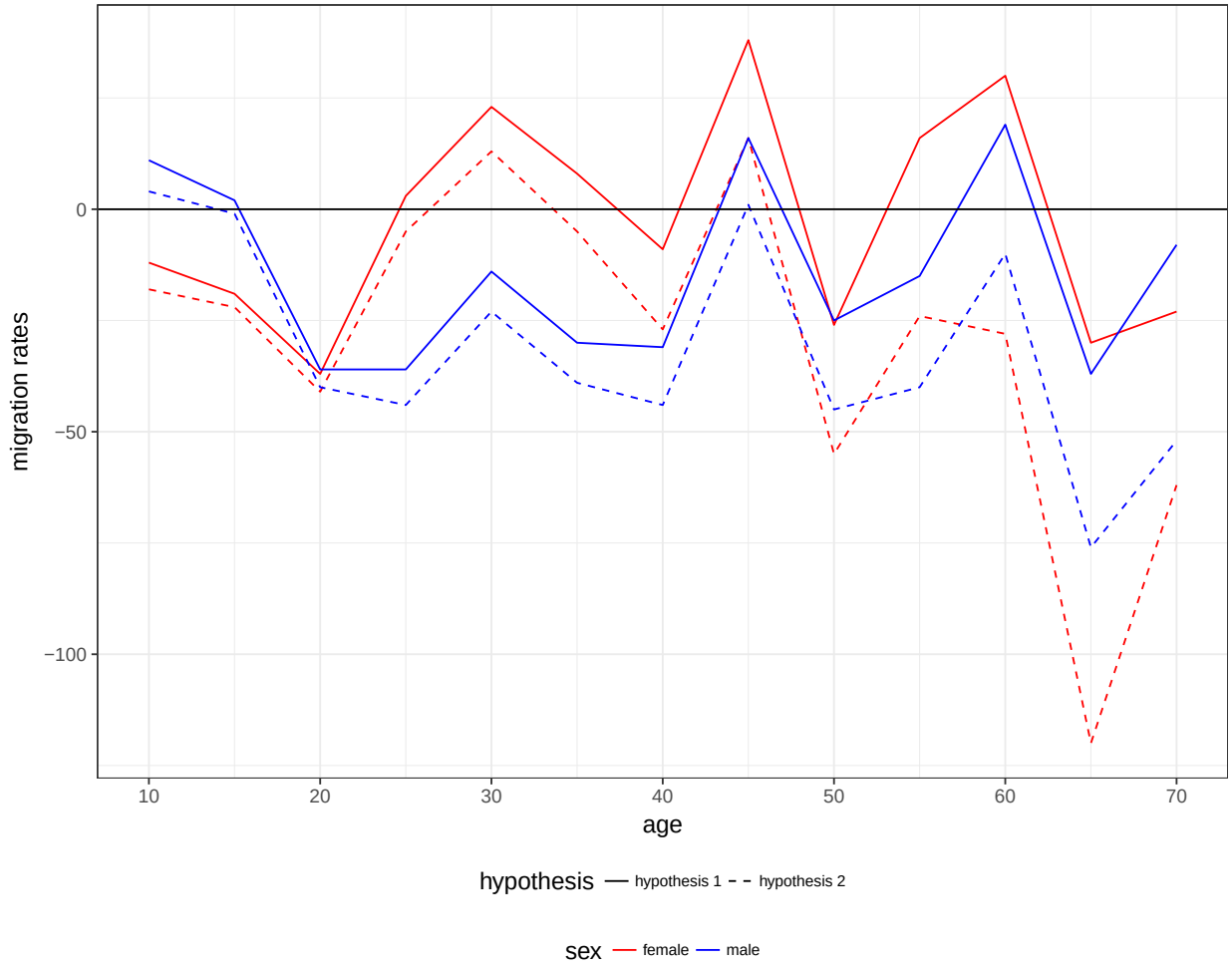

Figure 4.35: International net migration rates (‰) by age, sex and mortality hypothesis, 1980-1990, Brazil. Source: Carvalho, (1996)

equal census coverage for both censuses, while the latter adjusts for differential undercount in the censuses based on the PES estimates. The authors adjust the population of the 1991 Census in 9.08% and that of the 2000 Census in 5.84%.

The first limitation of this procedure is that it is known that census coverage is differential by age - see discussion in Section 4.2. The second is that the two omission rates used are not comparable. For the 1991 PES, the adjustment of 9.08% (or 8.32% undercount rate) refer to the overall omission rate, whereas the adjustment factor used for the 2000 PES (5.84%, or 5.52% undercount rate) refer to the net census coverage error, which subtracts duplications and erroneous enumerations from the overall omission rate. A figure comparable to the 8.32% undercount rate found in the 1991 Census for the 2000 Census is 7.87%, which indicates similar coverage in the two censuses, at least for the aggregate of the country. As

discussed in 4.2, the PES of the 1991 Census published no results about duplications. The use of similar adjustment factors would lead to an international migration closer to zero for the 1990s.

R. A. Garcia, (2013) estimates the negative net migration for the period 1995-2000 at around 250 thousands, divided almost equally among women and men. Carvalho and Campos, (2006) claim that the decline in the sex ratio of the net migration is attributed to the change in the nature of migration, in which there are more men in the first phases of the international migration, in this case in the 1980s, followed by an an equilibrium in the sex ratios, due to family reunification, among other reasons, which would have occurred after the 1990s.

Campos, (2011), again assuming that there was no difference in the coverage of the 2000 and 2010 censuses, and taking an official intercensal mortality estimate for the decade, suggests that the net migration in the 2000s was close to zero.

Despite the limitations of these techniques, there are evidences that international migration in Brazil have situated in very low levels compared to other countries in the world. The net migration flows seem to have been negative in the 1980 and moved towards zero in the next two decades.

Immigration

There are different ways to estimate immigration to Brazil based on census data, depending on the definition of the migration interval. Migrants can be defined according to their place of birth, place of previous residence and place of residence at a fixed past, for example five years before the census. The latter measure is commonly used as a reference for recent immigration. It includes both the non-native population and Brazilians who had emigrated in the past and returned within the five-year period prior to the census.

The number of immigrants in the five-year period prior to the census has increased from $66,218$ in 1991 to $143,644$ in 2000 and $268,298$ in 2010. In 2000 and 2010, more than 60% of those immigrants were Brazilians who had emigrated previously and have returned recently to Brazil. In 2010, almost 1/5 of those recent immigrants came from the US.

Immigration rates by age x and sex s for the period between years $y-5$ and y ($im_{x,s}^{y-5;y}$) are given by:

$$im_{x,s}^{y-5;y} = \frac{\mathrm{IM}_{x,s}^{y-5;y}}{K_{x,s}^{y}} \times 1000 \tag{4.51}$$

where $\mathrm{IM}_{x,s}^{y-5;y}$ is the number of immigrants for age group x and sex s between the years $y-5$ and y. This is the number of people who reported they were living in a different country five years prior to the census date. K_x^y is the population in the age group x and sex s enumerated in the census of the year y.

Figure 4.36 shows the five-year immigration rates ($im_{x,s}^{y-5;y}$) by age and sex for the years $y = 2000$ and $y = 2010$. The figure confirms the increase in the recent immigration between

the years 2000 and 2010. It also shows that the immigration is significantly higher among men older than 20 years old. The age pattern of the immigration reinforces the idea that an important part of this immigration is return migration. The figure also shows that immigration levels are extremely low. Even in the peak of immigration in 2010, around age 30, the proportion of the population that was living in a different country five years prior to the census ($im_{30,m}^{2005;2010}$) was less than 0.3% (or 3.0‰).

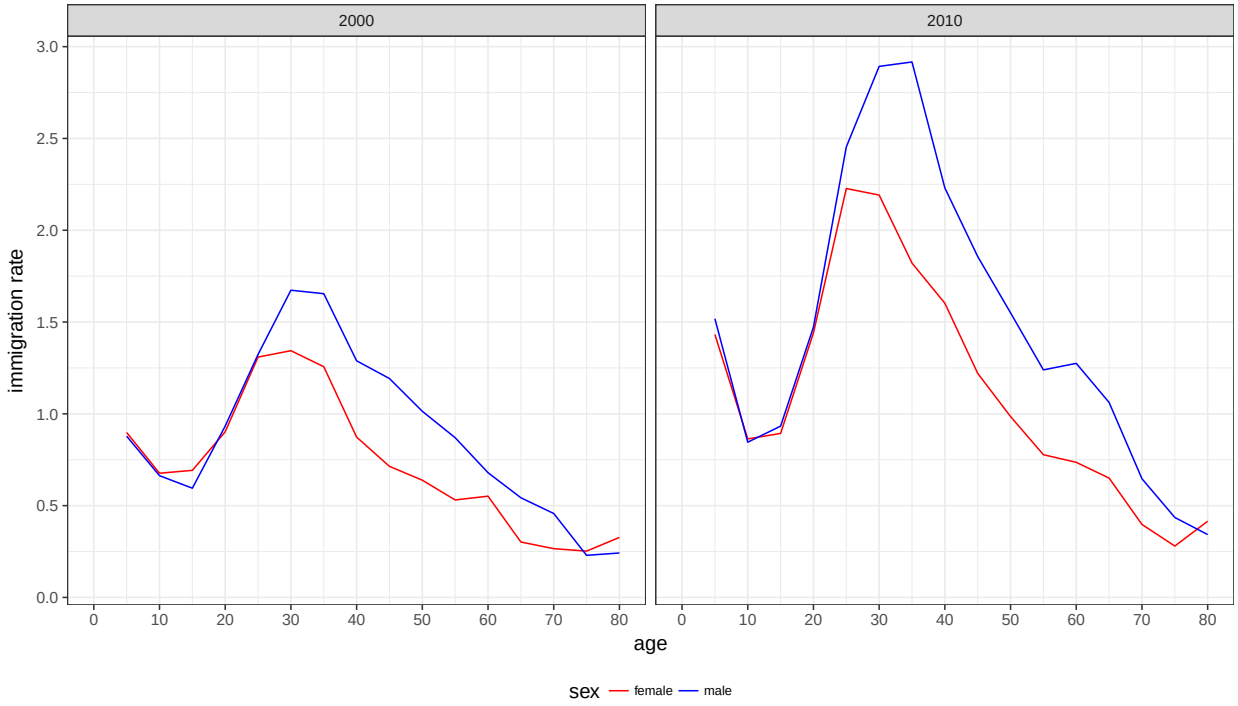

Figure 4.36: Immigration rate (‰) by age, sex and year, 2000 and 2010, Brazil. Source: IBGE, Censuses of 2000 and 2010

The overall proportion of people over age 5 who was living in a different country five years before the census date was 0.94‰ in 2000 and 1.52‰ in 2010.

Figure 4.37 shows the maps of this proportion for Brazilian states. This proportion ranges from close to zero for several states, particularly in the Northeast region, to more than 4‰ in MS and RR in 2000 and in PR and MS in 2010.

The map also shows an increase in the immigration for almost all states, more pronounced in a few states in the South and Southeast regions (ES, MG, SP, SC), in addition to GO, RO and DF.

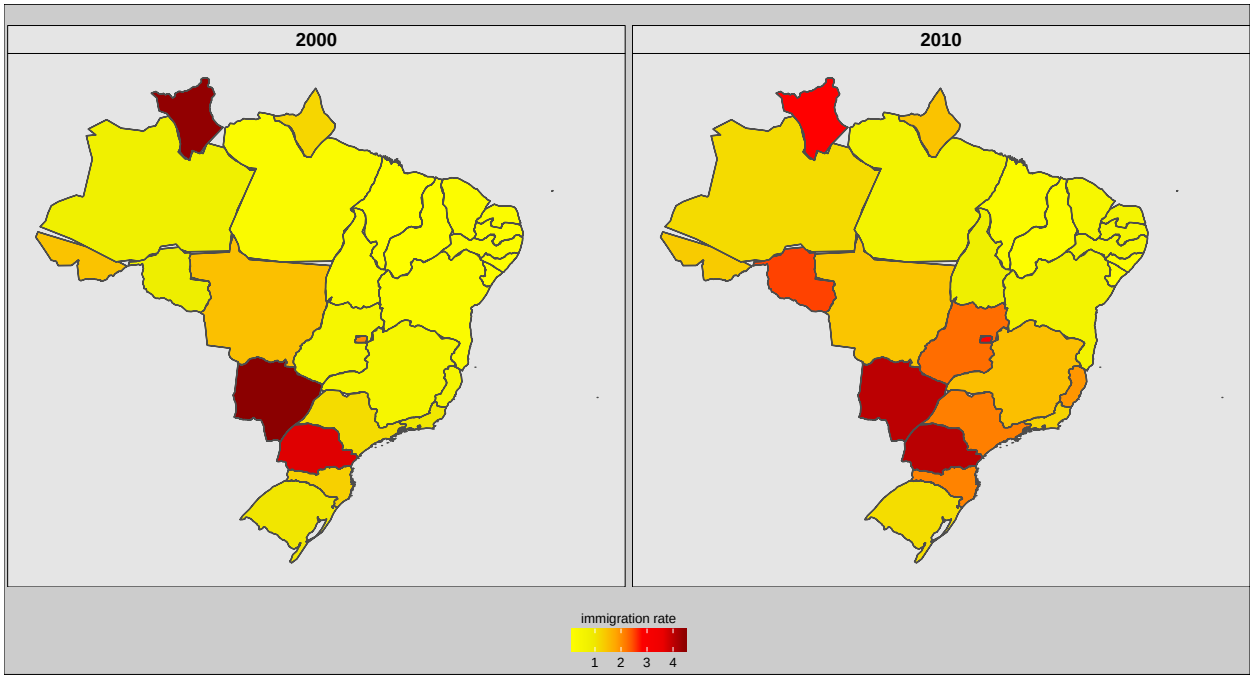

Figure 4.37: Map of the immigration rate (‰) by state and year, 2000 and 2010, Brazil. Source: IBGE, Censuses of 2000 and 2010

Emigration

As discussed above, there is much more uncertainty about emigration than immigration in Brazil. Official estimates of the number of Brazilians living abroad vary from a 1 to 3 million.

In 2010, the Brazilian census asked, for the first time, whether any household member has gone to live abroad, in which 491,645 people were reported. This figure is probably underestimated, primarily because households where all member migrated, including single-person households, had no respondent to report them. The number of emigrants reported in the Brazilian 2010 Census corresponds to only 34% and 24% respectively of the number of Brazilians reported in the America Community Survey (ACS) and in the Japanese census in the same year (Campos, 2014).

Despite the limitations, the results from these questions provide useful information about the country of destination of the emigrants, their origin in Brazil and their distribution by age and sex. The results show that around 70% of Brazilian emigrants are concentrated in only six countries: the US, Portugal, Spain, Japan, Italy and England (Campos, 2018; A. T. R. Oliveira, 2013).

The 2010 Census also asked when people left the country to live abroad. Around 60% of the emigrants reported living abroad were reported to have left in the period 2006-2010, around five years before the census date. This proportion is likely to be overestimated due to difficulties to remember old events or, more likely, a misunderstanding of the question about

when people left. Respondents may have reported, for example, the last time the migrants visited Brazil (Campos, 2018).

Emigration rates by age x and sex s for the period around five years before the census, people that left the country between 2006 and 2010, ($em_{x,s}^{2006;2010}$) are given by:

$$em_{x,s}^{2006;2010} = \frac{\mathrm{EM}_{x,s}^{2006;2010}}{K_{x,s}^{2010}} \times 1000 \tag{4.52}$$

where $\mathrm{EM}_{x,s}^{2006;2010}$ is the number of emigrants for age group x in 2010 and sex s that left the country between 2006 and 2010; $K_{x,s}^{2010}$ is the population in the age group x and sex s enumerated in the 2010 Census.

Figure 4.38 shows the emigration rates ($em_{x,s}^{2006;2010}$) for the period 2006-2010 by age and sex, calculated based on the question about the migration of a household member in the 2010 Census. Emigration is concentrated at working ages, around 25 years old, which is consistent with the well known migration age pattern. Emigration rates have a very similar age pattern for both sexes, but are consistently higher for women.

The limitations of this question in the census seem to introduce some bias in the results. The age pattern resulting from the question in the census reinforces the idea that that this question on migrant in the household captures more precisely individual migration than family migration. The age pattern of Brazilians living abroad captured by the 2010 Census is different from that collected by the censuses and surveys in the country of destination, particularly for those where family migration is predominant, for instance Japan. In this country, the number of children emigrants is largely underestimated, compared with the Japanese census (Campos, 2014).

Finally, the question in the 2010 Census provides useful information about the origin of the migrant population. This is perhaps the most useful result for population estimates, as it is the only data source for this type of information. Censuses and surveys in the countries of destination rarely ask questions about the detailed locations in the countries of origin.

Given this information, once emigration for Brazil is estimated, the emigrants may be reallocated in their respective states based on these results.

Figure 4.39 shows the map of the emigration rate by state, showing a significant spacial variation. All the states in the Northeast region have very low rates (below 1.2‰), and several states in the other regions have relatively high rates. GO is the state with the highest emigration rate: almost 6.0‰. It is worth noting that even in the state with the highest rate, this figure is still low (0.6%).

Internal Migration in Brazilian states from 1980 to 2010

The data available for internal migration in Brazil in the period 1940-1970 indicate that the most important flows were from the states of the Northeast region and from MG to the states with higher urban and industrial growth (mainly SP and RJ) and the frontiers

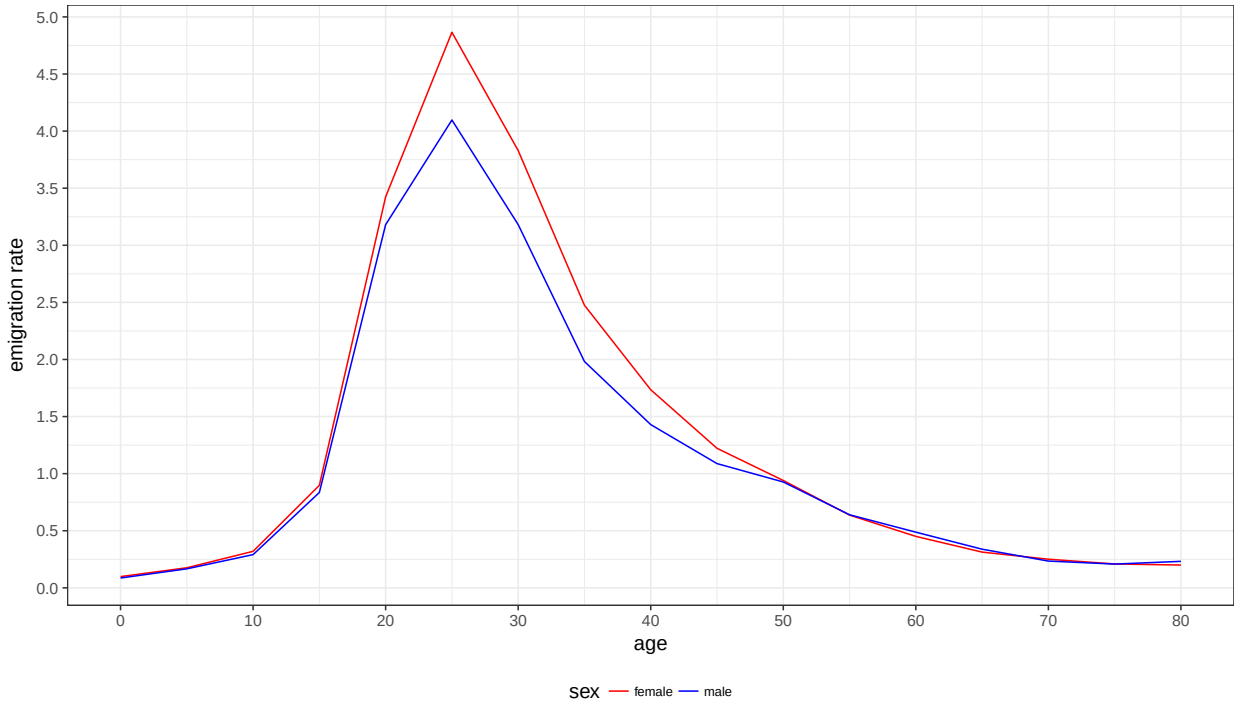

Figure 4.38: Emigration rate (‰) by age and sex, 2006-2010, Brazil. Source: IBGE, 2010 Censuses

of agricultural expansion in the North and Midwest regions. The hegemonic trajectory to SP remained while the migration to RJ reduced in the 1970s. In the 1980s, immigration to SP continued, but there was an important increase in the emigration from the state as well, partially due to return migration of immigration flows in previous periods (Brito, 2000)

Based on the question about the place of residence five years preceding the census date, asked in the 1991, 2000 and 2010 censuses, it is possible to estimate the interstate in-migration, out-migration and net internal migration.

Total in-migration ($im_g^{(y-5;y)}$), out-migration ($om_g^{(y-5;y)}$) and net internal migration ($nm_g^{(y-5;y)}$) rates by geographic region g for the period between years $y-5$ and y, the five-years period preceding the censuses, are defined as follows:

$$im_g^{(y-5;y)} = \frac{\mathrm{IM}_g^{(y-5;y)}}{K_{g,5+}} \times 1000 \tag{4.53}$$

$$om_g^{(y-5;y)} = \frac{\mathrm{OM}_g^{(y-5;y)}}{K_{g,5+}} \times 1000 \tag{4.54}$$

$$nm_g^{(y-5;y)} = \frac{\mathrm{NM}_g^{(y-5;y)}}{K_{g,5+}} \times 1000 \tag{4.55}$$

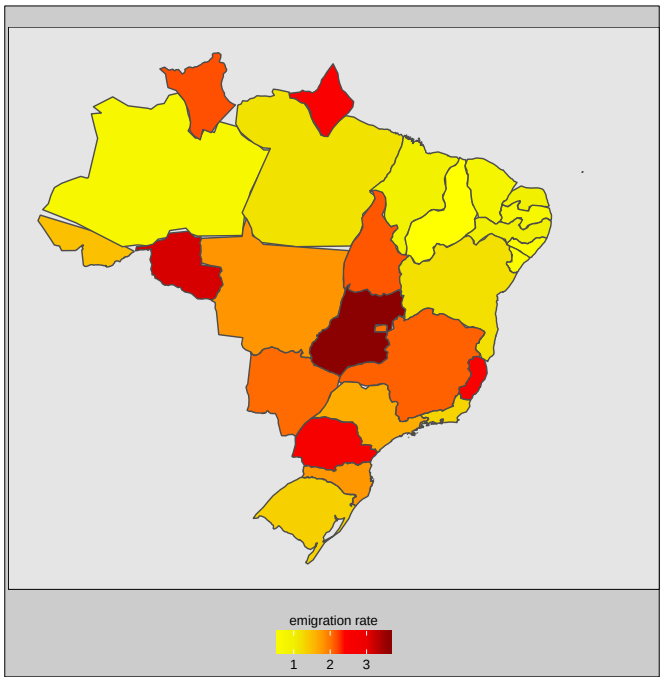

Figure 4.39: Emigration rate (‰) by state, 2010, Brazil. Source: IBGE, 2010 Censuses

where $\text{IM}_g^{(y-5;y)}$ is the number of people living in the state g on the census date that was living in a different state five years before the census; $\text{OM}_g^{(y-5;y)}$ is the number of people that was living in the state g five years before the census date and was living in a different state in the census date; $\text{NM}_g^{(y-5;y)}$ is the difference between in-migration and out-migration ($\text{IM}_g^{(y-5;y)} - \text{OM}_g^{(y-5;y)}$); $K_{g,5+}$ is the population aged 5 years or more in the census date enumerated in region g. The rates im^s and om^s are a proxy of the in-migration and out-migration flows to and from state g in the period of five years preceding the census.

Figure 4.40 shows $im_g^{(y-5;y)}$ and $om_g^{(y-5;y)}$ by state g and year for the last three Brazilian censuses ($y = 1991$, $y = 2000$ and $y = 2010$). The map indicates an overall reduction in migration rates. In fact, in 1991, 3.85% of the population over age 5 lived in a state five years prior to the census that was not the state where they were. In 2000, this proportion reduced to 3.39% and then to 2.62% in 2010.

Despite the reduction in migration rates, the regional pattern of the in-migration and out-emigration across states has remained very similar.

The states with the highest in-migration rates in all periods are those in the frontiers of the agricultural expansion in the North and Midwest regions. In the states of RR, DF, RO, MT, Tocantins (TO), more than 10% of their population lived in a different state five year preceding the 1991 Census. These states also had high in-migration rates in 2000 and 2010. SC was the only state that had an increase in the immigration rate between 2000 (40.9‰) and 2010 (51.6‰), mostly due to an emigration from RS. RS has been one of the states

with the lowest emigration and immigration rates, although there was a moderate increase in the emigration in the period (2005; 2010).

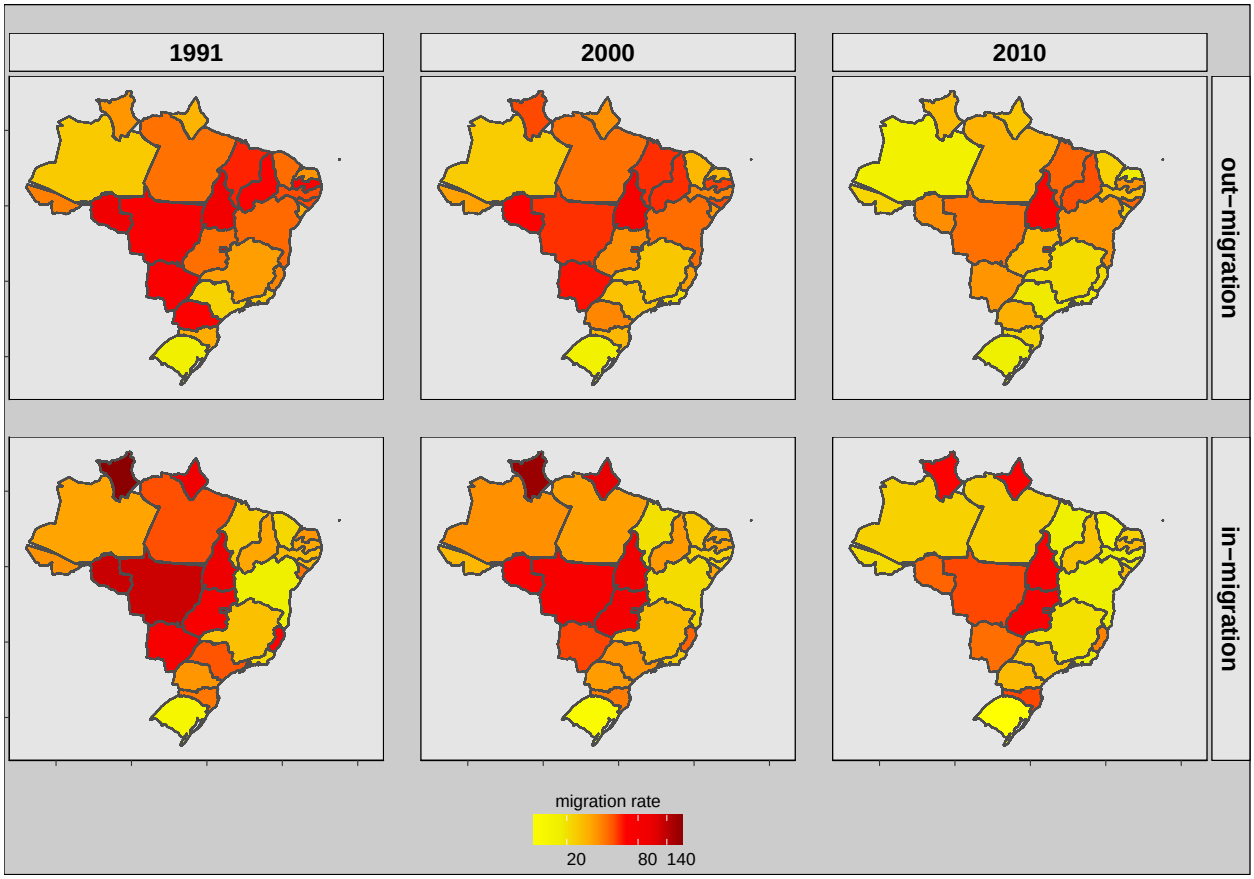

Figure 4.40: Map of the in-migration and out-migration rates by state, 1991, 2000 and 2010 (‰). Source: IBGE, Censuses of 1991, 2000 and 2010

Figure 4.41 shows the net migration rates by state, which are given by the difference between the rates in the two maps above (Figure 4.40). Despite a reduction in the net migration between 1991 and 2010, the states of the Northeast region show negative net migration. The states of MA, AL, PI and BA have had one of the lowest net migration rates. For 2010, net migration rates for these states was negative, meaning that out-migration exceeded in-migration in the period 2005-2010.

A few other states in the Northeast region, such as PB, PE and CE had high negative net migration rates in 1991, but they experienced an increase in the the next two censuses, although they have remained still negative. CE, for instance, had a net migration ($nm_{CE}^{(y-5;y)}$) of -22.3‰ in 1991 and only -3.59‰ in 2000 and -8.82‰ in 2010.

In addition to the Northeastern states, PR also had high negative migration in the period 1986-1991 (-27.4‰).

SP, the most populous state in Brazil, and one of the most developed and industrialized ones, has been primarily a receiving state, although the net migration has reduced significantly from 26.2‰ in 1991 to 10.0‰ in 2000 and 6.63‰ in 2010.

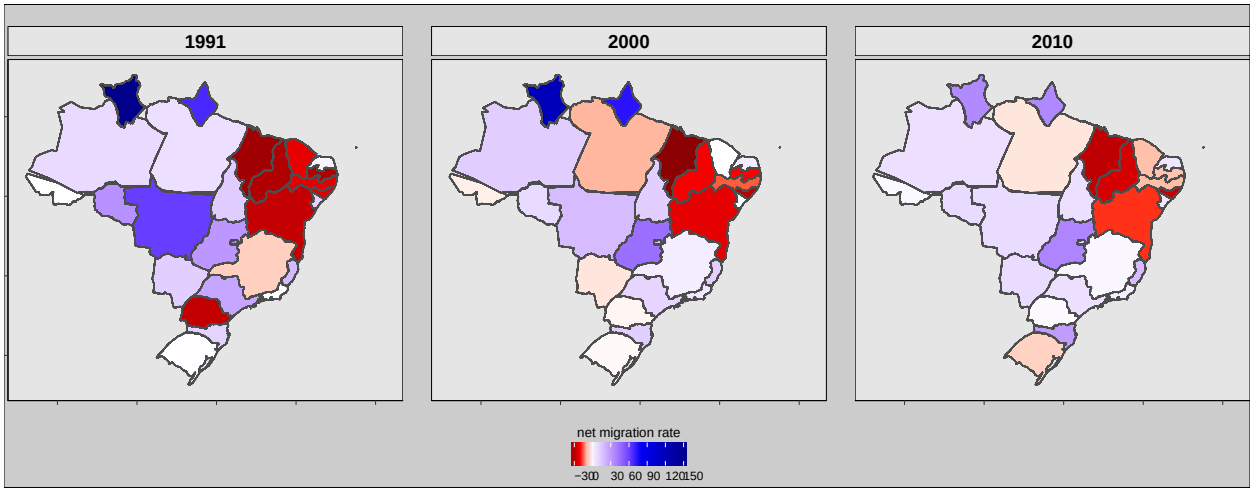

Figure 4.41: Map of the net migration rates by state, 1991, 2000 and 2010 (‰). Source: IBGE, Censuses of 1991, 2000 and 2010

Section C.1 in the Appendix C shows the in-migration, out-migration and net migration rates by sex and age for the 27 Brazilian states and the three censuses under analysis.

The plots show that the age pattern of in-migration is often different from that of out-migration. Out-migration rates peak around the age group 20-24 for the states with negative net migration, such as PI (Figure C.3). For the states with positive net migration, e.g. SP (Figure C.7), the peak of in-migration occurs in the same age groups. The in-migration in sending states and out-migration in receiving states occurs at older ages, around the age group 30-34, although these curves tend to be more spread out across different ages. This pattern is typical of return migration.

The combined effect of different age structure for in-migration and out-migration can lead to a shift in the sign of net migration across ages. The state of RN, for example, have had negative net migration around age group 20-24 due to high out-migration in this group, and positive net migration at older ages due to higher in-migration than out-migration, partially due to return migration.

Migration estimates by age and sex for the most populous states, such as those in the South and Southeast regions, for instance RJ, SP and PR (Figure C.7), are very stable. On the other hand, large fluctuation are seen in small states, such as AP (Figure C.2), due to sampling variation.

Conclusion

This section presented an analysis of international and internal migration in Brazil for the period 1980-2010. International migration has been low, compared to other countries in the world. On the other hand, interstate migration has had an important impact on the population growth and age distribution at subnational level in Brazil.

It is broadly recognized that emigration from Brazil to other countries peaked in the 1980s and reduced in the following two decades. On the other hand, immigration has grown in the same period, with Brazil appearing as a destination of new migration flows and, more importantly, as a result of the return of Brazilians who had emigrated in the first periods of international migration flows. As a result, the net migration rate seems to have been negative in the 1980s and have approximated to zero in the decade 2000/2010.

International net migration rates have been estimated by using indirect residual methods. These methods have been useful to point out to the magnitude and direction of migration flows, but have several limitations: they are extremely sensitive to the choice of the intercensal mortality rates and the measures of census undercount. These often lead to an estimated age structure which is inconsistent with the expected pattern of migration, and may be actually indicating problems of census coverage, age misstatement and/or errors in mortality estimates.

Thus, estimates of mortality and census coverage, with their respective measures of uncertainty, should be incorporated in a unified framework along with with independent information on immigration and emigration. Then, plausible scenarios of migration can be derived. A proposal of methods to do so is discussed in Chapter 3 and an application to Brazil is shown in Chapter 5.

Contrary to international population flows, interstate migration has shaped demographic changes at the subnational level in Brazil. Despite the reduction in internal migration in the last decades, this is still a relevant phenomenon for almost all Brazilian states. Internal migration flow are also better documented due to a rich questionnaire on this topic in Brazilian censuses.

Internal migration has occurred primarily from poor states in the Northeast region to more developed states in the Southeast region. Furthermore, important flows have been reported to the frontiers of agricultural expansion in the North and Midwest regions. Internal migration is, obviously, much more complex, and include other important flows of return migration, e.g. to the Northeast, and new flows, e.g. to SC.

The question about place of residence five years prior to the census seems to provide a a useful measure of recent in-migration and out-migration for population estimates. Since this question has been included only in the long form of the censuses, it contains high sampling variation, particularly for the less populous states. Thus, sampling error should be used as a measure of uncertainty in these estimates.

4.6 Summary

This chapter presented several estimates of the coverage of censuses and CRVS, as well as mortality, fertility and migration, discussing the advantages and limitations of each one. The next section will present the final estimates and the corresponding measures of uncertainty to be used in the application of the methods discussed in Chapter 3.

Chapter 5

Case Study from Brazil

This chapter shows an application of the method described in the previous chapters to Brazil and three selected states (RS, PB and RJ). Brazilian states are interesting case studies to test the validity of the method under different circumstances, since they are remarkably diverse demographically and socioeconomically. They also differ in regard to the quality of their statistics.

5.1 Method

Chapter 3 discussed in detail the methods used to make inference about the demographic parameter of interest, which consists of first estimating population, fertility, mortality and migration, and then reconciling these premodel posterior distributions estimates by using the demographic balancing equation.

This section discusses the specifics of the application of the methods based on the data presented in chapter 4 and proposes an extension of the Bayesian melding approach for reconciling information from three years.

This extension combines the general idea of the Bayesian melding approach with the current practice of deterministic methods for reconciling demographic data from multiple years, for instance the demographic reconciliation approach (Chackiel, 2009; Gerland, 2014) and the "three-census method" proposed by Ntozi, (1978).

In the illustration presented in this chapter, reconciliation is performed by using Brazilian demographic data from 1990 to 2010. See a detailed discussion of the data and estimates available for this period in chapter 4.

The estimates are produced for the female population by five year age groups and five year intervals for Brazil and three states with different demographic and data quality characteristics.

Inference about the parameters of interest for the three-census case is also done in two steps. First, samples from the posterior distributions of the population, fertility, mortality

and migration for the years 1990, 2000 and 2010, called premodel posterior distributions, are drawn.

Then, the 1990 population is projected to 2000 and population of 2010 is backprojected to the same year. The premodel posterior distribution for the population in 2000 is also used. It is estimated based on the 2000 census and its census coverage information from the PES. Reconciliation consists of harmonizing the premodel posterior distribution for 2000 with the two probability distributions induced by the premodel posterior distributions of 1990 and 2010. These are projected and backprojected using the distributions of the demographic events during the two intercensal periods. Reconciliation is done by using the Sampling/Importance Resampling (SIR) algorithm.

Another extension to the original method proposed in this dissertation is to estimate simultaneously several age groups. In the first step, when the premodel posterior distributions are estimated, there is correlation between age groups, for instance when mortality is modeled by using the TOPALS relational model. This is important to stabilize estimates, particularly for small population groups, and also to take into account the known pattern of demographic rates.

In the second step, reconciliation is done independently by age group. The intuition behind this is related to the particularities of inconsistencies by age group. The population of children under age five in 2000 backprojected from 2010, for example, may be more reliable than the estimated based on the 2000 census, due to high undercount among this group. On the other hand, the cohort aged 10-14 in 2000 and 20-24 in 2010 is perhaps better estimated based on the 2000 Census.

The posterior distributions of the adapted Bayesian melding approach is approximated by drawing samples according to the following steps:

1. Draw a sample of j values of the inputs ($\boldsymbol{A^y}$, $\boldsymbol{K(1990)}$, $\boldsymbol{K(2010)}$) and the output ($\boldsymbol{K(2000)}$) from their likelihoods and prior distributions. $\boldsymbol{A^y}$ contains information about fertility, mortality and migration for the periods $y = \{1990/1995;\ 1995/2000;\ 2000/2005;\ 2005/2010\}$. For the projection years (1990/1995 and 1995/2000), $\boldsymbol{A^y}$ is the Leslie matrix, plus migration. For the backprojected period (2000/2005; 2005/2010), $\boldsymbol{A^y}$ contains only mortality and migration information.
2. Calculate the unnormalized premodel posterior distributions of population in 1990 and 2010 and demographic events between 1990 and 2000 and between 2000 and 2010, which is formed of the sample of size j.
3. Calculate the unnormalized premodel posterior distributions in 2000, which is also formed of the sample of size j.
4. Determine the induced sample on the output for each of the j values generated in *Step 2* by running the demographic projection and backprojection models. In the extension proposed in this dissertation, this is a joint probability distribution formed by the two induced distributions.
5. Use nonparametric two-dimensional Kernel Density Estimation (KDE) to obtain estimates of the distributions of both populations in 2000 to be harmonized: the joint induced distribution and the premodel posterior distribution for the year 2000.

6. Form the importance sampling weights to be used for melding the two posteriors in 2000 using logarithmic pooling with polling weight 0.5.
7. Sample k values from the discrete premodel posterior distributions samples in *Step 2* with values proportional to the importance sampling weights calculated in *Step 6*.

The resampling process described in *Step 7* is done for the corresponding cohorts used for generating the populations in 2000. The weights calculated for the cohort aged 20-24 in 2000, for example, are used to resample the population aged 10-14 in 2000, as well as the migration and mortality estimates for the same cohort in the intercensal periods. There is a one to one correspondence between the estimated population in 2000 and the populations in 1990 and 2010 and the intercensal migration and mortality estimates.

The weights for the age group 0-4 are used to reestimate fertility for the period 1995/2000 and the weights for the age group 5-9 are used to reestimate fertility for 1990-1995. This is only an approximation, since the estimated population of children also depends on the number of women at reproductive ages, as well as child mortality. Since fertility estimates for the period 2000/2010 do not enter in the model, they are not resampled, and the estimates presented in the results for this period are the premodel posterior distributions.

5.2 Population Estimates for Brazil from 1990 to 2010

The main objective of this dissertation is to produce demographic estimates that are consistent over time and across data sources and estimation methods. Figure 5.1 shows the populations enumerated in the three Brazilian censuses under analysis [1] and the projected and backprojected populations from the other two censuses. Projections and backprojections use mortality calculated via using VS adjusted by DDM and fertility estimates from the adjusted P/F ratio method. These estimates assume no international migration. As the figure shows, there are serious inconsistencies in the estimates. In 2000, for instance, the population of children below 10 years old enumerated in the census is significantly lower than that estimated by backprojecting the population aged 10 to 19 in 2010 and that generated by projecting the population from the 1990, which includes fertility estimates. The population from 1990 is even higher than the backprojected from 2010. This may be due to an overestimate of fertility rates or overestimation of the population at reproductive ages.

The figure shows that women at reproductive ages projected from 1990 is higher than the enumerated population in the 2000 census and in the backprojected population from the 2010 Census. This is probably indicating that international migration needs to be incorporated in the population estimates, and disregarding the negative net migration flows in the 1990s overestimates the population of young adults in 2000. It is also clear in the plots the underestimation of the cohort of young children in 1990. At older ages, there seems to exist an underestimation of women in the 50s and 60s, possibly due to age misstatement. There was probably age understatement of women around 50 in 1990

[1] interpolated to guarantee exactly ten year intervals

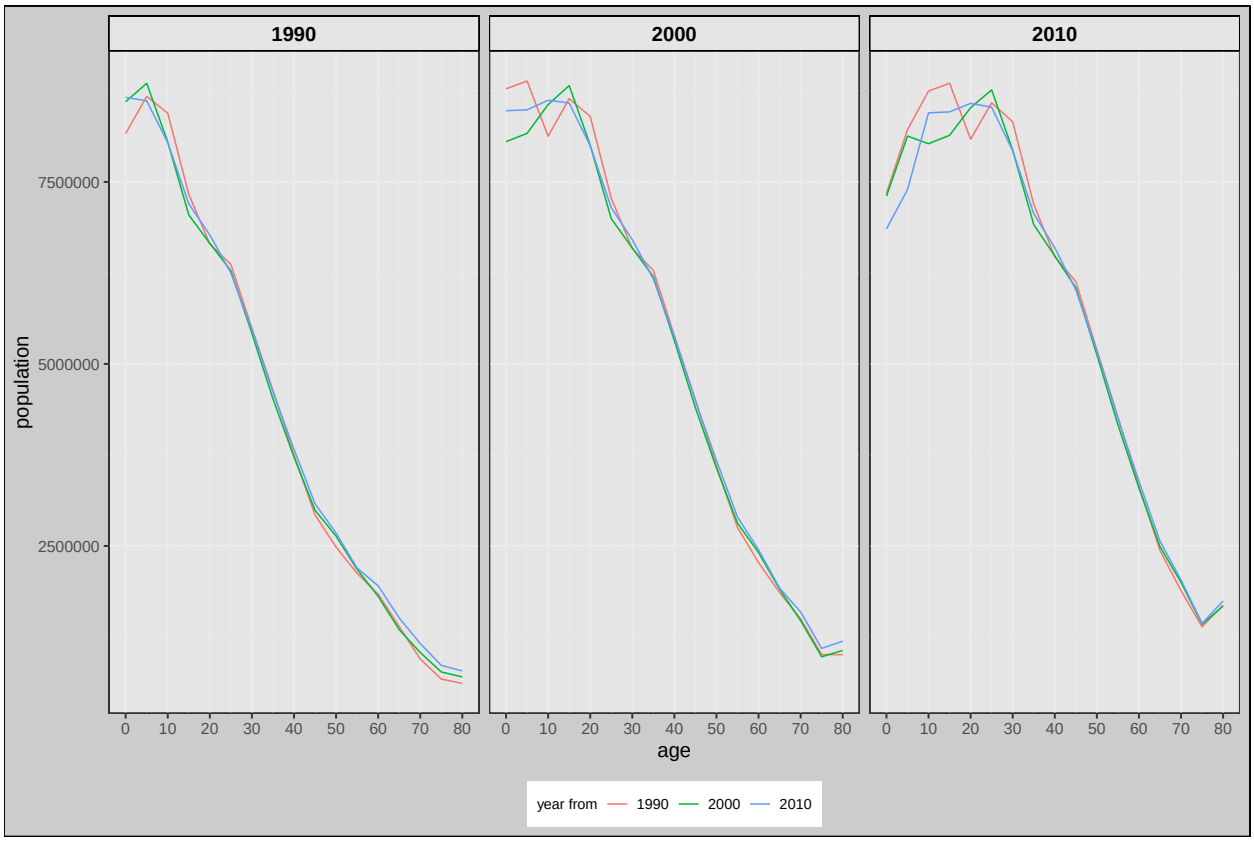

Figure 5.1: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

Population

The illustration for Brazil uses the censuses of 1991, 2000 and 2010, extrapolated or interpolated to January 1st of 1990, 2000 and 2010. Prior estimates for the census undercount are derived from the results of the PES of the 1991 and 2000 censuses. The 2010 PES has not been published yet.

As discussed in chapter 4, the indicators published by the PES of different years are not always consistent. For example, the 1991 PES published no estimates about erroneous inclusion. The use of the published indicator would overestimate the net undercount of that census. The overall undercount rate is very similar for the 1991 and 2000 censuses. Demographic analysis has also shown that the coverage of Brazilian censuses since 1980 seems to be similar. Given this scenario, the hyperparameters of the beta distributions for 1990, 2000 and 2010 were chosen based on the overall omission rate for the 2000 census.

An adjustment factor increasing which increases with age, starting at age 50 is applied to take into account overcount due to age misstatement. The mean of this factor varies linearly from 1 to 1.1 between ages 50 and 80, and are modeled as a gamma distribution with variance equals 0.001. These factors are roughly the factors found by (IBGE, 2013b)

for Brazil in 2000. These probability distributions are fixed for all states.

Fertility

Fertility estimates comes basically from two sources. First, the TFR is estimated based on the the adjusted P/F ratio method. The TFR estimated for the censuses are then interpolated to the years 1990, 2000 and 2010. The estimated TFR for these years in Brazil are 3.07, 2.38 and 1.74. These average TFR are combined with a variance of 1% to form the Gamma distribution to which fertility is modeled. The distribution for 2000, for example, has a 10th percentile of 2.25 and the 90th percentile of 2.51.

The second data source used to estimate fertility is the proactive search survey, which gives an estimate of the completeness of registered births. The estimated completeness of registered births is 0.925 for 2000 and 0.959 for 2010. These estimates are used as the mean of the Beta distribution to model the completeness of registered births, which are assign a variance of 0.05%. These lead to the 10th percentile of 0.895 and the 90th percentile of 0.952 for 2000 and 0.916 and 0.991 for 2010, respectively.

The number of registered births, B_c^{obs}, is then modeled by a Poisson distribution, as discussed in Chapter 3:

$$B_c^{obs} \sim Poisson(K_c \cdot f_c \cdot \beta_c) \tag{5.1}$$

where K_c is the population at age group c, f_c is the ASFR and β_c is the completeness of registered births. Prior distribution for fertility rates and completeness of registered births are defined as described above. Instead of modeling each ASFR, TFR is modeled and then apportioned deterministically proportional to the participation of each age group in the TFR observed in the respective censuses. The completeness of registered births is considered to be constant by age, so that $\beta_c = \beta$ for every c.

Mortality

Estimates to produce mortality indicators come from multiple sources. Two measures of completeness of registered deaths are used. First, the completeness of the registered deaths relative to the completeness of the census is estimated by the application of the DDM: SEG for the period 1980/1990 and the combined GGB-SEG for the other periods. The estimated intercensal factors are interpolated to the census years. The hyperparameters a_c^D and b_c^D of the gamma distribution are estimated so that the mean of the gamma distributions match the mean of the estimate given by the DDM method. Given the limitations of these methods, these results enter the model through relatively weak priors, which are found by choosing relatively high variance.

The second estimate of the completeness of registered deaths come from the proactive search survey. This is modeled as a Beta distribution, with mean given by the survey result and variance equal to 0.02. The estimated completeness is 0.91 for 2000 and 0.94 for

2010. Since there is no information for 1990, an ignorance prior is used: a beta distribution $Beta(1,1)$, which is equivalent to a $Unif(0,1)$.

In addition to the completeness of registered deaths, child and adult mortality are also estimated through indirect demographic. Child mortality (${}_5q_0$) is estimated by using the Brass method and adult mortality (${}_{45}q_{15}$) is estimated though the maternal orphanhood method, as described in Chapter 4.

Based on these estimates, the complete life tables are estimated by finding a correspondent model life table. In this study, the Coale and Demeny West life table is used (Coale and Demeny, 1966). These mortality rates then enter the model through a relatively weak prior, as the following Gamma distribution:

$$h_c \sim Gamma(10, \frac{10}{h_c}) \tag{5.2}$$

Finally, mortality rates are smooth by using the TOPALS relational model with a standard mortality schedule given by the female life table published by IBGE for 2000.

Thus, mortality is estimated by using several data sources, and uncertainty around the estimates should reflect the availability and reliability of the estimates. Estimates for 1990 should have more uncertainty, since there is no direct estimate of completeness of registered deaths.

Migration

International immigration and emigration are estimated separately, and modeled by a gamma distribution. Data to estimate immigration come from censuses, which provides point estimates for i_c^{obs} from the question about place of residence five years prior to the census reference day. Since this question is available for all censuses of the period under analysis, they can be used to estimate immigration to Brazil. The hyperparameters for international immigration are similar to those used to model mortality rates.

There are much more uncertainty about international emigration from Brazil. The 2010 Census asked for the first time about emigration. Thus, this information is used to produce an estimate of the emigration flows for the five years prior to the 2010 Census. For 2000, the same age structure of 2010 is assumed, with emigration rates twice as high and that of 2010, procedure replicated for the year 1991. A much weaker prior is assigned to the emigration rates. These weak priors would allow emigration rates to be reestimated based on the consistency between censuses, at the same time they set plausible age structure for the phenomenon.

The distributions for i_c and e_c are defined as below:

$$i_c \sim Gamma(10, \frac{10}{i_c^{obs}}) \tag{5.3}$$

$$e_c \sim Gamma(1, \frac{1}{e_c^{obs}}) \tag{5.4}$$

Results

Figure 5.2 shows the premodel posterior distributions for the three years under analysis, as well as the consistent postmodel posterior distributions. The figure shows that the estimated population tends to be higher than the enumerated populations in the censuses, particularly for children and young adults, which indicate high omission rates at those ages. The postmodel posterior distribution is slightly narrower than the premodel distribution, indicating that the reconciliation of data from different sources reduces uncertainty. The reconciliation reduces significantly the uncertainty about the population of children, for example.

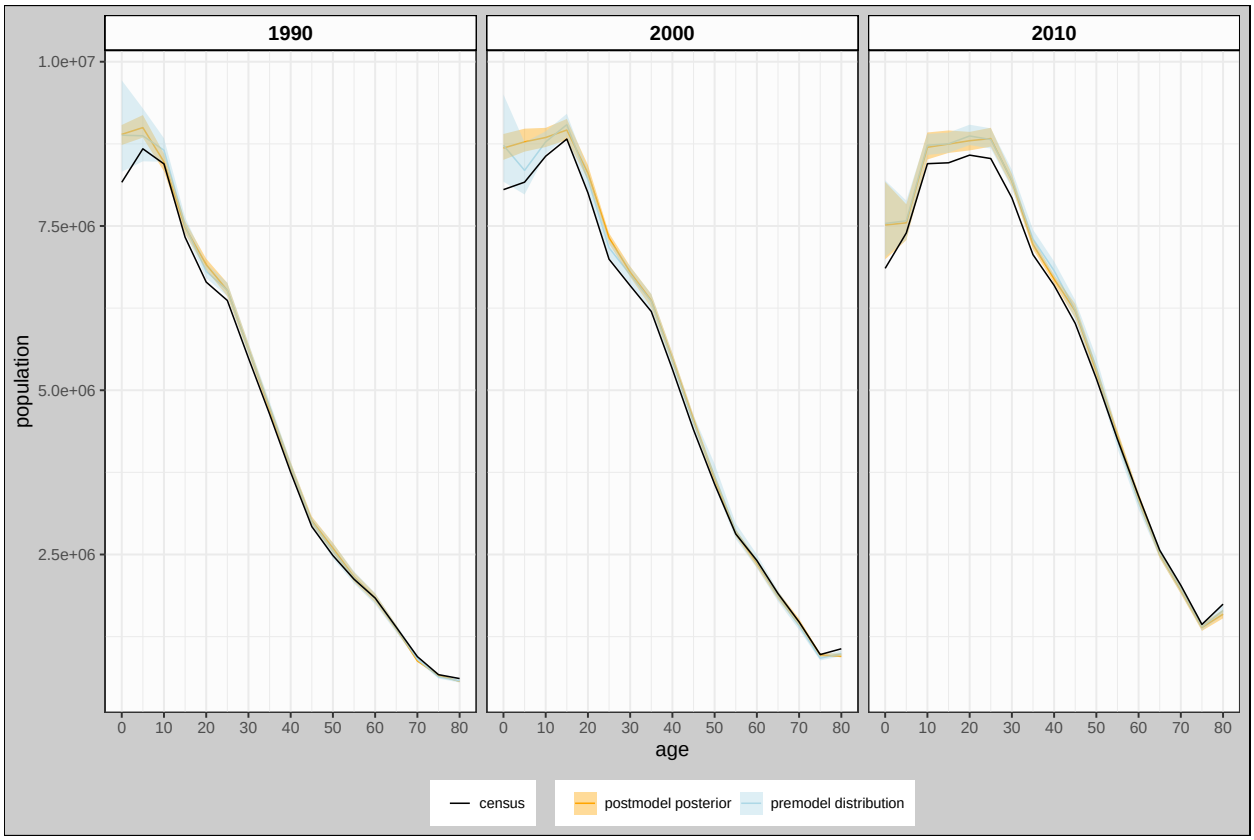

Figure 5.2: Population estimates by age group: comparison between the census, the premodel and the postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

These estimates result in a median undercount of 2.75% for 1990, with 80% prediction interval of [1.06%; 4.47%]. The median undercount was 3.26% for 2000 [1.60%; 5.03%] and 2.40% [0.16%; 4.48%] for 2010. These results confirm that there is no evidence of significant differential undercount across different censuses in Brazil. The census coverage rates by age group are very similar for all years, and the differences in the overall census undercount reflect more the changes in the age structure of the population. The fact that there are fewer children in 2010 contributes to the lower overall undercount rate in this year.

Figure 5.3 shows the premodel posterior distributions of net migration for the four quinquennium under analysis, as well as the consistent postmodel posterior distributions. The figure shows that posterior distributions are not very different from the premodel distributions and confirms the reduction in the international emigration in Brazil. For the periods 2005/2010, the postmodel distribution has slightly more negative migration in the age group 25-29 and more positive migration at older ages, possibly return migration, than the premodel distribution. Net migration in the period 2000/2005 is younger than the premodel distribution. The postmodel distribution for the 1990-1995 has a peak in the age 25, that should be investigated to confirm whether this is likely to be real or a census error that was not entirely captured by the model.

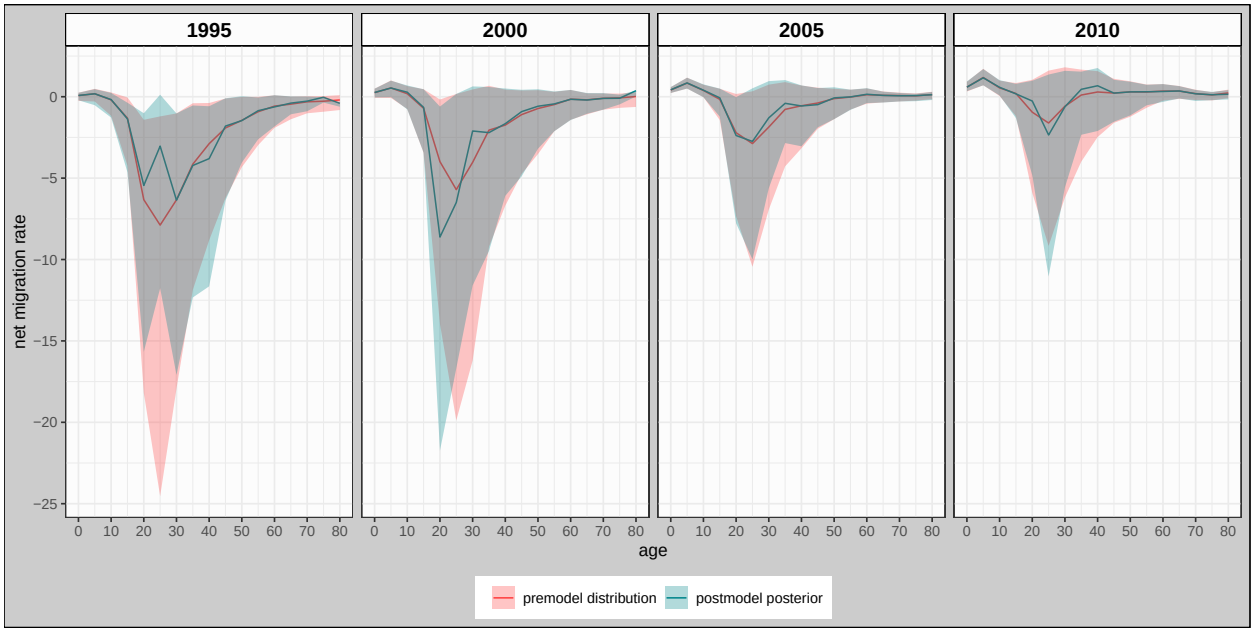

Figure 5.3: Net migration rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

Figure 5.4 compares both distributions for mortality estimates. The figure shows that there is not much uncertainty regarding mortality estimates, and both the premodel and the postmodel distributions are similar. Uncertainty is higher in the 1990s, when there is less

information. For example, there is no direct estimates of completeness of registered deaths for this period.

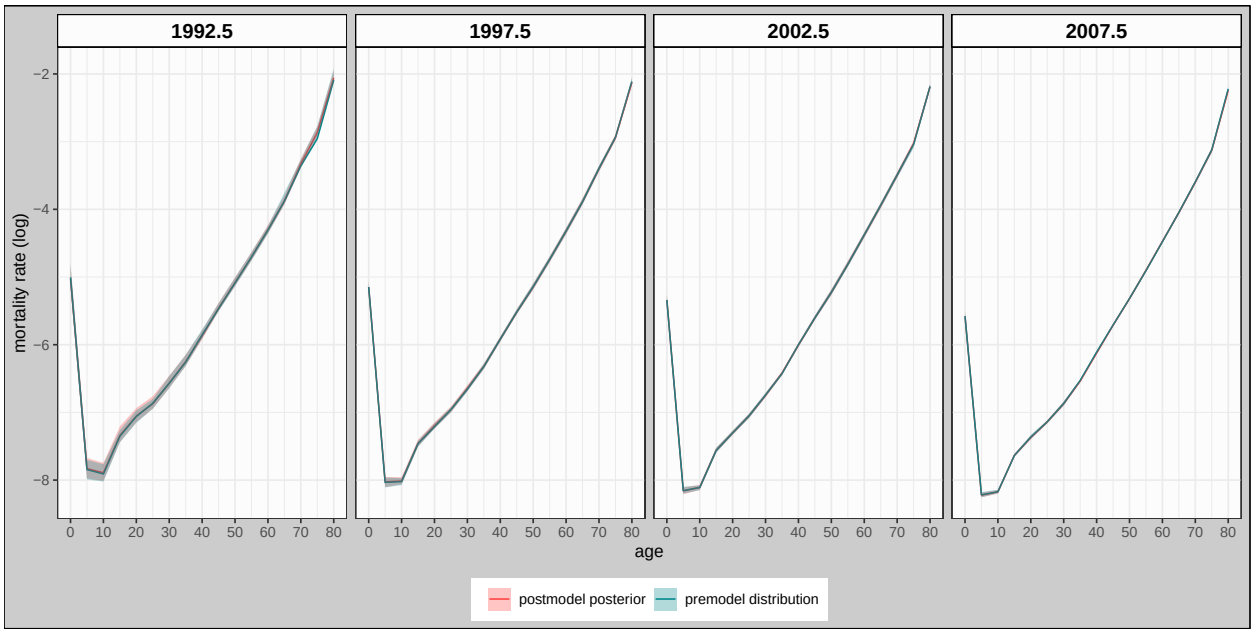

Figure 5.4: Age-specific mortality rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

Figure 5.5 shows the comparison between the official estimates of life expectancy and the estimates produced by the methods proposed in this dissertation. Official estimates are consistently lower than the estimated here, even below the lower bound of the prediction interval, with higher difference for early years. This is probably due to the difference in the estimates of completeness of registered deaths, which tend to be lower in IBGE estimates.

Finally, figure 5.6 compares both distributions for fertility estimates, showing that reconciliation did not change much the premodel distribution. The figure shows that there is not much uncertainty regarding fertility estimates as well. Fertility has decline significantly in the period under analysis, and the age schedule has also varied slightly, although the modal age of the fertility schedule remains in the age group 20-24.

Figure 5.7 shows the comparison between the official estimates of the TFR with the postmodel posterior estimates. IBGE estimates are almost always within the prediction interval.

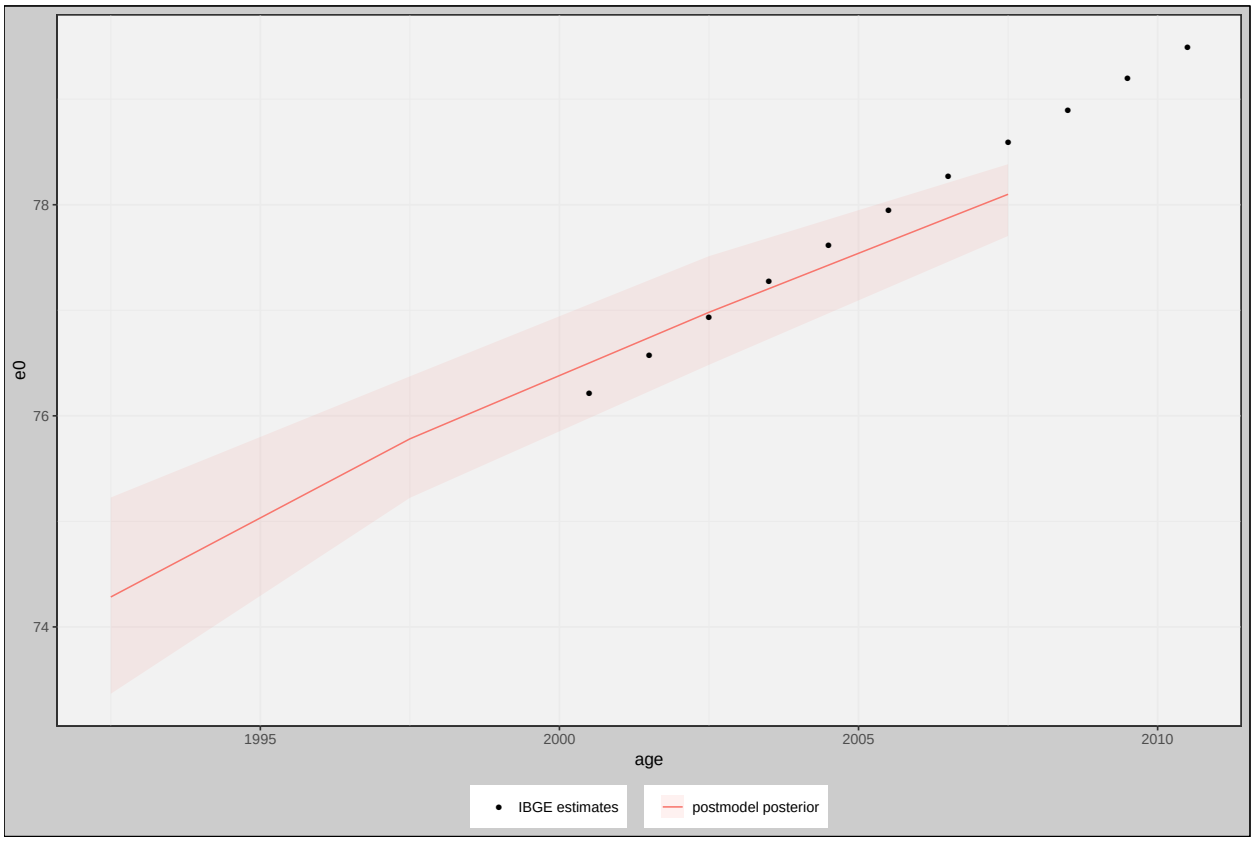

Figure 5.5: Life expectancy at birth: comparison between postmodel posterior with IBGE estimates, Brazil, female population, 1990, 2000 and 2010

5.3 Population Estimates for Brazilian states from 1990 to 2010

Similar to what was shown for Brazil, section E.1 in the Appendix E shows plots for the enumerated, projected and backprojected populations for 1990, 2000 and 2010 for the 27 Brazilian states. The projections and backprojections use the best information available for intercensal fertility, mortality and migration and make no adjustment in the censuses populations.

The figures show that most of the states have important inconsistencies, such as RO, AP, PI, PB, RJ. Almost all states have inconsistent estimates for the population of children, as well as among adults. Only a few states have roughly consistent estimates, such as RS and SC. The application of the method proposed in this study to adjust for these inconsistencies will be discussed below with illustrations for three of these states.

The estimation procedures and the data used for Brazilian states are similar to those used for the country as a whole. The main difference is regarding migration, which procedure for

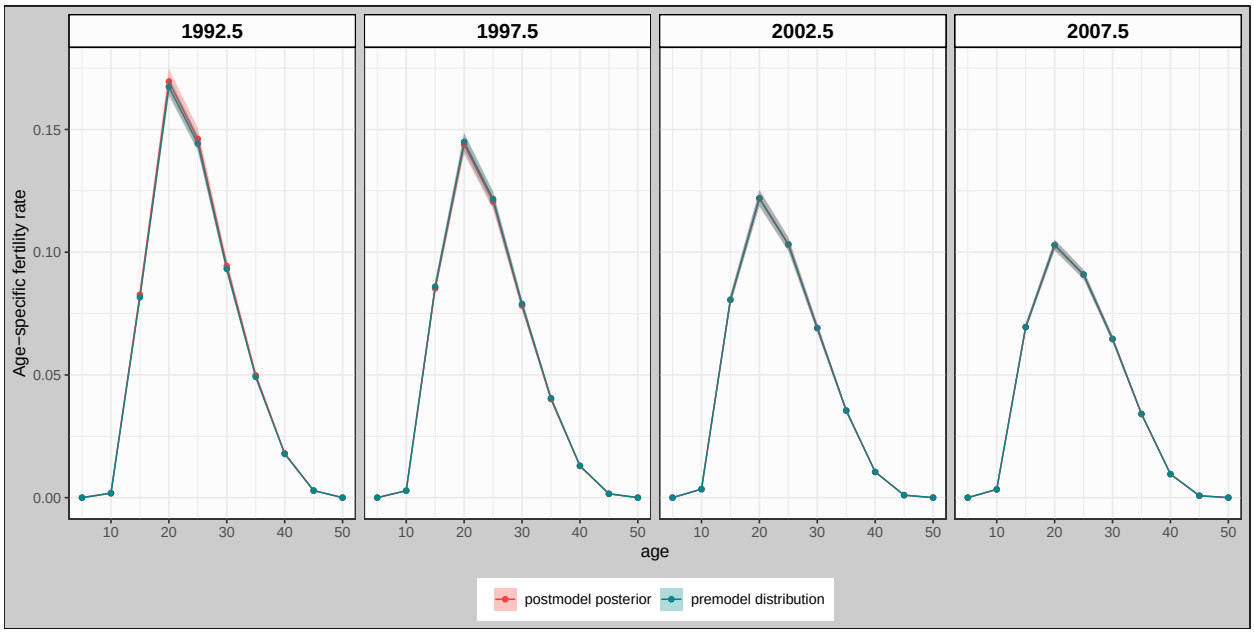

Figure 5.6: Age-specific fertility rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

producing subnational estimates is described below.

Modeling internal migration

Before applying the method to Brazilian states, internal migration needs to be model. Brazilian censuses provide information on in-migration and out-emigration for the five year periods preceding the census. Although this is a very useful information for population estimates, it has a few limitations.

First, there are large fluctuations in internal migration estimates for small states, as shown in Figure C.2, due to sampling variation. To overcome this issue, in-migration and out-migration rates were modeled by using the TOPALS relational model. As discussed in chapter 5, TOPALS relates rates of a certain region being modeled to a standard schedule by using splines. The standard migration schedule used to model in-migration and out-migration is the total internal migration flow calculated by using the 1996 Count, which was the only time this question was asked to the total of the Brazilian population in the short form of the census.

The second limitation is that this data is only available for the quinquennium before the censuses. Thus, migration needs to be estimated for the other five-year intercensal periods. The results for two consecutive censuses show that interstate migration flows have been fairly regular in Brazil across years, indicating that migration for this period of lacking data can

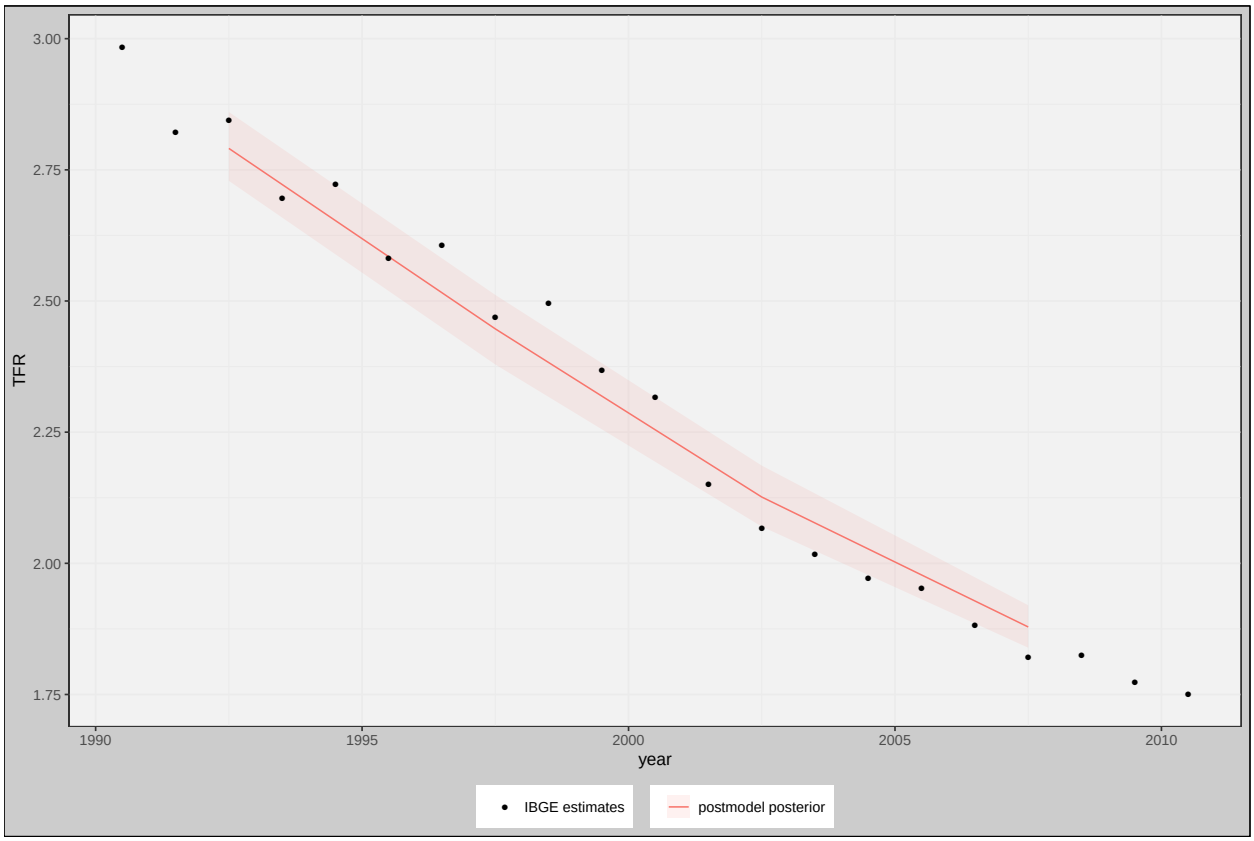

Figure 5.7: TFR: comparison between postmodel posterior with IBGE estimates, Brazil, 1990, 2000 and 2010

be interpolated by using the available information for two consecutive censuses. Instead of a simple linear interpolation, this study proposes a random walk interpolation, where the intercensal migration flows are normally distributed with mean given by the interpolated mean between the two censuses and the variance is given by the average variance between the two censuses multiplied by the distance between the interpolation year and the nearest census.

Section E.2 in the Appendix E shows the fitted five-year in-migration and out-migration rates for all Brazilian states from 1990 to 2010. The figures show much smoother rates, with higher uncertainty at older ages and for smaller states, for instance AC, RO and AP. Uncertainty is higher for early years, e.g. 1990, as well as for the years further from the census years, e.g. 2005.

Two limitations remain. First, although the method estimates uncertainty for regions with higher variation, sampling variation is not explicitly incorporated into the model. Future work should investigate ways to incorporate sampling variation into the results by using the sampling design of each census. The second limitation is inherent to the data and is related

to differential census undercount across states. If there is an imbalance in census undercount in the state of origin and the states of destination, the net migration will be biased. More specifically, in-migration tends to be unbiased, even if there is census undercount, since the numerator is consistent wit the denominator [2]. Out-migration rates, on the other hand, depend on the census coverage in the states of destination, as the number of out-migrants used to calculate the out-migration rates is the sum of the migrants reported in all states but that where the rates are being calculated. To give an example, take a state in the Northeast with high migration rates and relative low census undercount. The in-migration rates to that state is nearly unbiased, whereas the out-migration rates might be biased downwards if the direction of the migration flows is towards states in the Southeast region, particularly to highly populated cities with higher census undercount.

The migration data that comes from the censuses, adjusted and interpolated as described below, are used as prior distributions for the model. However, to take into account the limitations below, extra variability was incorporated to the estimates. This extra variance allows migration being re-estimated based on the consistency between censuses. Thus, the mean in-migration and out-migration are estimated as previously described, but the variance is given by a normal distribution $N(0, 0.5)$ is the log scale. This also allows international migration to be incorporated into the model, which is harder to estimate for states.

The subsections below show the results for three Brazilian states

[2] under the assumption that undercount is not differential by migration status

Demographic estimates for Rio Grande do Sul

RS is a developed state located in the Southern region of Brazil. It has relatively good vital statistics, low census undercount and low migration rates (see Chapter 4).

Estimates of completeness of registered births and deaths indicate that vital statistics systems in this state are nearly complete.

The figure for RS (E.8) in the section E.1 of Appendix E shows that population enumerated in the 2000 Census is consistent with the population projected from 1990 and backprojected from 2010. The main demographic inconsistency in RS is among women in the age groups 25-29 and 30-34 in 2000. For these groups, there are fewer women projected from 1990 than the enumerated in 2000 and backprojected from 2010.

The results of the application of the proposed method for the population estimates for the years 1990, 2000 and 2010 are shown in Figure 5.8. As discussed below, population estimates for RS were already highly consistent, even before the application of the method. The figure shows that the premodel posterior distributions for 1990, 2000 and 2010 were very close to the consistent postmodel posterior estimates. Prediction intervals are narrower in the postmodel posterior than in the premodel posterior. By using the information for the 2000 and 2010 censuses, and the plausible scenarios for the intercensal demographic events, the method was able to adjust for one of the few inconsistencies that existed, that is, the population around age 20 in 1990.

The median undercount for the 1990, 2000 and 2010 censuses were 1.44%, 1.26% and 1.29%, confirming that undercount rates of censuses in RS are low and similar across years.

Figure 5.9 compares the estimated median population and the 10th and 90th percentiles with official estimates produced by IBGE for the period under analysis. IBGE estimates are within the 90th percentile prediction interval, but the point estimates produced by IBGE are higher than the median postmodel posterior estimates. This difference is probably due to the method used by IBGE (IBGE, 2018) to adjust the base population, which applies to all states the same factor that accounts for census errors found in Brazil. Since the PES of the censuses for RS indicate low census undercount, population estimated for this state by using the method proposed in this study is lower.

Figure 5.10 shows the premodel and postmodel estimates of the five-year net migration rates by age groups. The two sets of estimates are not very different, but the postmodel estimates are more unstable, which is partially due to the low migration rates in RS.

Figure 5.11 compares mortality estimates from both distributions and shows that there is not much uncertainty regarding mortality estimates.

Figure 5.12 shows the comparison between the official estimates of life expectancy and the estimates produced by the methods proposed in this dissertation. IBGE estimates are within the 80% prediction interval, but the time trend is different. IBGE estimates have a steeper increase in life expectancy than the estimated by this study.

Premodel and postmodel fertility estimates are also consistent for RS, as show by figure 5.13. Figure 5.14 shows that posterior estimates of the TFR are also consistent with IBGE estimates for most of the period of comparison.

Figure 5.8: Population estimates by age group: comparison between the census, the premodel and the postmodel posterior distributions, RS, female population, 1990, 2000 and 2010

Demographic estimates for Rio de Janeiro

RJ is also one of the most industrialized and rich states un Brazil. It also has relatively reliable vital statistics. Contrary to RS, RJ seems to have had relatively high census undercount, according to the PES. The population of the state is highly concentrated in the metropolitan area of the capital, which is presumably a hard-to-count area. The state of RJ also has relatively high in-migration rates, mostly from states in the Northeast region.

The figure for RJ (E.7) in the section E.1 of Appendix E shows that there are serious inconsistencies between the three censuses under analysis and the intercensal demographic estimates. The figures indicate high undercount of children under age 10 for both the 1990 and 2000 censuses. Data are also inconsistent for the population below age 40 in 2000. Population backprojected from 2010 is consistently higher than the populations in 1990 and 2000. There are also important differences in estimates for the population aged around 60 years old in 2000.

To produce the premodel posterior distributions, prior distributions for the 1991 and 2000 census coverage come from the PES of those censuses. Since there is no information

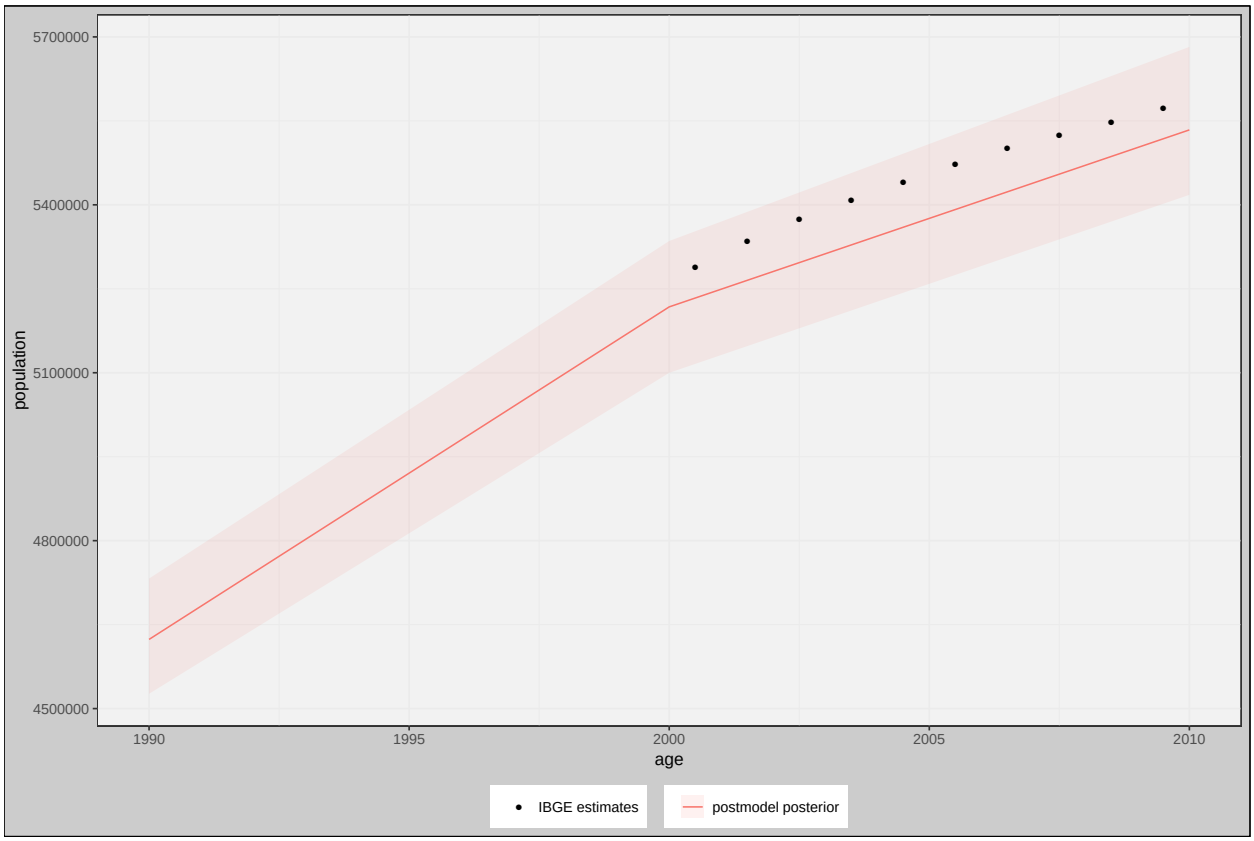

Figure 5.9: Total population: comparison between the postmodel posterior distributions with IBGE estimates, RS, female population, 1990, 2000 and 2010

about the coverage of the 2010 Census, the estimates of the 2000 Census are used, even though the consistency plot indicate that undercount is lower for 2010. The reconciliation step should be able to reestimate undercount of this census is this is the case.

The results of the application of the proposed method for the population estimates for the years 1990, 2000 and 2010 are shown in Figure 5.15. Results indicate high census undercount in 1990 of children, young adults and woman aged around 50 years old. There are significant children and adult undercount in 2000 as well. The 2010 postmodel posterior estimates for 2010 are situated around the enumerated census counts. The postmodel estimates for 2000 are adjusted upwards after the reconciliation step and the estimates are adjusted downwards for 2010 after the reconciliation.

The median overall census coverage is of 3.18% in 1990, 1.31% in 2000, and -0.81% in 2010, which indicates a small net overcount in the 2010 Census. It is worth mentioning that RJ was one of the states with the highest count imputation in 2010 (2.26%), which may explain the apparent low undercount in 2010.

Figure 5.16 compares the estimated median and the 10th and 90th percentiles with official

Figure 5.10: Net migration rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, RS, female population, 1990, 2000 and 2010

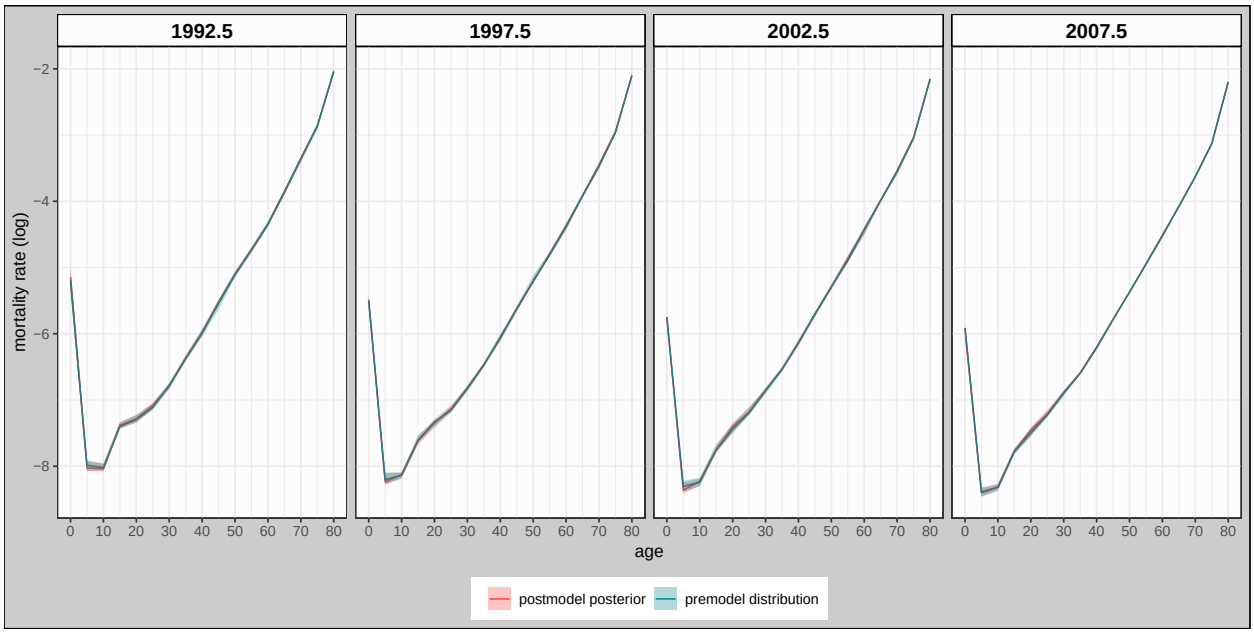

Figure 5.11: Age-specific mortality rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, RS, female population, 1990, 2000 and 2010

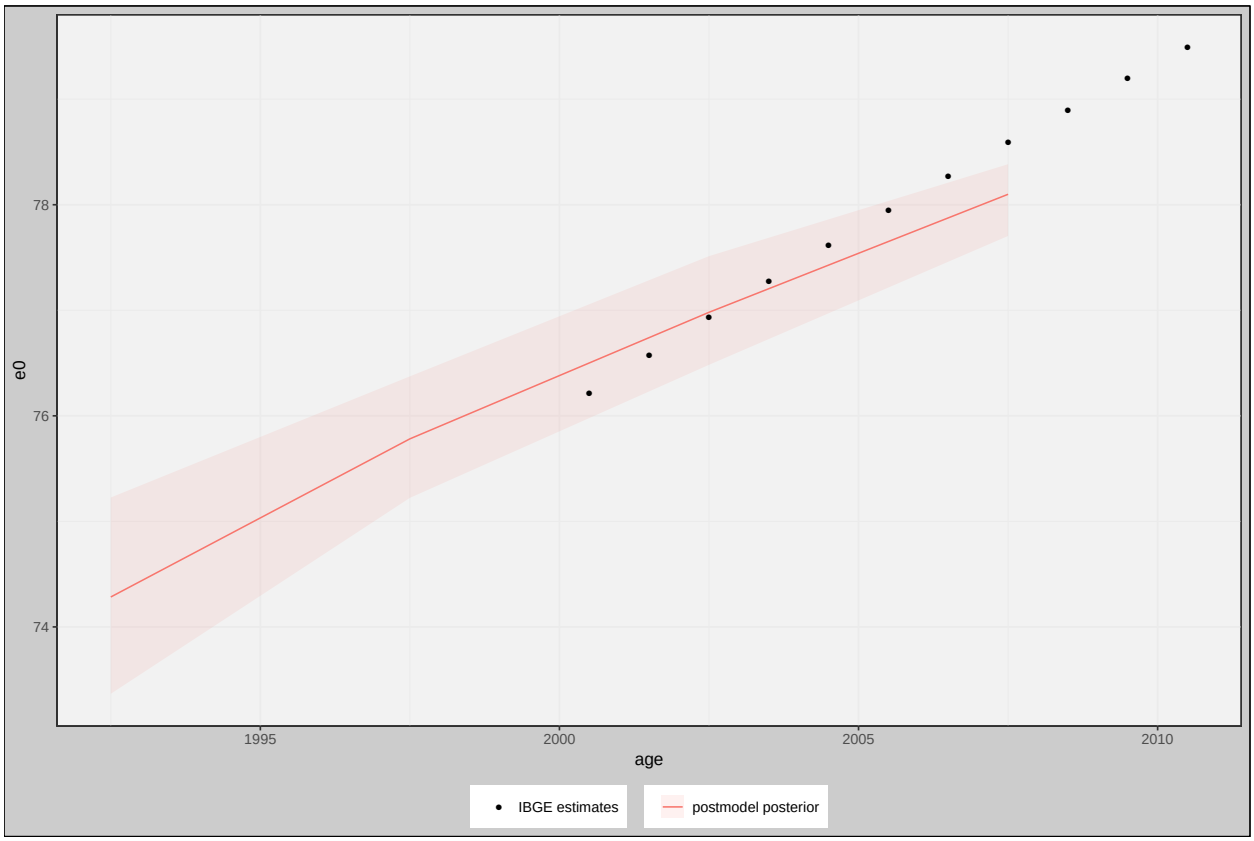

Figure 5.12: Life expectancy at birth: comparison between postmodel posterior with IBGE estimates, RS, female population, 1990, 2000 and 2010

estimates produced by IBGE. IBGE estimates tend to be higher than the median postmodel posterior distribution estimated in this study and this difference is higher for the period close to 2010 than for the years close to 2000.

Figure 5.17 shows the premodel and postmodel estimates of the five-year net migration by age groups. The postmodel estimates are significantly higher than the premodel estimates for the 2000/2005 and 2005/2010 periods, possibly due to the inconsistencies in the measure of in-migration and out-migration discussed in section 5.3.

Figure 5.18 shows the comparison of mortality rates, which also seem to be consistent and with low uncertainty.

Inconsistency between official estimates of life expectancy and the estimates produced by the methods proposed in this dissertation are also found in RJ (Figure 5.19). In this case, official estimates are higher than the postmodel posterior. These inconsistencies have been also reported by Schmertmann and Gonzaga, (2018) and deserve further research to investigate their causes.

Premodel and postmodel fertility estimates are also consistent for RJ, as show by figure

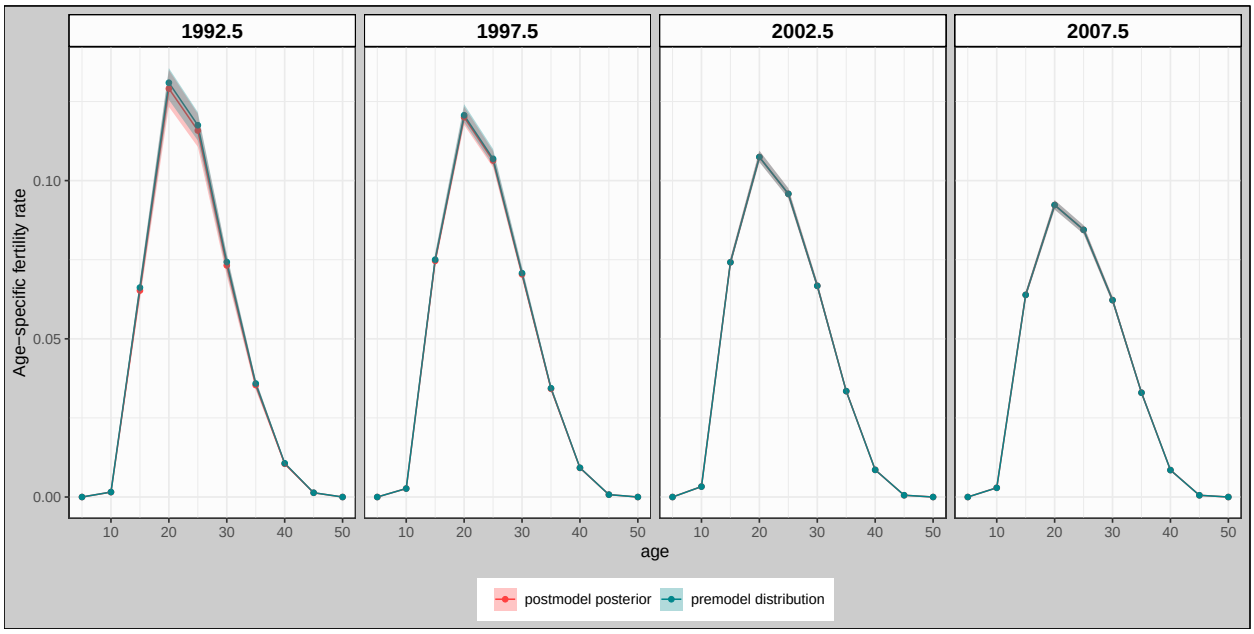

Figure 5.13: Age-specific fertility rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

5.20. Figure 5.21 shows that posterior estimates of the TFR are higher than official estimates for most of the period of comparison.

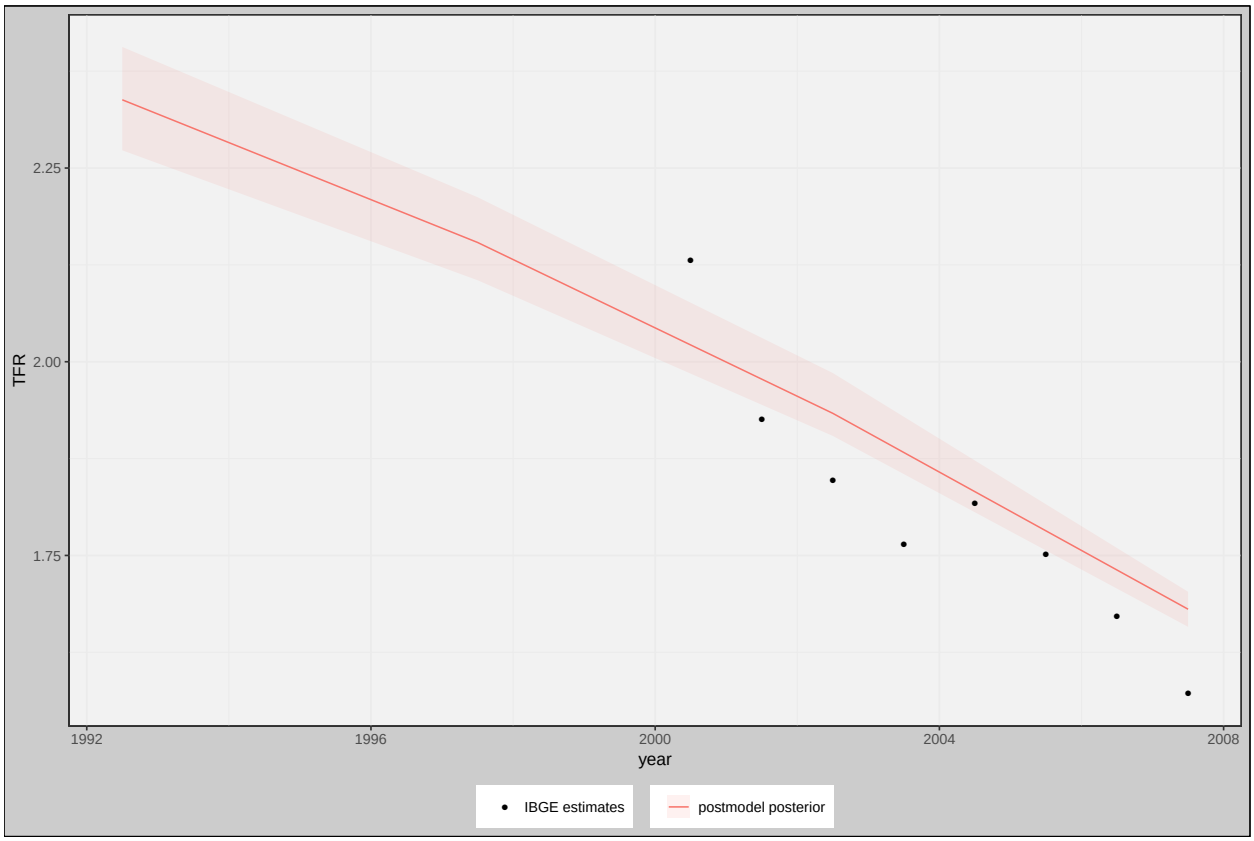

Figure 5.14: TFR: comparison between postmodel posterior with IBGE estimates, Brazil, 1990, 2000 and 2010

Demographic estimates for Paraíba

PB is a poor state in the Northeast region of Brazil, with high infant mortality and deficient vital statistics systems. The PES results show that it had low census undercount in 2000.

The figure for PB (E.4) in the section E.1 of Appendix E shows inconsistencies mostly between the 1990 Census with the other two censuses. The population of children generated from the 1990 projection is much higher than the enumerated population in 2000 and the backprojected population from 2010. This could be due to an overestimation of fertility rates in this state for the 1990s, but it may also be related to an overestimation of women at reproductive ages. This overestimation can be related to imprecision in migration estimates, as discussed in section 5.3.

The results of the application of the proposed method for the population estimates are shown in Figure 5.22. The postmodel posterior estimates are narrower than the premodel posterior distributions. The median undercount for the 1990, 2000 and 2010 censuses were 0.61%, 2.33% and 0.62%. Posterior estimates are close to the censuses counts for all years in almost all age groups. The higher undercount rates for 2000 is mostly driven by the high

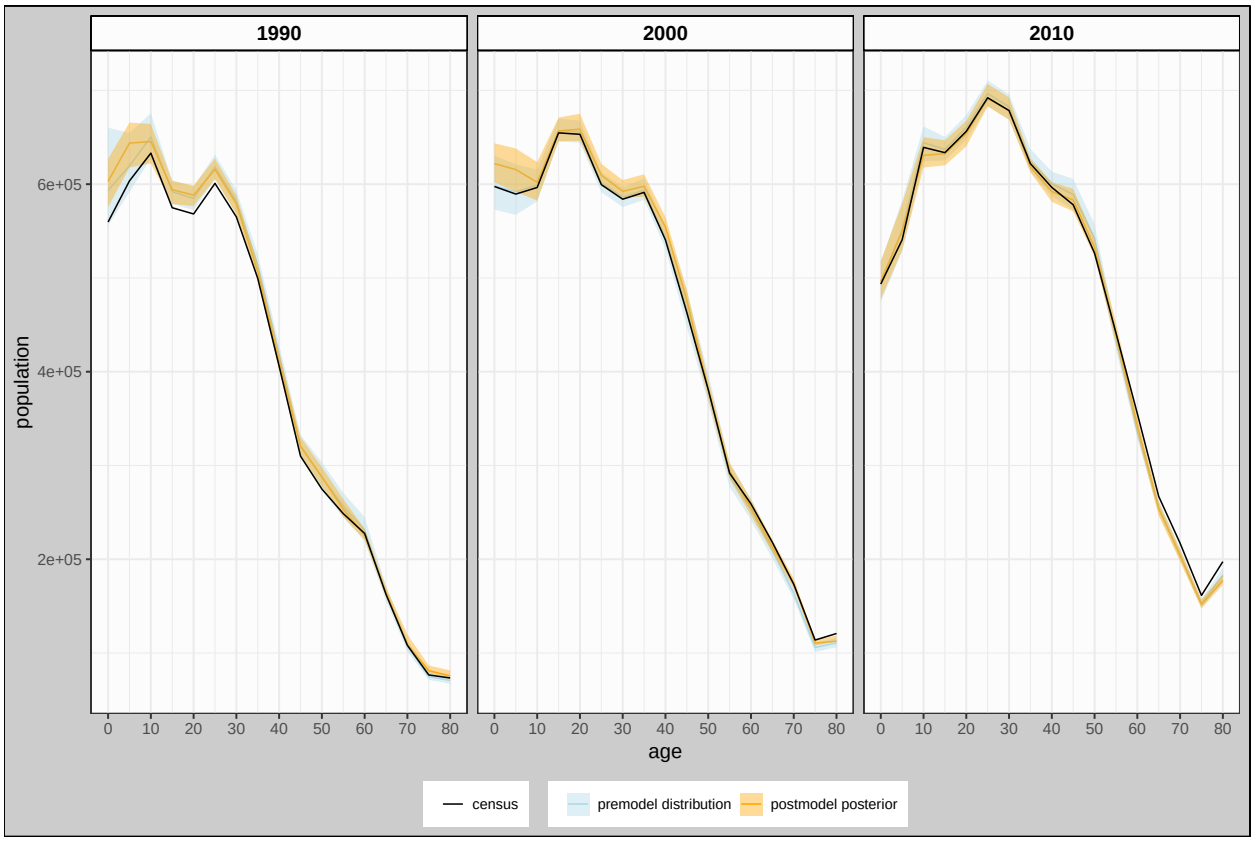

Figure 5.15: Population estimates by age group: comparison between the census, the premodel and the postmodel posterior distributions, RJ, female population, 1990, 2000 and 2010

undercount of children in this census. This can be an artifact of overestimated fertility, that needs to be further investigated.

Figure 5.23 compares estimated populations with official estimates produced by IBGE. IBGE estimates are within the 90th percentile prediction interval for the entire period, but difference is higher in the second half of the 2000s.

Similarly to what occurs in RJ, the premodel and postmodel estimates of the five-year net migration rates are significantly different. In this case, promodel migration estimates seem to underestimate the negative migration flows (Figure 5.24). This may be also a case of inconsistency between reported in-migrants and out-migrants. It seems that in-migration in PB is better captured than out-migration of people from PB to other states.

Mortality estimates in the the premodel and postmodel posterior distributions are also very similar for PB (Figure 5.25).

Comparison with IBGE, however, show high inconsistencies, with much lower life expectancies in IBGE estimates (Figure 5.26).

Premodel and postmodel fertility estimates are also consistent for PB (Figure 5.27), as

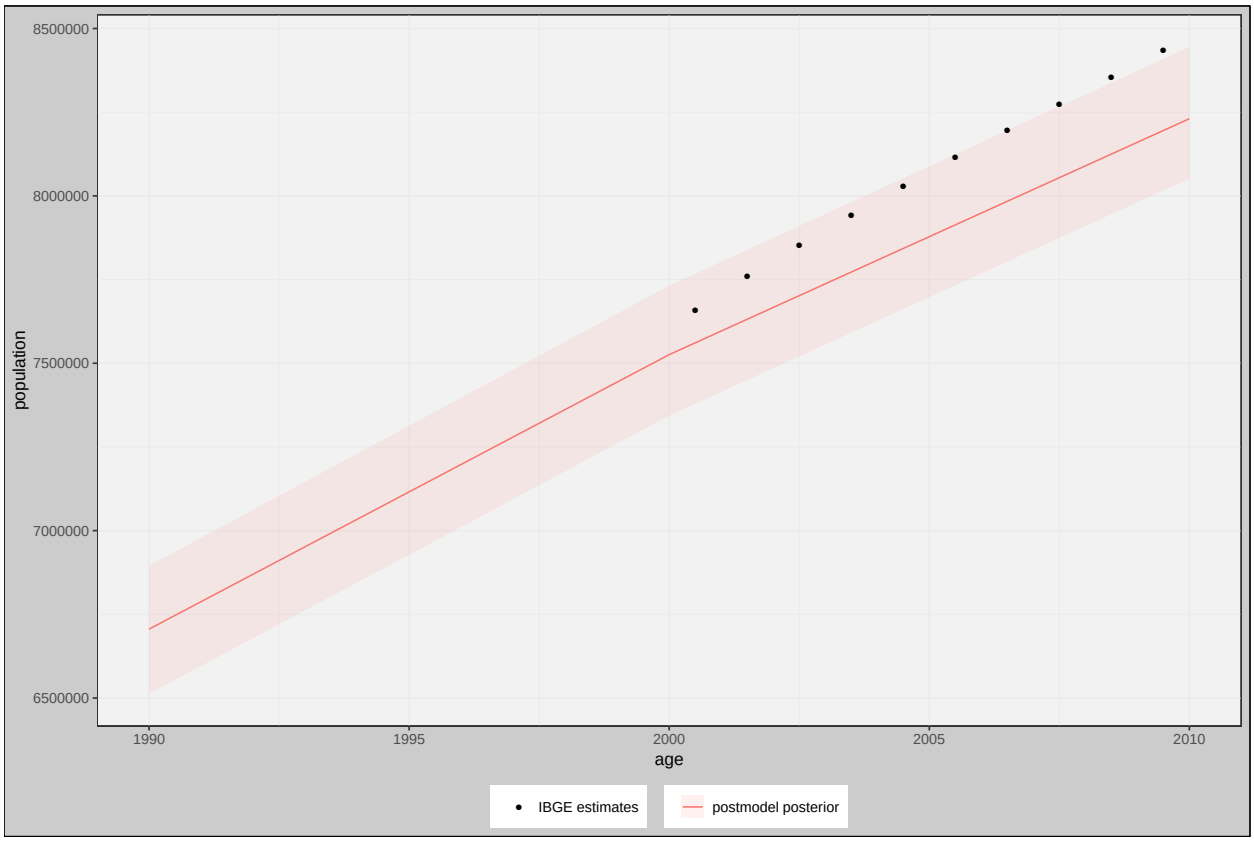

Figure 5.16: Total population: comparison between the postmodel posterior distributions with IBGE estimates, RJ, female population, 1990, 2000 and 2010

well as the comparison with official estimates (Figure 5.28).

5.4 Summary

The illustrations presented in this chapter for Brazil and three selected states show that the methods developed in this dissertation to produce demographic estimates and reconcile inconsistent demographic data seem to work satisfactorily in different contexts, with varying quality of vital statistics, census coverage, and level of migration flows.

The illustrations for different time periods also show that the method is flexible enough to be applicable to contexts with varying data availability.

There are several advantages of working with population estimates in an integrated demographic approach. The reconciliation has changed some previous knowledge about the demographic data. The re-estimation of internal migration flows in the states of PB and RJ is a clear example that indicates biases in the prior information that would be extremely difficult to identify otherwise. Fertility estimates are also improved by using multiple data

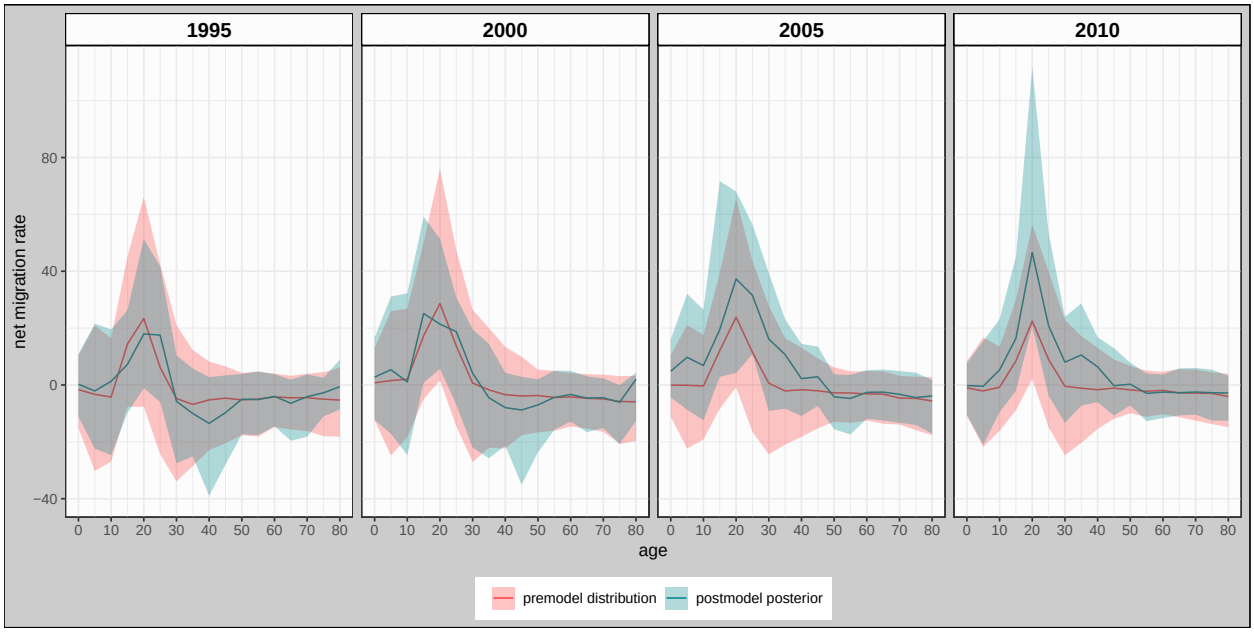

Figure 5.17: Net migration rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, RJ, female population, 1990, 2000 and 2010

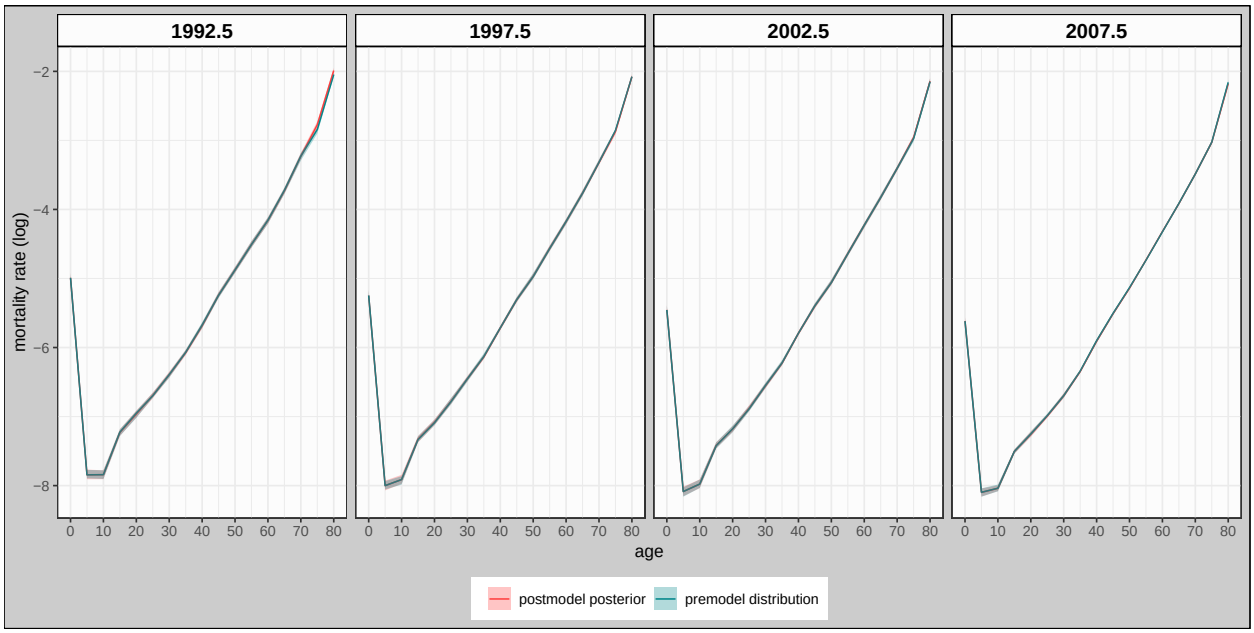

Figure 5.18: Age-specific mortality rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, RJ, female population, 1990, 2000 and 2010

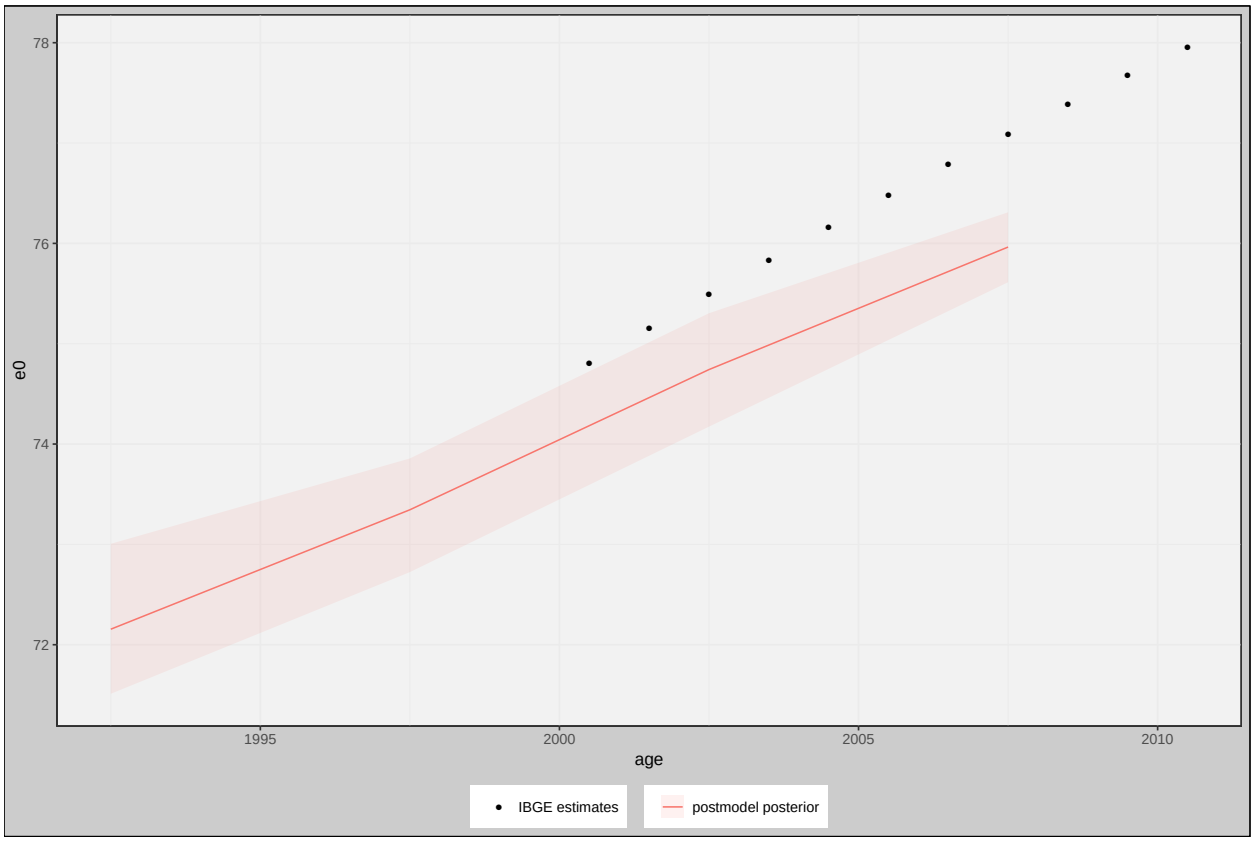

Figure 5.19: Life expectancy at birth: comparison between postmodel posterior with IBGE estimates, RJ, female population, 1990, 2000 and 2010

sources, which include the population of children enumerated in the censuses and their measures of undercount. Prior information on fertility and on the population of children in the subsequent census, improves the estimation of children, in turn.

There are a few limitations in the procedures adopted and avenues for future research. Adjustment factors to take into account age misstatement could use prior information from several other studies that have estimated this indicator. It should also vary by state, possibly proportional to the quality of the reported age in the census.

For international migration, the only information used was the number of emigrants in the 2010 Census, which is known to be biased, for instance due to the missing children. Even though it is difficult to estimate international emigration from Brazil, collecting information of Brazilians in other countries might be helpful to define an age schedule of international emigration.

Future studies should also apply the method to male population.

Finally, the estimates provided in these illustrations were subject to several subjective decisions and are not necessarily the best estimates available. The main objective of this

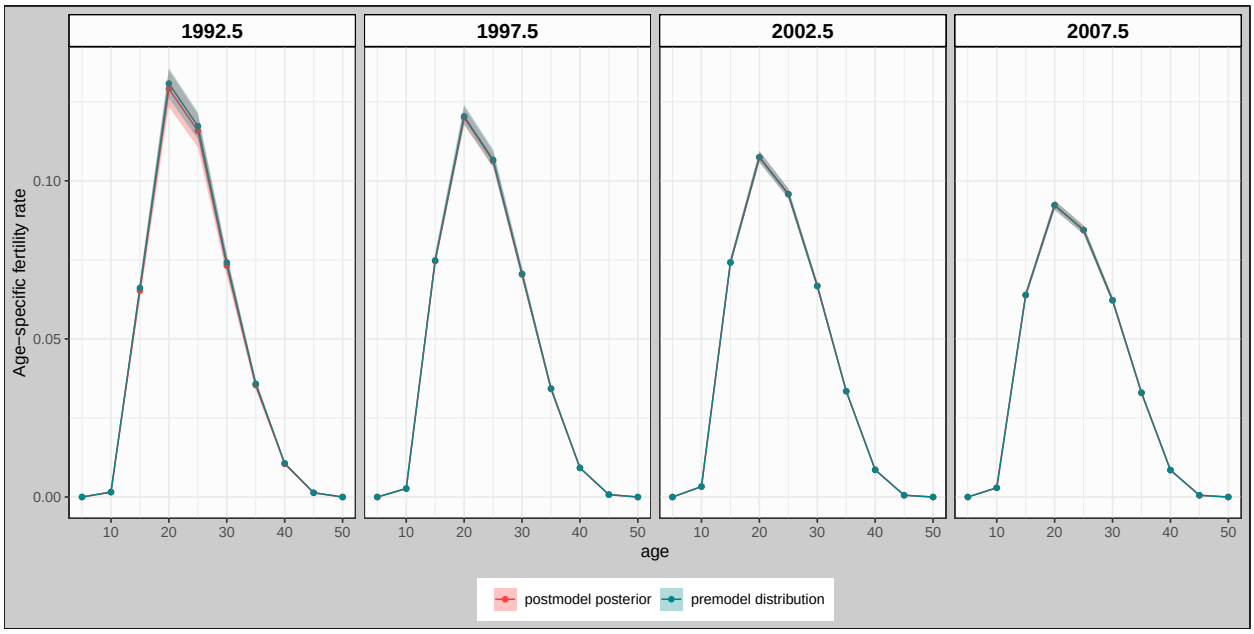

Figure 5.20: Age-specific fertility rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, Brazil, female population, 1990, 2000 and 2010

study was to developed a method that could be used by gathering prior information from experts and then reconciling all the data available to produce consistent estimates, with the associated uncertainty measures.

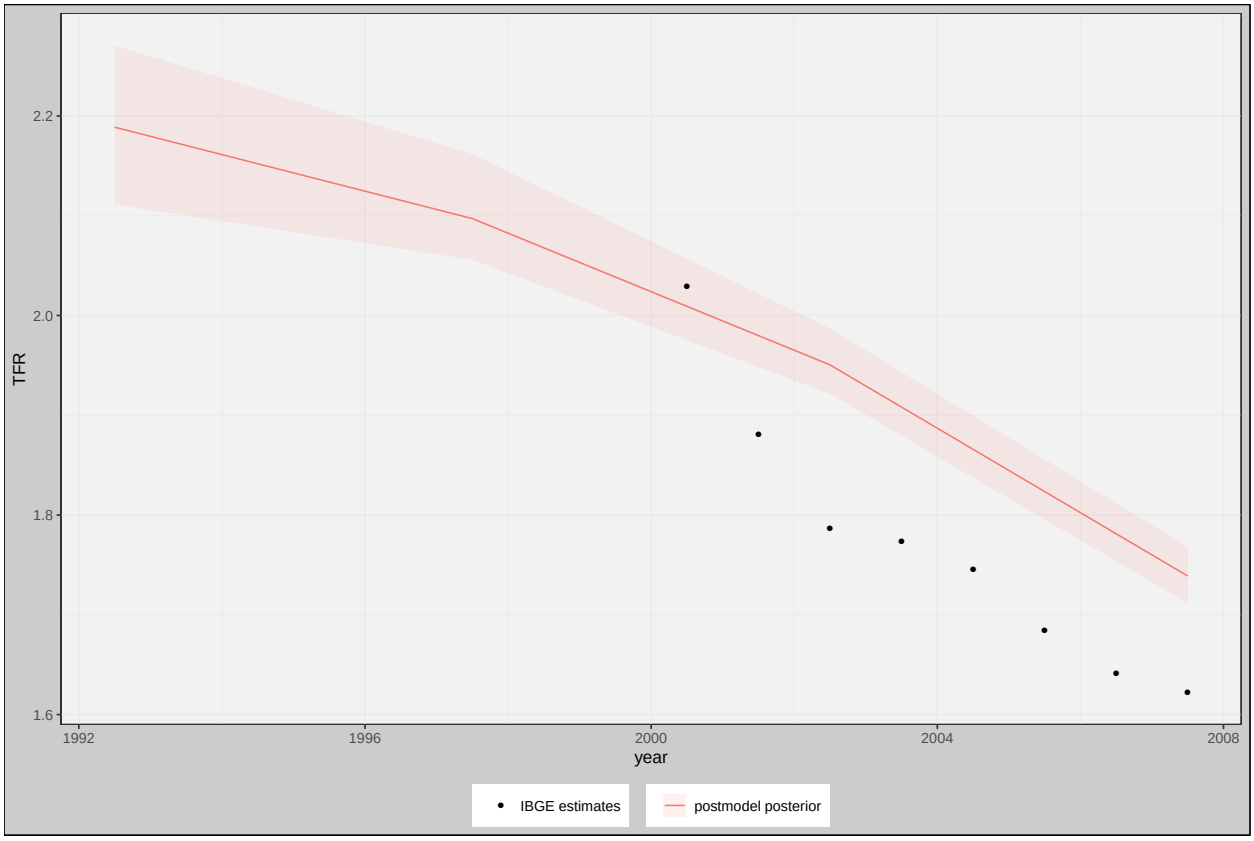

Figure 5.21: TFR: comparison between postmodel posterior with IBGE estimates, RJ, 1990, 2000 and 2010

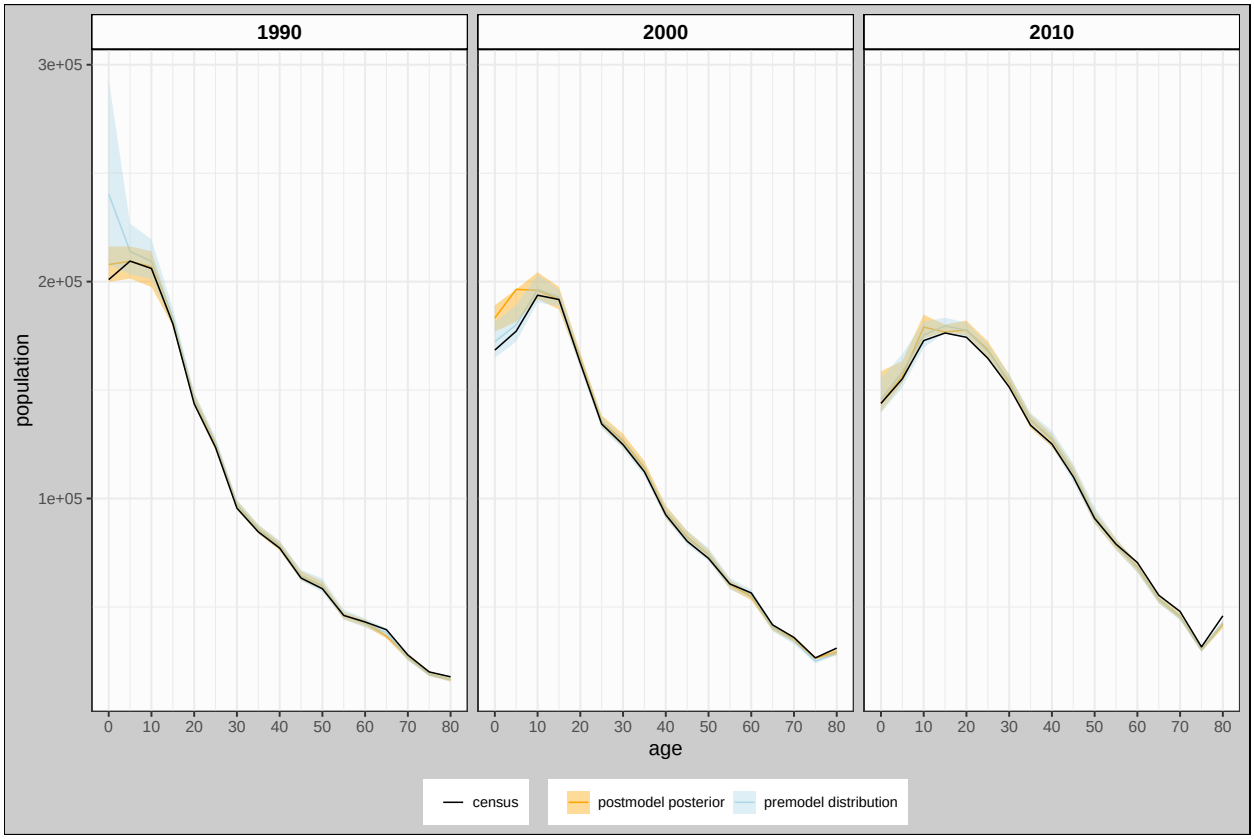

Figure 5.22: Population estimates by age group: comparison between the census, the premodel and the postmodel posterior distributions, PB, female population, 1990, 2000 and 2010

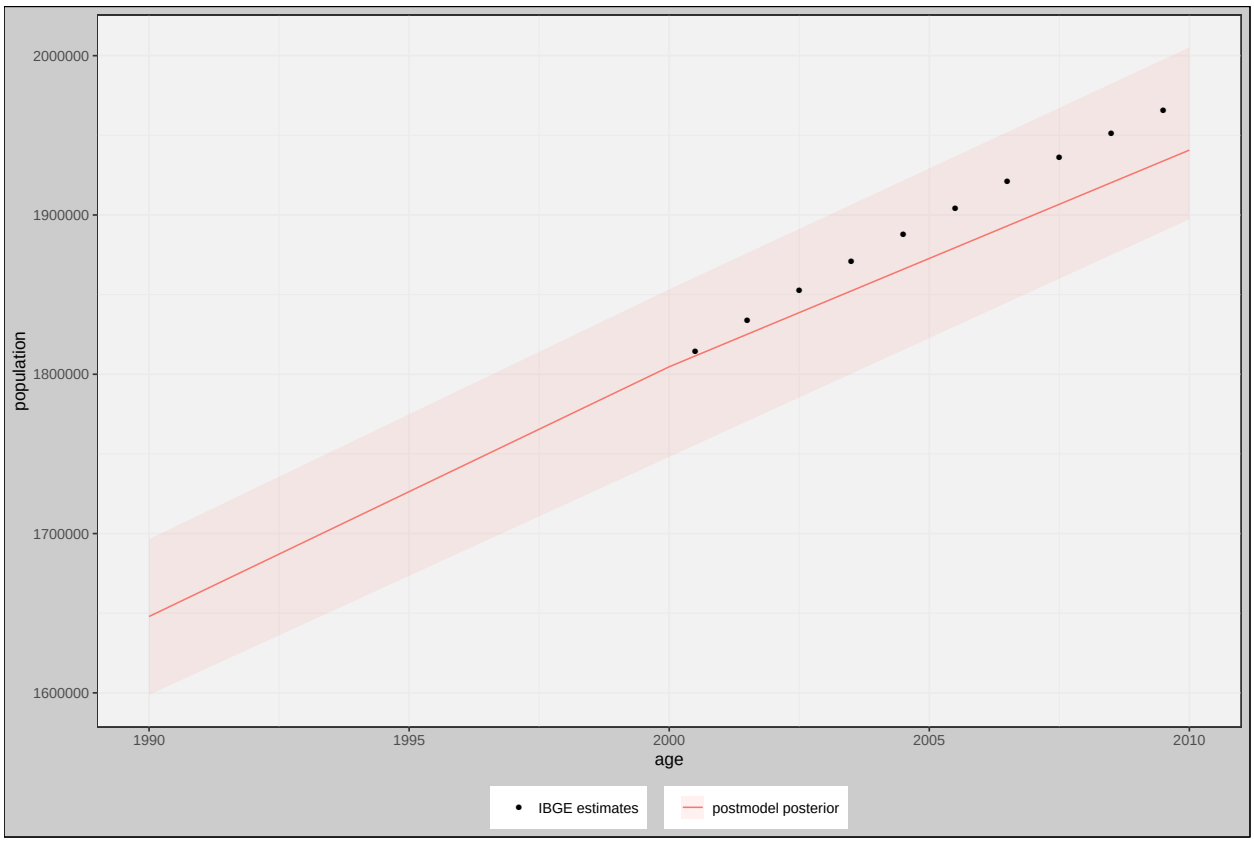

Figure 5.23: Total population: comparison between the postmodel posterior distributions with IBGE estimates, PB, female population, 1990, 2000 and 2010

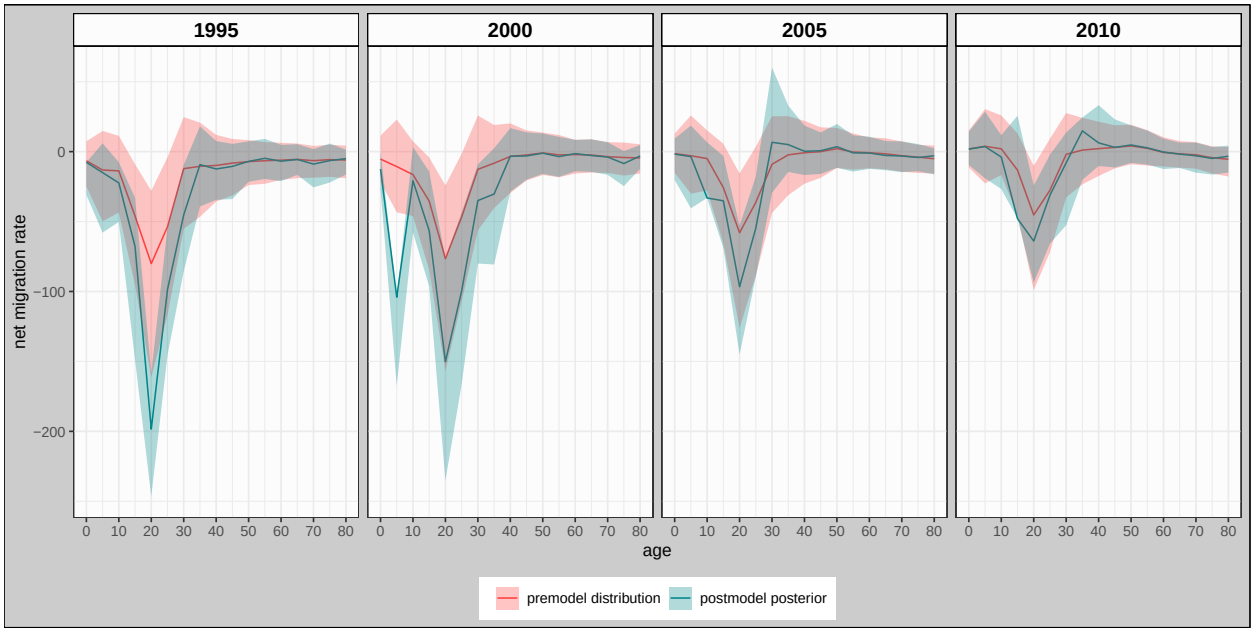

Figure 5.24: Net migration rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, PB, female population, 1990, 2000 and 2010

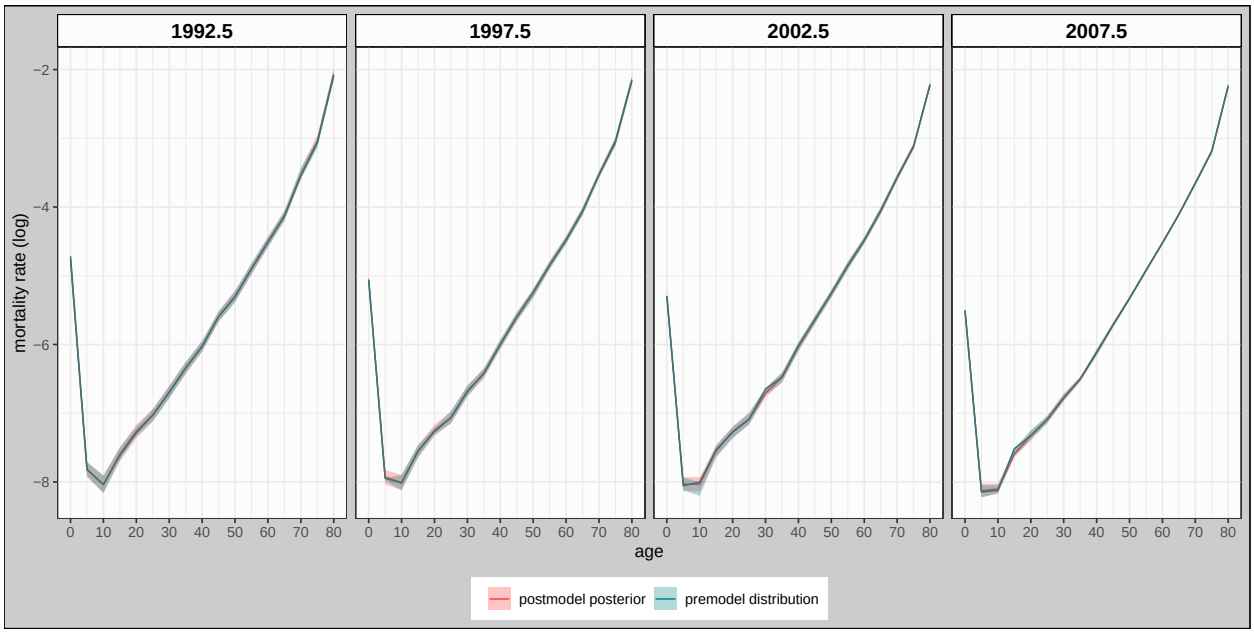

Figure 5.25: Age-specific mortality rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, PB, female population, 1990, 2000 and 2010

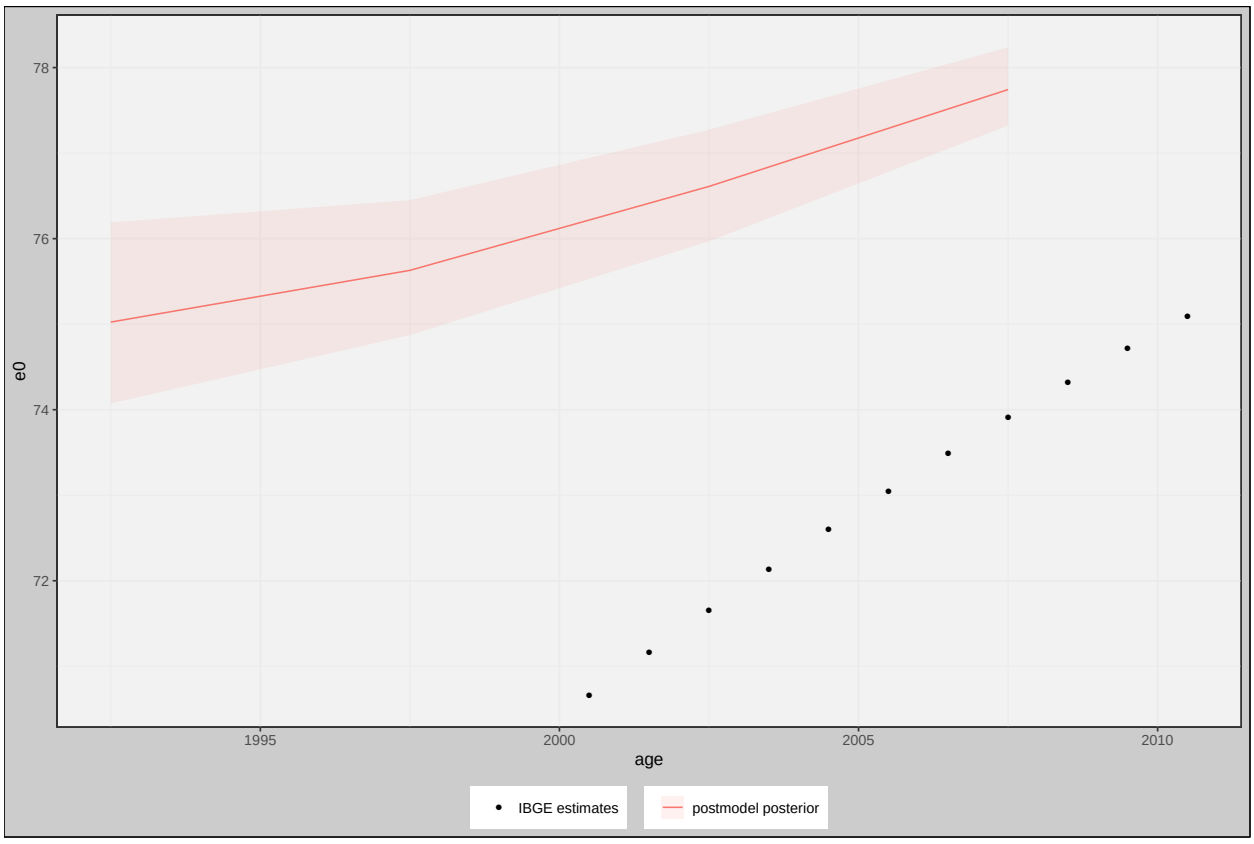

Figure 5.26: Life expectancy at birth: comparison between postmodel posterior with IBGE estimates, PB, female population, 1990, 2000 and 2010

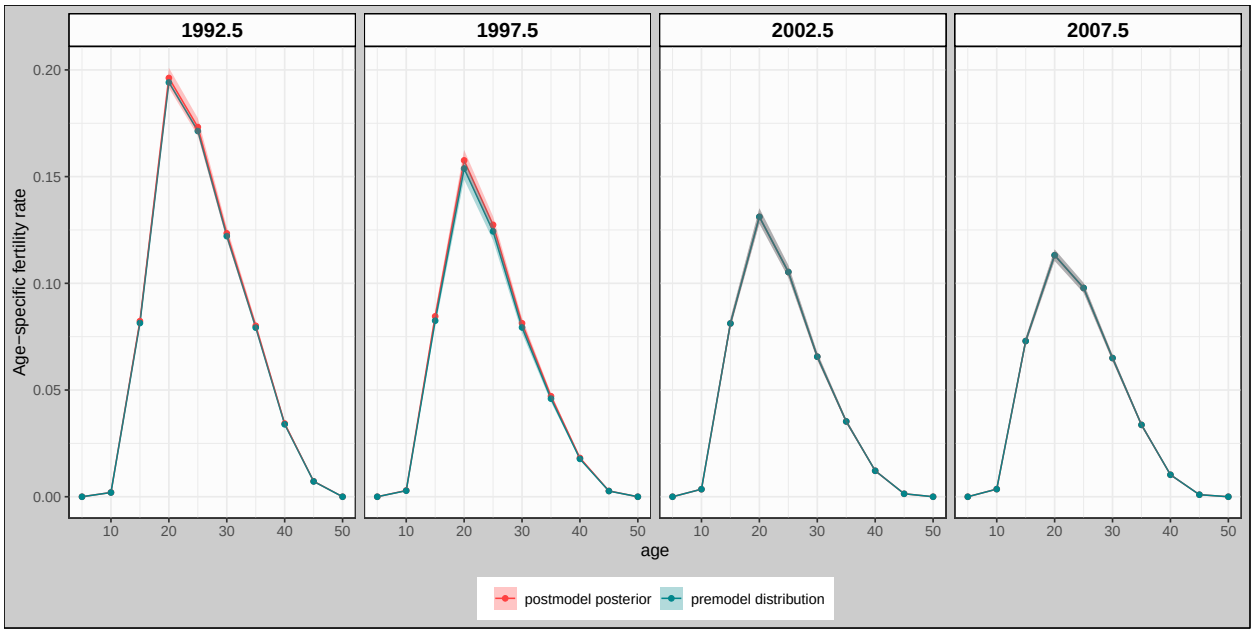

Figure 5.27: Age-specific fertility rates estimates by age group: Comparison between the premodel and postmodel posterior distributions, PB, female population, 1990, 2000 and 2010

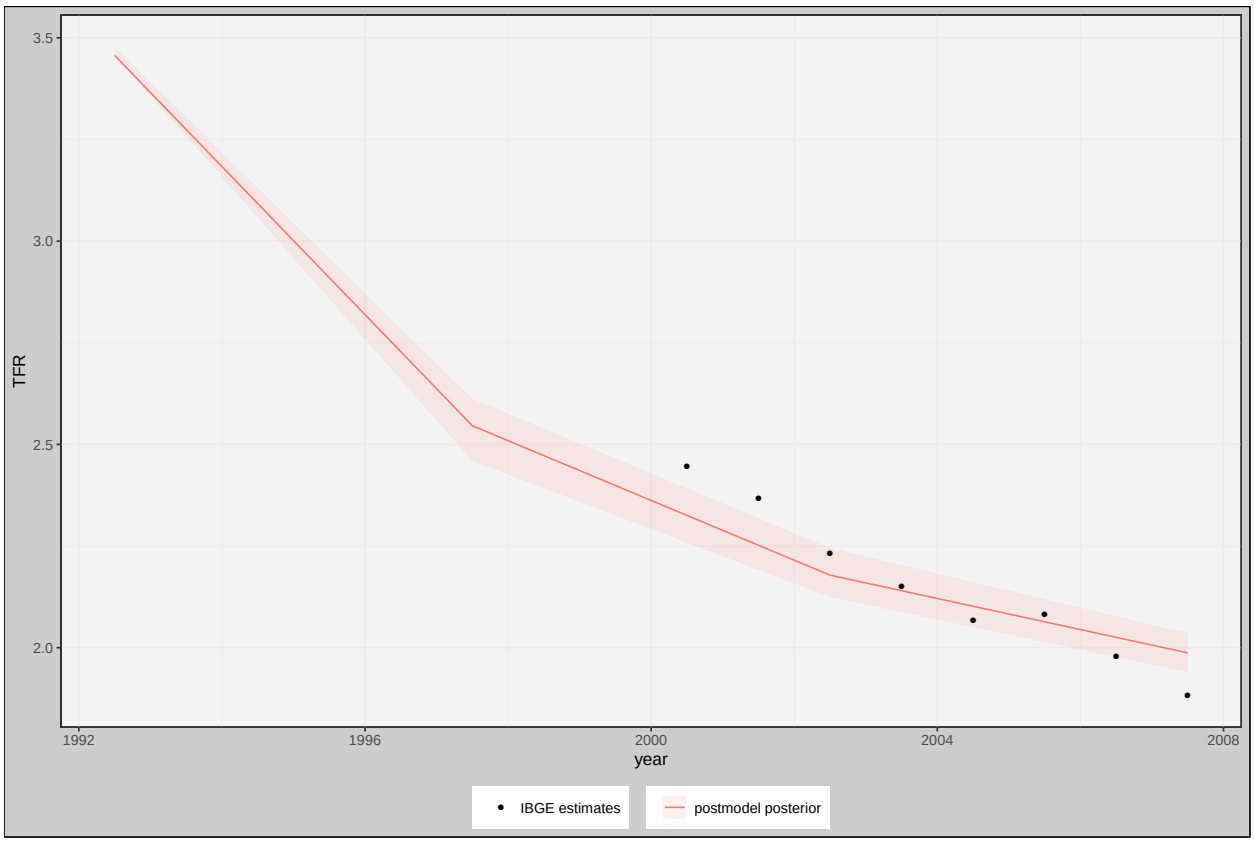

Figure 5.28: TFR: comparison between postmodel posterior with IBGE estimates, Brazil, 1990, 2000 and 2010

Bibliography

Abbott, Owen (2009). "2011 UK Census Coverage Assessment and Adjustment Methodology". en. In: *Population Trends* 137.1, pp. 25–32. ISSN: 2040-1590. DOI: 10.1057/pt.2009.31.

A'Hearn, Brian, Jörg Baten, and Dorothee Crayen (2006). "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital". In: *The Journal of Economic History* 69.3, pp. 783–808.

Alexander, Monica and Leontine Alkema (2018). "Global Estimation of Neonatal Mortality Using a Bayesian Hierarchical Splines Regression Model". In: *Demographic Research* 38.15, pp. 335–372. DOI: 10.4054/DemRes.2018.38.15.

Alexander, Monica, Emilio Zagheni, and Magali Barbieri (2017). "A Flexible Bayesian Model for Estimating Subnational Mortality". en. In: *Demography* 54.6, pp. 2025–2041. ISSN: 0070-3370, 1533-7790. DOI: 10.1007/s13524-017-0618-7.

Alkema, Leontine, Doris Chou, et al. (2016). "National, Regional, and Global Levels and Trends in Maternal Mortality between 1990 and 2015 with Scenario-Based Projections to 2030: A Systematic Analysis by the United Nations Maternal Mortality Estimation Inter-Agency Group". In: *Lancet (London, England)* 387.10017, pp. 462–474. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(15)00838-7.

Alkema, Leontine, Jin Rou New, et al. (2014). "Child Mortality Estimation 2013: An Overview of Updates in Estimation Methods by the United Nations Inter-Agency Group for Child Mortality Estimation". In: *PLoS ONE* 9.7. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0101112.

Alkema, Leontine, Adrian Raftery, and Samuel J. Clark (2007). "Probabilistic Projections of HIV Prevalence Using Bayesian Melding". In: *The Annals of Applied Statistics* 1.1, pp. 229–248. ISSN: 1932-6157.

Alkema, Leontine, Adrian Raftery, Patrick Gerland, et al. (2012). "Estimating Trends in the Total Fertility Rate with Uncertainty Using Imperfect Data: Examples from West Africa". In: *Demographic research* 26.15. ISSN: 1435-9871. DOI: 10.4054/DemRes.2012.26.15.

Anderson, Charles, Thomas Bratcher, and Khalil Kutran (1994). "Bayesian Estimation of Population Density and Visibility". In: *Texas Journal of Science* 46.1, pp. 1–12.

Bangha, Martin W. (2010). "Estimating Adult Mortality in Cameroon from Census Data on Household Deaths: 1976-1987". In:

Banister, Judith and Kenneth Hill (2004). "Mortality in China 1964-2000". eng. In: *Population Studies* 58.1, pp. 55–75. ISSN: 0032-4728. DOI: 10.1080/0032472032000183753.

Bennett, Neil G. and Shiro Horiuchi (1981). "Estimating the Completeness of Death Registration in a Closed Population". In: *Population index*, pp. 207–221.

Berquó, Elza S. and Suzana M. Cavenaghi (2014). "Notas sobre os diferenciais educacionais e econômicos da fecundidade no Brasil". pt. In: *Revista Brasileira de Estudos de População* 31.2, pp. 471–482. ISSN: 1980-5519.

Blumenthal, Saul and Ram C. Dahiya (1981). "Estimating the Binomial Parameter N". In: *Journal of the American Statistical Association* 76.376, pp. 903–909. ISSN: 0162-1459. DOI: 10.1080/01621459.1981.10477739.

Booth, H and Patrick Gerland (2016). *Demographic Techniques: Data Adjustment and Correction*. en-AU. http://demography.cass.anu.edu.au/research/publications/demographic-techniques-data-adjustment-and-correction.

Borges, G. M. and L. Silva (2015). "Fontes de Dados de Fecundidade No Brasil: Características, Vantagens e Limitações". In: *Mudança Demográfica No Brasil No Início Do Século XXI: Subsídios Para as Projeções Da População*. Rio de Janeiro: IBGE, pp. 138–151.

Brass, William (1964). "Uses of Census or Survey Data for the Estimation of Vital Rates". In: *UN. ECA African Seminar on Vital Statistics*. Addis Ababa, Ethiopia: United Nations. Economic and Social Counci.

— (1971). "Mortality Estimates from Children Ever Born and Children Surviving: In Methods for Estimating Fertility and Mortality from Limited and Defective Data". In: *The University of North Carolina at Chapel Hill*, pp. 50–59.

— (1975). "Methods for Estimating Fertility and Mortality from Limited and Defective Data." In: *Methods for estimating fertility and mortality from limited and defective data.*

Brass, William and Ansley Coale (1968). "Methods of Analysis and Estimation". In: *Demography of Tropical Africa*. Princeton University Press, pp. 88–150.

Brass, William and Kenneth Hill (1973). "Estimating Adult Mortality from Orphanhood." In: *Population Studies*. Vol. 3. Liege, Belgium: IUSSP, pp. 111–23.

Brillinger, David R. (1986). "A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics". In: *Biometrics* 42.4, pp. 693–734. ISSN: 0006-341X. DOI: 10.2307/2530689.

Brito, Fausto (2000). "Brasil, final do século: a transição para um novo padrão migratório?" pt. In: *ABEP*, pp. 1–44.

Bryan, Thomas and Robert Heuser (2004). "Collection and Processing of Demographic Data". In: *The Methods and Materials of Demography*. Second. San Diego, California: Elsevier Academic Press.

Bryant, John and Patrick Graham (2013). "Bayesian Demographic Accounts: Subnational Population Estimation Using Multiple Data Sources". EN. In: *Bayesian Analysis* 8.3, pp. 591–622. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/13-BA820.

Bryant, John and Patrick Graham (2015). "A Bayesian Approach to Population Estimation with Administrative Data". In: *Journal of Official Statistics* 31.3, pp. 475–487. ISSN: 2001-7367. DOI: 10.1515/jos-2015-0028.

Bulatao, Rodolfo A and John Bongaarts (2000). *Beyond Six Billion:: Forecasting the World's Population*. National Academies Press. ISBN: 0-309-06990-4.

Camarda, Carlo G. (2012). "MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines". en-US. In: *Journal of Statistical Software* 50.1. DOI: 10.18637/jss.v050.i01.

Campos, Marden Barbosa de (2011). "Reversão do saldo migratório internacional negativo do Brasil? Evidências preliminares com base nos dados do censo 2010". por. In: *Revista Paranaense de Desenvolvimento* 121, pp. 189–200. ISSN: 2236-5567.

— (2014). "Medidas de emigración internacional basadas en la información proporcionada por personas que convivieron con los emigrantes: la experiencia brasileña con el Censo Demográfico de 2010". es. In:

— (2018). "Vale a Pena Manter as Perguntas Sobre Ex-Moradores Residindo No Exterior Nos Censos Demográficos Brasileiros?" In: *Revista Brasileira de Estudos de População* 35.3. ISSN: 0102-3098. DOI: 10.20947/s0102-3098a0044.

Carpenter, Bob et al. (2017). "Stan: A Probabilistic Programming Language". en-US. In: *Journal of Statistical Software* 76.1. DOI: 10.18637/jss.v076.i01.

Carroll, Raymond J. and F. Lombard (1985). "A Note on N Estimators for the Binomial Distribution". In: *Journal of the American Statistical Association* 80.390, pp. 423–426. ISSN: 0162-1459. DOI: 10.1080/01621459.1985.10478134.

Carvalho, José Alberto Magno de (1982). "Aplicabilidade Da Técnica de Fecundidade de Brass Quando a Fecundidade Está Declinando Ou Quando a População Não é Fechada". In: *III Encontro Nacional de Estudos Populacionais*.

— (1996). "O Saldo Dos Fluxos Migratórios Internacionais Do Brasil Na Década de 80 - Uma Tentativa de Estimação". en. In: *Revista Brasileira de Estudos de População* 13.1, pp. 3–14. ISSN: 1980-5519.

Carvalho, José Alberto Magno de and Marden Barbosa de Campos (2006). "A Variação Do Saldo Migratório Internacional Do Brasil". In: *Estudos Avançados* 20.57, pp. 55–58. ISSN: 0103-4014. DOI: 10.1590/S0103-40142006000200005.

Carvalho, José Alberto Magno de, G. Q. Gonçalves, and L. Silva (2018). "Application of P/F Brass Ratio Method in the Context of Fast-Paced Adolescent Fertility Decline". en. In: *Revista Brasileira de Estudos de População* 35.1, pp. 1–26. ISSN: 1980-5519. DOI: http://dx.doi.org/10.20947/S102-3098a0052.

Castanheira, Helena and Hans-Peter Kohler (2015). "It Is Lower Than You Think It Is: Recent Total Fertility Rates in Brazil and Possibly Other Latin American Countries". In: *PSC Working Paper Series, WPS 15-5*.

Cavenaghi, Suzana M. and José Eustáquio Diniz Alves (2016). "Qualidade Das Informações Sobre Fecundidade No Censo Demográfico de 2010". In: *Revista Brasileira de Estudos de População* 33.1, pp. 189–205. ISSN: 0102-3098. DOI: 10.20947/S0102-309820160010.

CELADE (1968). "Métodos de Evaluación En Los Censos de Población: Algunas Aplicaciones Hechas Por CELADE". In: *A/83, Santiago*.

Chackiel, Juan (2009). *Evaluación y Estimación de La Cobertura En Los Censos de Población: La Experiencia Latinoamericana*.

Chatterjee, Kiranmoy and Diganta Mukherjee (2015). "An Improved Estimator of Omission Rate for Census Count: With Particular Reference to India". In: *Communications in Statistics - Theory and Methods* 45, pp. 1047–1062. DOI: 10.1080/03610926.2013.854911.

Cho, Lee Jay, Robert D. Retherford, and MK Choe (1986). *The Own-Children Method of Fertility Estimation*. Honolulu, Hawaii: The East-West Center.

Clark, Samuel J. and David J. Sharrow (2011). *Contemporary Model Life Tables for Developed Countries An Application of Model-Based Clustering*. Working Paper 107. Seattle: University of Washington.

Clark, Samuel J., Jason R. Thomas, and Le Bao (2012). "Estimates of Age-Specific Reductions in HIV Prevalence in Uganda: Bayesian Melding Estimation and Probabilistic Population Forecast with an HIV-Enabled Cohort Component Projection Model". In: *Demographic research* 27. ISSN: 1435-9871. DOI: 10.4054/DemRes.2012.27.26.

Coale, Ansley and Graziella Caselli (1990). "Estimation of the Number of Persons at Advanced Ages from the Number of Deaths at Each Age in the given Year and Adjacent Years". In: *Genus*, pp. 1–23.

Coale, Ansley and P Demeny (1966). *Regional Model Life Tables and Stable Populations*. New Jersey: Princeton Univ. Press.

CONAPO (2012). *Proyecciones de La Población de México, 2010-2050. Documento Metodológico*.

Costa, M et al. (2001). "Mortalidade infantil e condições de vida: a reprodução das desigualdades sociais em saúde na década de 90". pt. In: *Cadernos de Saúde Pública* 17, pp. 555–567. ISSN: 0102-311X, 0102-311X, 1678-4464. DOI: 10.1590/S0102-311X2001000300011.

Das Gupta, Prithwis (1975). "A General Method of Correction for Age Misreporting in Census Populations". en. In: *Demography* 12.2, pp. 303–312. ISSN: 0070-3370, 1533-7790. DOI: 10.2307/2060767.

DasGupta, A. and Herman Rubin (2005). "Estimation of Binomial Parameters When Both n, p Are Unknown". In: *Journal of Statistical Planning and Inference*. Herman Chernoff: Eightieth Birthday Felicitation Volume 130.1, pp. 391–404. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2004.02.019.

de Beer, Joop (2012). "Smoothing and Projecting Age-Specific Probabilities of Death by TOPALS". In: *Demographic Research* 27.20, pp. 543–592. DOI: 10.4054/DemRes.2012.27.20.

de Carvalho, Luiz Max et al. (2015). "Choosing the Weights for the Logarithmic Pooling of Probability Distributions". In: *arXiv:1502.04206 [stat]*. arXiv: 1502.04206 [stat].

Del Popolo, Fabiana (2000). *Los Problemas En La Declaración de La Edad de La Población Adulta Mayor En Los Censos*. CEPAL. ISBN: 92-1-321668-8.

Demeny, P and F. C. Shorter (1968). "Estimating Turkish Mortality, Fertility, and Age Structure: Application of Some New Techniques". In: *Faculty of Economics Pub. No. 218, University of Istanbul*.

Dorrington, Rob (2013). "Synthetic Extinct Generations Methods". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Dorrington, Rob, I. M. Timaeus, and Simon Gregson (2006). "Adult Mortality in Southern Africa Using Deaths Reported by Households: Some Methodological Issues and Results". In: *Annual Conference of the Population Association of America. Los Angeles*. Vol. 30.

Dorrington, Rob and Ian M. Timæus (2008). "Death Distribution Methods for Estimating Adult Mortality: Sensitivity Analysis with Simulated Data Errors, Revisited". In: *Population Association of America 2008*. New Orleans.

Draper, N. and I. Guttman (1971). "Bayesian Estimation of the Binomial Parameter". In: *Technometrics* 13.3, pp. 667–673. ISSN: 0040-1706. DOI: 10.1080/00401706.1971.10488827.

Dvorzak, Michaela and Helga Wagner (2016). "Sparse Bayesian Modelling of Underreported Count Data". en. In: *Statistical Modelling* 16.1, pp. 24–46. ISSN: 1471-082X. DOI: 10.1177/1471082X15588398.

Ewbank, Douglas C. (1981). "Age Misreporting and Age-Selective Underenumeration: Sources Patterns and Consequences for Demographic Analysis." In:

Feehan, Dennis M., Mary Mahy, and Matthew J. Salganik (2017). "The Network Survival Method for Estimating Adult Mortality: Evidence From a Survey Experiment in Rwanda". en. In: *Demography* 54.4, pp. 1503–1528. ISSN: 0070-3370, 1533-7790. DOI: 10.1007/s13524-017-0594-y.

Freedman, David A and Kenneth W Wachter (2001). "Census Adjustment: Statistical Promise or Illusion?" en. In: *Society* 39.1, p. 26. ISSN: 1936-4725. DOI: 10.1007/BF02712617.

— (2003). "On the Likelihood of Improving the Accuracy of the Census through Statistical Adjustment". In: *Lecture Notes-Monograph Series* 40, pp. 197–230. ISSN: 0749-2170.

Gakidou, E. and G. King (2006). "Death by Survey: Estimating Adult Mortality without Selection Bias from Sibling Survival Data". In: *Demography* 43.3, pp. 569–585.

Garcia, Andres J. et al. (2015). "Modeling Internal Migration Flows in Sub-Saharan Africa Using Census Microdata". In: *Migration Studies* 3.1, pp. 89–110. ISSN: 2049-5838. DOI: 10.1093/migration/mnu036.

Garcia, R. A. (2013). "Estimates of International Migrants Leaving Brazil between 1995 and 2000: An Application of the Intercensal Survival Method". In: *Revista Brasileira de Estudos de População* 30.1, pp. 99–123. ISSN: 0102-3098. DOI: 10.1590/S0102-30982013000100006.

Gelman, Andrew et al. (2013). *Bayesian Data Analysis, Third Edition*. en. CRC Press. ISBN: 978-1-4398-4095-5.

Genest, Christian, Kevin J. McConway, and Mark J. Schervish (1986). "Characterization of Externally Bayesian Pooling Operators". In: *The Annals of Statistics* 14.2, pp. 487–501. ISSN: 0090-5364.

Genest, Christian and James V. Zidek (1986). "Combining Probability Distributions: A Critique and an Annotated Bibliography". In: *Statistical Science* 1.1, pp. 114–135. ISSN: 0883-4237.

Gerland, Patrick (2014). "UN Population Division's Methodology in Preparing Base Population for Projections: Case Study for India". In: *Asian Population Studies* 10.3, pp. 274–303. ISSN: 1744-1730. DOI: 10.1080/17441730.2014.947059.

Givens, Geof H. and Paul J. Roback (1999). "Logarithmic Pooling of Priors Linked by a Deterministic Simulation Model". In: *Journal of Computational and Graphical Statistics* 8.3, pp. 452–478. ISSN: 1061-8600. DOI: 10.1080/10618600.1999.10474826.

Glei, Dana, Hans Lundström, and John Wilmoth (2017). *About Mortality Data for Sweden*. Tech. rep. Berkeley: Human Mortality Database.

Gompertz, Benjamin (1825). "On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies". In: *Philosophical Transactions of the Royal Society of London* 115, pp. 513–583. ISSN: 0261-0523.

Gonzaga, Marcos R. and Carl P. Schmertmann (2016). "Estimating Age- and Sex-Specific Mortality Rates for Small Areas with TOPALS Regression: An Application to Brazil in 2010". In: *Revista Brasileira de Estudos de População* 33.3, pp. 629–652. ISSN: 0102-3098. DOI: 10.20947/s0102-30982016c0009.

Goodkind, Daniel (2011). "Child Underreporting, Fertility, and Sex Ratio Imbalance in China". en. In: *Demography* 48.1, pp. 291–316. ISSN: 0070-3370, 1533-7790. DOI: 10.1007/s13524-010-0007-y.

Grabill, Wilson R. and Lee Jay Cho (1965). "Methodology for the Measurement of Current Fertility From Population Data on Young Children". en. In: *Demography* 2.1, pp. 50–73. ISSN: 0070-3370, 1533-7790. DOI: 10.2307/2060106.

Guzmán, José Miguel et al. (2006). "The Demography of Latin America and the Caribbean since 1950". In: *Population* 61.5, pp. 519–620.

Haupt, Arthur, Thomas Kane, and Carl Haub (2011). *PRB's Population Handbook*.

Heligman, L. and J. H. Pollard (1980). "The Age Pattern of Mortality". en. In: *Journal of the Institute of Actuaries* 107.1, pp. 49–80. ISSN: 2058-1009, 0020-2681. DOI: 10.1017/S0020268100040257.

Hill, Kenneth (1987). "Estimating Census and Death Registration Completeness". In: *Asian and Pacific Population Forum/East-West Population Institute, East-West Center*. Vol. 1, p. 8. ISBN: 0891-2823.

— (2013). "Indirect Estimation of Child Mortality". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

— (2017). *Analytical Methods to Evaluate the Completeness and Quality of Death Registration: Current State of Knowledge*. Tech. rep. Technical Paper No. 2017/2. New York: United Nations Department of Economic and Social Affairs Population Division.

Hill, Kenneth, Yoonjoung Choi, and Ian M. Timæus (2005). "Unconventional Approaches to Mortality Estimation". In: *Demographic Research* 13.12, pp. 281–300.

Hill, Kenneth and Rob Dorrington (2013). "Introduction to Migration". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Hill, Kenneth and J. Trussell (1977). "Further Developments in Indirect Mortality Estimation". In: *Population Studies* 31.2, pp. 313–334. ISSN: 0032-4728. DOI: 10.2307/2173920.

Hill, Kenneth, Danzhen You, and Yoonjoung Choi (2009). "Death Distribution Methods for Estimating Adult Mortality: Sensitivity Analysis with Simulated Data Errors". In: *Demographic Research* 21.9, pp. 235–254. DOI: 10.4054/DemRes.2009.21.9.

Hill, Kenneth, Danzhen You, Mie Inoue, et al. (2012). "Child Mortality Estimation: Accelerated Progress in Reducing Global Child Mortality, 1990–2010". en. In: *PLOS Medicine* 9.8, e1001303. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001303.

Hobbs, Frank (2004). "Age and Sex Composition". In: *The Methods and Materials of Demography*. Second. San Diego, California: Elsevier Academic Press.

Hook, E. B. and R. R. Regal (1995). "Capture-Recapture Methods in Epidemiology: Methods and Limitations". eng. In: *Epidemiologic Reviews* 17.2, pp. 243–264. ISSN: 0193-936X.

IBGE (2008). "Projeção Da População Por Sexo e Idade: 1980-2050, Revisão 2008". In: *Estudos e Pesquisa – Informação Demográfica e Socioeconômica* 24.

— (2012). *Censo Demográfico 2010: Características Gerais Da População, Religião e Pessoas Com Deficiência*. Rio de Janeiro: IBGE.

— (2013a). *Metodologia Do Censo Demográfico 2010*. Tech. rep. 41. Rio de Janeiro: IBGE, p. 712.

— (2013b). *Projeção Da População Do Brasil e Unidades Da Federação*. Vol. 40. Séries Relatórios Metodológicos. Rio de Janeiro: IBGE.

— (2013c). *Tábuas Abreviadas de Mortalidade Por Sexo e Idade: Brasil, Grandes Regiões e Unidades Da Federação: 2010*. Estudos e pesquisas. Informação demográfica e socioeconômica 30. Rio de Janeiro: IBGE. ISBN: 978852404296.

— (2018). *Projeções Da População: Brasil e Unidades Da Federação (Revisão 2018)*. Second. Vol. 40. Séries Relatórios Metodológicos. Rio de Janeiro: IBGE.

INDEC (2013). *Estimaciones y Proyecciones de Población 2010-2040 - Total Del País*. Tech. rep. 35. Buenos Aires: Instituto Nacional de Estadística y Censos - INDEC.

Jaynes, E. T. (1968). "Prior Probabilities". In: *IEEE Transactions on Systems Science and Cybernetics* 4.3, pp. 227–241. ISSN: 0536-1567. DOI: 10.1109/TSSC.1968.300117.

Jorge, Maria Helena Prado de Mello, Ruy Laurenti, and Sabina Léa Davidson Gotlieb (2007). "Quality Analysis of Brazilian Vital Statistics: The Experience of Implementing the SIM and SINASC Systems". In: *Ciência & Saúde Coletiva* 12.3, pp. 643–654. ISSN: 1413-8123. DOI: 10.1590/S1413-81232007000300014.

Kéry, Marc and Michael Schaub (2012). "Chapter 11 - Estimation of Demographic Rates, Population Size, and Projection Matrices from Multiple Data Types Using Integrated Population Models". In: *Bayesian Population Analysis Using WinBUGS*. Boston: Academic Press, pp. 347–381. ISBN: 978-0-12-387020-9. DOI: 10.1016/B978-0-12-387020-9.00011-0.

Kintner, Hallie J. and David A. Swanson (1993). "Measurement Errors in Census Counts and Estimates of Intercensal Net Migration". In: *Journal of Economic and Social Measurement* 19.2, pp. 97–120.

Lee, Ronald D. (1974). "Estimating Series of Vital Rates and Age Structures from Baptisms and Burials: A New Technique, with Applications to Pre-Industrial England". In: *Population Studies* 28.3, pp. 495–512. ISSN: 0032-4728. DOI: 10.2307/2173642.

— (1982). "Correcting Census Age Distributions: Extensions and Applications of the Demeny-Shorter Technique." Berkeley.

— (1985). "Inverse Projection and Back Projection: A Critical Appraisal, and Comparative Results for England, 1539 to 1871". In: *Population Studies* 39.2, pp. 233–248. ISSN: 0032-4728.

— (1998). "Probabilistic Approaches to Population Forecasting". In: *Population and Development Review* 24, pp. 156–190. ISSN: 0098-7921.

— (2004). "Reflections on Inverse Projection: Its Origins, Development, Extensions, and Relation to Forecasting". en. In: *Inverse Projection Techniques*. Ed. by Dr Elisabetta Barbi, Professor Salvatore Bertino, and Professor Eugenio Sonnino. Demographic Research Monographs. Springer Berlin Heidelberg, pp. 1–9. ISBN: 978-3-642-05892-9 978-3-662-08016-0. DOI: 10.1007/978-3-662-08016-0_1.

Lee, Ronald D. and Lawrence R. Carter (1992). "Modeling and Forecasting U.S. Mortality". In: *Journal of the American Statistical Association* 87.419, pp. 659–671. ISSN: 0162-1459. DOI: 10.1080/01621459.1992.10475265.

Lee, Ronald D. and D. Lam (1983). "Age Distribution Adjustments for English Censuses, 1821 to 1931". In: *Population Studies* 37.3, pp. 445–464. ISSN: 0032-4728. DOI: 10.1080/00324728.1983.10408872.

Leslie, P. H. (1945). "On the Use of Matrices in Certain Population Mathematics". In: *Biometrika* 33.3, pp. 183–212. ISSN: 0006-3444. DOI: 10.2307/2332297.

Liu, P and Adrian Raftery (2017). "Accounting for Measurement Error in Bayesian Probabilistic Projection of the Total Fertility Rate". In: *PAA 2017 Annual Meeting*. Chicago: PAA.

Luther, Norman Y. and Robert D. Retherford (1988). "Consistent Correction of Census and Vital Registration Data". In: *Mathematical population studies* 1.1, pp. 1–20.

Malta, Deborah Carvalho et al. (2011). "Apresentação Do Plano de Ações Estratégicas Para o Enfrentamento Das Doenças Crônicas Não Transmissíveis No Brasil, 2011 a 2022". In: *Epidemiologia e Serviços de Saúde* 20.4, pp. 425–438. ISSN: 1679-4974. DOI: 10.5123/S1679-49742011000400002.

Masquelier, B. (2012). "Adult Mortality from Sibling Survival Data: A Reappraisal of Selection Biases". In: *Demography*, pp. 1–22.

McElreath, Richard (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. en. Boca Raton, FL: CRC Press.

Mikkelsen, Lene et al. (2015). "A Global Assessment of Civil Registration and Vital Statistics Systems: Monitoring Data Quality and Progress". en. In: *The Lancet* 386.10001, pp. 1395–1406. ISSN: 01406736. DOI: 10.1016/S0140-6736(15)60171-4.

Moreno, Elías and Javier Girón (1998). "Estimating with Incomplete Count Data A Bayesian Approach". In: *Journal of Statistical Planning and Inference* 66.1, pp. 147–159. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(97)00073-6.

Mortara, Giorgio (1941). "Estudos Sobre a Utilização Do Censo Demográfico Para a Reconstrução Das Estatísticas Do Movimento Da População Do Brasil: VI Sinopse Da Dinâmica Da População Do Brasil Nos Últimos Cem Anos". In: *Rev. bras. Estat* 2, pp. 267–276.

Moultrie, Tom A. (2013a). "General Assessment of Age and Sex Data". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

— (2013b). "Introduction to Fertility Analysis". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

— (2013c). "Overview of Fertility Estimation Methods Based on the P/F Ratio". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Moultrie, Tom A. and Rob Dorrington (2008). "Sources of Error and Bias in Methods of Fertility Estimation Contingent on the P/F Ratio in a Time of Declining Fertility and Rising Mortality". English. In: *Demographic Research; Rostock* 19, pp. 1635–1662. ISSN: 14359871.

Moultrie, Tom A. and B Zaba (2013). "Cohort Parity Comparison with Vital Registration Data". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

MRE (2016). *Estimativas Populacionais Das Comunidades Brasileiras No Mundo - 2015*.

Murray, Christopher J. L. et al. (2010). "What Can We Conclude from Death Registration? Improved Methods for Evaluating Completeness". In: *PLoS Medicine* 7.4. ISSN: 1549-1277. DOI: 10.1371/journal.pmed.1000262.

Noumbissi, Amadou (1992). "L'indice de Whipple Modifié: Une Application Aux Données Du Cameroun, de La Suède et de La Belgique". In: *Population (French Edition)* 47.4, pp. 1038–1041. ISSN: 0032-4663. DOI: 10.2307/1533772.

Ntozi, James Patrick Manyenye (1978). "The Demeny-Shorter and Three-Census Methods for Correcting Age Data". en. In: *Demography* 15.4, pp. 509–521. ISSN: 0070-3370, 1533-7790. DOI: 10.2307/2061203.

O'Hare, William P (2015). *The Undercount of Young Children in the US Decennial Census*. Springer. ISBN: 3-319-18917-4.

Oliveira, A. T. R. (1996). "Notas Sobre a Migração Internacional No Brasil Na Década de 80". In: *Migrações Internacionais: Herança XX, Agenda XXI*. Campinas, pp. 239–57.

— (2013). "A Panorama of International Migration Based on the 2010 Demographic Census". In: *REMHU: Revista Interdisciplinar da Mobilidade Humana* 21.40, pp. 195–210. ISSN: 1980-8585. DOI: 10.1590/S1980-85852013000100012.

— (2015). "O perfil geral dos imigrantes no Brasil a partir dos censos demográficos 2000 e 2010". pt. In: *PÉRIPLOS. Revista de Pesquisa sobre Migrações* 1.2. ISSN: 2448-1076.

— (2017). "A reforma deformada". pt. In: *Cadernos de Saúde Pública* 33, e00052317. ISSN: 0102-311X, 0102-311X, 1678-4464. DOI: 10.1590/0102-311x00052317.

Oliveira, A. T. R. (2018). "Panorama Das Estatísticas Vitais No Brasil". In: *Sistemas de Estatísticas Vitais No Brasil : Avanços, Perspectivas e Desafios*. Estudos e análises. Informação demográfica e socioeconômica 7. Rio de Janeiro: IBGE, Coordenação de População e Indicadores Sociais. ISBN: 2236-5265.

Oliveira, G. L., Rosangela Helena Loschi, and Renato Martins Assunção (2017). "A Random-Censoring Poisson Model for Underreported Data". en. In: *Statistics in Medicine* 36.30, pp. 4873–4892. ISSN: 1097-0258. DOI: 10.1002/sim.7456.

Oliveira, L. C. S., M. P. S. de Freitas, et al. (2003). *Censo Demográfico 2000: Resultados Da Pesquisa de Avaliação Da Cobertura Da Coleta*. Textos Para Discussão 9. Rio de Janeiro: IBGE.

Oliveira, L. C. S., L. B. Indá, et al. (1996). *Avaliação Da Cobertura Da Coleta Do Censo Demográfico de 1991*. Textos Para Discussão 84. Rio de Janeiro: IBGE.

Olkin, Ingram, A. John Petkau, and James V. Zidek (1981). "A Comparison of n Estimators for the Binomial Distribution". In: *Journal of the American Statistical Association* 76.375, pp. 637–642. ISSN: 0162-1459. DOI: 10.1080/01621459.1981.10477697.

ONS (2012). *2011 Census: Population and Household Estimates for England and Wales*.

Otis, David L. et al. (1978). "Statistical Inference from Capture Data on Closed Animal Populations". In: *Wildlife Monographs* 62, pp. 3–135. ISSN: 0084-0173.

Passel, Jeffrey (2007). "Unauthorized Migrants in the United States: Estimates, Methods, and Characteristics". en. In: ISSN: 1815-199X. DOI: 10.1787/110780068151.

Pizarro, Jorge Martínez and Miguel Villa (2005). "International Migration in Latin America and the Caribbean: A Summary View of Trends and Patterns". In: *UN Expert Group Meeting on International Migration and Development, New York, July*. Citeseer.

Poole, D and Adrian Raftery (2000). "Inference for Deterministic Simulation Models: The Bayesian Melding Approach". In: *Journal of the American Statistical Association* 95.452, pp. 1244–1255. ISSN: 0162-1459. DOI: 10.1080/01621459.2000.10474324.

Preston, Samuel H, Ansley Coale, et al. (1980). "Estimating the Completeness of Reporting of Adult Deaths in Populations That Are Approximately Stable". In: *Population Index*, pp. 179–202. ISSN: 0032-4701.

Preston, Samuel H and Irma T Elo (1999). "Effects of Age Misreporting on Mortality Estimates at Older Ages". In: *Population studies* 53.2, pp. 165–177. ISSN: 0032-4728.

Preston, Samuel H, Irma T Elo, et al. (1998). "Reconstructing the Size of the African American Population by Age and Sex, 1930–1990". en. In: *Demography* 35.1, pp. 1–21. ISSN: 0070-3370, 1533-7790. DOI: 10.2307/3004023.

Preston, Samuel H, Patrick Heuveline, and Michel Guillot (2001). *Demography: Measuring and Modeling Population Processes*. Malden, MA: Blackwell Publishers.

Preston, Samuel H and Kenneth Hill (1980). "Estimating the Completeness of Death Registration". eng. In: *Population Studies* 34.2, pp. 349–366. ISSN: 0032-4728. DOI: 10.1080/00324728.1980.10410395.

Queiroz, Bernardo Lanza and Diana O. T. Sawyer (2012). "What Can the Mortality Data from the 2010 Census Tell Us?" In: *Revista Brasileira de Estudos de População* 29.2, pp. 225–238. ISSN: 0102-3098. DOI: 10.1590/S0102-30982012000200002.

Raftery, Adrian (1988). "Inference for the Binomial N Parameter: A Hierarchical Bayes Approach". en. In: *Biometrika* 75.2, pp. 223–228. ISSN: 0006-3444. DOI: 10.1093/biomet/75.2.223.

Raftery, Adrian et al. (2012). "Bayesian Probabilistic Population Projections for All Countries". en. In: *Proceedings of the National Academy of Sciences* 109.35, pp. 13915–13921. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1211452109.

Rajaratnam, Julie Knoll et al. (2010). "Neonatal, Postneonatal, Childhood, and under-5 Mortality for 187 Countries, 1970–2010: A Systematic Analysis of Progress towards Millennium Development Goal 4". English. In: *The Lancet* 375.9730, pp. 1988–2008. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(10)60703-9.

Raymer, James et al. (2013). "Integrated Modeling of European Migration". In: *Journal of the American Statistical Association* 108.503, pp. 801–819. ISSN: 0162-1459. DOI: 10.1080/01621459.2013.789435.

Ribeiro, Patrícia Dias and Moema Gonçalves Bueno Fígoli (2008). "Análise econômica e social da introdução do fator previdenciário na nova regra de cálculo dos benefícios da previdência social brasileira". pt. In: *Estudos sobre Previdência Social no Brasil: diagnóstico e propostas de reforma*. Vol. 1. E-book. Belo Horizonte: ABEP; UNFPA, pp. 37–62. ISBN: 978-85-85543-20-4.

RIPSA (2013). *Indicadores e Dados Básicos - Brasil - 2013*.

Robinson, J Gregory et al. (1993). "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis". In: *Journal of the American Statistical Association* 88.423, pp. 1061–1071. ISSN: 0162-1459. DOI: 10.1080/01621459.1993.10476375.

Romero, Dalia and Anitza Freitez (2008). "Problemas de Calidad de La Declaración de La Edad de La Población Adulta Mayor En Los Censos de América Latina de La Ronda Del 2000". In: *Trabajo Presentado En El III Congreso de La Asociación Latinoamericana de Población, Córdoba (Argentina)*. Vol. 24.

Schmertmann, Carl P. (2014). "Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data[1]". In: *Revista Brasileira de Estudos de População* 31.2, pp. 291–307. ISSN: 0102-3098. DOI: 10.1590/S0102-30982014000200004.

Schmertmann, Carl P., Suzana M. Cavenaghi, et al. (2013). "Bayes plus Brass: Estimating Total Fertility for Many Small Areas from Sparse Census Data". eng. In: *Population Studies* 67.3, pp. 255–273. ISSN: 1477-4747. DOI: 10.1080/00324728.2013.795602.

Schmertmann, Carl P. and Marcos R. Gonzaga (2018). "Bayesian Estimation of Age-Specific Mortality and Life Expectancy for Small Areas With Defective Vital Records". en. In: *Demography*, pp. 1–26. ISSN: 0070-3370, 1533-7790. DOI: 10.1007/s13524-018-0695-2.

Sekar, C. Chandra and W. Edwards Deming (1949). "On a Method of Estimating Birth and Death Rates and the Extent of Registration". In: *Journal of the American Statistical Association* 44.245, pp. 101–115. ISSN: 0162-1459. DOI: 10.1080/01621459.1949.10483294.

Sevcikova, Hana, Adrian Raftery, and Paul A. Waddell (2007). "Assessing Uncertainty in Urban Simulations Using Bayesian Melding". In: *Transportation Research Part B: Methodological* 41.6, pp. 652–669. ISSN: 0191-2615. DOI: 10.1016/j.trb.2006.11.001.

Sharrow, David J. et al. (2013). "The Age Pattern of Increases in Mortality Affected by HIV: Bayesian Fit of the Heligman-Pollard Model to Data from the Agincourt HDSS Field Site in Rural Northeast South Africa". In: *Demographic research* 29, pp. 1039–1096. ISSN: 1435-9871. DOI: 10.4054/DemRes.2013.29.39.

Siler, William (1979). "A Competing-Risk Model for Animal Mortality". en. In: *Ecology* 60.4, pp. 750–757. ISSN: 1939-9170. DOI: 10.2307/1936612.

— (1983). "Parameters of Mortality in Human Populations with Widely Varying Life Spans". en. In: *Statistics in Medicine* 2.3, pp. 373–380. ISSN: 1097-0258. DOI: 10.1002/sim.4780020309.

Silva, Andréa Diniz da, Marcos Paulo Soares de Freitas, and Djalma Galvão Carneiro Pessoa (2015). "Assessing Coverage of the 2010 Brazilian Census". In: *Statistical Journal of the IAOS* 31.2, pp. 215–225. ISSN: 1874-7655. DOI: 10.3233/sji-150897.

Silva, Romesh (2012). "Child Mortality Estimation: Consistency of Under-Five Mortality Rate Estimates Using Full Birth Histories and Summary Birth Histories". en. In: *PLOS Medicine* 9.8, e1001296. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001296.

Smith, Philip J. (1991). "Bayesian Analyses for a Multiple Capture-Recapture Model". en. In: *Biometrika* 78.2, pp. 399–407. ISSN: 0006-3444. DOI: 10.1093/biomet/78.2.399.

Soares, Weber and Dimitri Fazito (2008). "Quando o "Direito de Escolha" Não é Um Direito: Da Distinção Estrutural Entre Migrantes Internacionais Regulares e Irregulares Em Governador Valadares". In: *XVI Encontro Nacional de Estudos Populacionais*. Caxambu-MG: ABEP.

Spoorenberg, Thomas (2007). "Quality of Age Reporting : Extension and Application of the Modified Whipple's Index". en. In: *Population* Vol. 62.4, pp. 729–741. ISSN: 0032-4663.

SSA (2012). *Census 2011 - Post-Enumeration Survey: Results and Methodology*. Tech. rep. 03-01-46. Pretoria: Statistics South Africa, p. 84.

Stamey, James D., Dean M. Young, and Tom L. Bratcher (2006). "Bayesian Sample-Size Determination for One and Two Poisson Rate Parameters with Applications to Quality Control". In: *Journal of Applied Statistics* 33.6, pp. 583–594. ISSN: 0266-4763. DOI: 10.1080/02664760600679643.

Szwarcwald, Célia Landmann (2008). "Strategies for Improving the Monitoring of Vital Events in Brazil". eng. In: *International Journal of Epidemiology* 37.4, pp. 738–744. ISSN: 1464-3685. DOI: 10.1093/ije/dyn130.

Szwarcwald, Célia Landmann, Paulo Germano de Frias, et al. (2014). "Correction of Vital Statistics Based on a Proactive Search of Deaths and Live Births: Evidence from a Study of the North and Northeast Regions of Brazil". In: *Population Health Metrics* 12.1, p. 16. ISSN: 1478-7954. DOI: 10.1186/1478-7954-12-16.

Szwarcwald, Célia Landmann, Maria do Carmo Leal, et al. (2002). "Infant Mortality Estimation in Brazil: What Do Ministry of Health Data on Deaths and Live Births Say?" In: *Cadernos de Saúde Pública* 18.6, pp. 1725–1736. ISSN: 0102-311X. DOI: 10.1590/S0102-311X2002000600027.

Szwarcwald, Célia Landmann, OL Morais Neto, et al. (2010). "Busca Ativa de Óbitos e Nascimentos No Nordeste e Na Amazônia Legal: Estimação Das Coberturas Do SIM e

Do SINASC Nos Municípios Brasileiros". In: *Ministério da Saúde, organizador. Saúde Brasil*.

Team, Stan Development (2017). *Stan Modeling Language: User's Guide and Reference Manual*. Tech. rep. Version 2.17.0.

Thomas, Kevin and Kenneth Hill (2007). "What Can Data on Household Deaths Tell Us about Adult Mortality in Lesotho and Botswana?" In: *Union for African Population Studies (UAPS)*. Arusha, Tanzania.

Timæus, Ian M. (2013a). "Indirect Estimation of Adult Mortality from Data on Siblings". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

— (2013b). "Indirect Estimation of Adult Mortality from Orphanhood". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Timæus, Ian M., Rob Dorrington, and Kenneth Hill (2013). "Introduction to Adult Mortality Analysis". In: *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Trindade, J. E. O., L. F. L. Costa, and A. T. R. Oliveira (2018). "Aplicação Do Método Captura-Recaptura Aos Dados de Estatísticas Vitais: Estudo Empírico". In: *Sistemas de Estatísticas Vitais No Brasil : Avanços, Perspectivas e Desafios*. Estudos e análises. Informação demográfica e socioeconômica 7. Rio de Janeiro: IBGE, Coordenação de População e Indicadores Sociais. ISBN: 2236-5265.

Trussell, J. and G. Rodríguez (1990 Nov-Dec). "A Note on the Sisterhood Estimator of Maternal Mortality". eng. In: *Studies in Family Planning* 21.6, pp. 344–346. ISSN: 0039-3665.

UN (1955). *Age and Sex Patterns of Mortality Model Life-Tables for Under- Developed Countries*. Tech. rep. 22. New York: United Nations.

— (1982). *Model Life Tables for Developing Countries*. Tech. rep. 77. New York: United Nations.

— (1983). *Manual X: Indirect Techniques for Demographic Estimation*. ST/ESA/SER.A/81. Population Studies 81. New York: United Nations.

— (2014). *World Population Prospects: The 2012 Revision, Methodology of the United Nations Population Estimates and Projections*. Tech. rep. ESA/P/WP.235. United Nations, Department of Economic and Social Affairs, Population Division.

UNSD (2004). *Handbook on the Collection of Fertility and Mortality Data*. Vol. 92. United Nations Publications. ISBN: 92-1-161462-7.

— (2008). *Principles and Recommendations for Population and Housing Censuses*. United Nations Publications. ISBN: 92-1-161505-4.

— (2010). *Post Enumeration Surveys: Operational Guidelines*. Technical Report. New York.

US Bureau of the Census (1985). *Evaluating Censuses of Population and Housing*. Statistical Training Document ISP-TR-5. Washington, D. C.

Wachter, Kenneth W (1986). "Ergodicity and Inverse Projection". In: *Population Studies* 40.2, pp. 275–287. ISSN: 0032-4728. DOI: 10.1080/0032472031000142076.

Wachter, Kenneth W (2014). *Essential Demographic Methods*. Harvard University Press. ISBN: 0-674-04557-2.

Wachter, Kenneth W and David A Freedman (2000). "The Fifth Cell: Correlation Bias in U.S. Census Adjustment". en. In: *Evaluation Review* 24.2, pp. 191–211. ISSN: 0193-841X. DOI: 10.1177/0193841X0002400202.

West, Kirsten K and J Gregory Robinson (1999). *What Do We Know About the Undercount of Children?* US Department of Commerce, Economics and Statistics Administration.

Wheldon, Mark C et al. (2013). "Reconstructing Past Populations with Uncertainty from Fragmentary Data". In: *Journal of the American Statistical Association* 108.501, pp. 96–110. ISSN: 0162-1459.

— (2016). "Bayesian Population Reconstruction of Female Populations for Less Developed and More Developed Countries". In: *Population studies*, pp. 1–17. ISSN: 0032-4728.

Wilmoth, John et al. (2012). "A Flexible Two-Dimensional Mortality Model for Use in Indirect Estimation". In: *Population Studies* 66.1, pp. 1–28. ISSN: 0032-4728. DOI: 10.1080/00324728.2011.611411.

Wolter, Kirk M. (1986). "Some Coverage Error Models for Census Data". In: *Journal of the American Statistical Association* 81.394, pp. 337–346. ISSN: 0162-1459. DOI: 10.1080/01621459.1986.10478277.

Abbreviations

UK United Kingdom. 44, 48, 114

US United States. 44, 48, 114, 117, 119

NSO National Statistical Office. 9, 44, 47, 50

IBGE Brazilian Institute of Geography and Statistics. v–vii, 50–53, 56, 58, 60, 63–65, 67, 68, 71, 83, 84, 87–90, 108, 118, 119, 121–124, 132, 135, 136, 138, 140, 142, 144, 146–148, 150, 152, 154, 156, 158, 198–200

CELADE Latin American and Caribbean Demographic Center. 47

MCMC Markov Chain Monte Carlo. 35

HMC Hamiltonian Monte Carlo. 14, 19, 23, 24, 31, 35, 38

PES Post-Enumeration Survey. ix, 45–48, 51–59, 67–69, 72, 116, 117, 128, 130, 140, 141, 146

DSE Dual System Estimator. ix, 45, 46

DA Demographic Analysis. ix, 45, 47, 48, 52, 56, 58, 59, 66–69

CSR Cohort Survival Ratio. iii, vii, viii, 66, 67, 115, 201–210

SR Sex Ratio. iii, vii, 64, 65, 197–200

RO Rondônia. 54, 56, 118, 122, 136, 138

AC Acre. 41, 54, 56, 96, 138

AM Amazonas. 56

RR Roraima. 41, 82, 87, 118, 122

AP Amapá. 41, 54, 96, 124, 136, 138

TO Tocantins. 122

DHS Demographic and Health Survey. 5, 70, 89

TFR Total Fertility Rate. iv–vii, 11, 31, 32, 36, 84, 86–88, 131, 135, 138, 140, 145, 146, 152, 158

ASFR Age-Specific Fertility Rate. 29, 75–77, 82, 131

IMR Infant Mortality Rate. 91, 94

DDM Death Distribution Methods. 7, 29, 99, 101, 103, 105, 129, 131

SEG Synthetic Extinct Generations. 7, 99, 101–103, 105, 131

GGB General Growth Balance. 100, 102, 103, 105, 131

Appendix A

Probability distributions

This appendix show some features and properties of the main statistical distributions used in this dissertation.

A.1 Poisson distribution

The Poisson distribution is often used to express the probability of a certain number of events occurring in a fixed period of time. It has been used to model demographic events such as births and deaths, as well as population counts (Brillinger, 1986).

The density function of a random variable θ Poisson-distributed with rate parameters λ, for $\lambda > 0$, $\theta \sim Poisson(\lambda)$ is:

$$p(\theta) = \frac{\lambda^{\theta} e^{-\lambda}}{(\theta!)} \tag{A.1}$$

An interesting feature of the Poisson distribution is that the mean is equal to the variance:

$$E(\theta) = \lambda \tag{A.2}$$

$$var(\theta) = \lambda \tag{A.3}$$

Non-informative prior distributions for the parameter λ

A natural conjugate prior for the Poisson distribution is the gamma distribution. However, in cases where almost no information on λ is available, this parameter requires a non-informative prior.

One of the most commonly used approaches to define a non-informative prior distribution is to use Jeffreys' invariance principle, which states that the prior is invariant to transformations in the parameter. Jeffreys' priors $p(\lambda)$ are defined as:

$$p(\lambda) \propto \sqrt{I(\lambda)} \tag{A.4}$$

where $I(\lambda)$ is Fisher information for λ.

For the Poisson distribution with rate parameter λ, $\theta \sim Poisson(\lambda)$, the Fisher information is given by: $I(\lambda) = \frac{1}{\lambda}$. Thus the Jeffreys' prior is:

$$p(\lambda) \propto \frac{1}{\sqrt{\lambda}} \tag{A.5}$$

which is an improper prior.

Another option of non-informative prior for λ was proposed by Jaynes, (1968) and it is given by:

$$p(\lambda) \propto \frac{1}{\lambda} \tag{A.6}$$

This prior distribution implies a uniform distribution for $log(\lambda)$ and is also an improper prior. This has been used in several studies due to its properties in conjugacy with the gamma (Stamey, Young, and T. L. Bratcher, 2006) and beta distributions (Moreno and Girón, 1998).

In the case of this dissertation, when a non-informative prior distributions for the parameter λ is needed, the most sensible choice is the improper uniform prior, which does not have a heavily concentrated mass of probabilities close to zero, as the other priors described above do:

$$p(\lambda) \propto 1 \tag{A.7}$$

A.2 Beta distribution

The beta distribution has been used to represent the variability of a parameter over the range $[0; 1]$. It has the advantage of being highly flexible.

The density function of a random variable θ beta-distributed with shape parameters a and b, $\theta \sim Beta(a, b)$ is:

$$p(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a, b)} \tag{A.8}$$

where $B(a, b)$ is a normalizing constant to make the density function integrates to one and it is given by the following ratio of the gamma functions Γ: $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

The mean and the variance of θ beta-distributed are given by:

$$E(\theta) = \frac{a}{a+b} \tag{A.9}$$

$$var(\theta) = \frac{ab}{(a+b)^2(a+b+1)} \tag{A.10}$$

For a proper density, both a and b need to be positive. If $a = 1$ and $b = 1$, the density is equal to the uniform density. If a>1 and b>1, the mode of the density is:

$$mode(\theta) = \frac{a-1}{a+b-2} \tag{A.11}$$

Given the mode of the beta distribution (A.11), $(a-1)$ can be interpreted as the number os successes and $(b-1)$ the number of failures.

When $a = b$, equation A.11 simplifies to $\frac{1}{2}$ and the distribution is symmetric. If $a \neq b$, the distribution is skewed. Most of the beta distributions used in this study are negatively skewed, meaning that the distribution is highly concentrated near one. In these cases, a is much greater than b. This is more common when modeling census coverage, which tend to have higher coverage. A sensible example of beta distribution to model census coverage of a certain age group c is $\kappa_c \sim Beta(50, 2)$, which leads to a mean of 0.96 and mode of 0.98 (see Figure A.1). For completeness of registered births and deaths, it is more common to find a more diverse pattern, with some population groups having extremely low coverage, whereas other regions show almost complete vital registration. For a population group with completeness of registered deaths of about 80% a reasonable distribution will be similar to $\delta_c \sim Beta(50, 10)$ (see Figure A.1).

Figure A.1 shows six examples of beta distributions with different values of a and b and their implied mean, variance and mode.

Method of moments

The shape parameters a and b of $\theta \sim Beta(a, b)$ can be estimated by the method of moments, using a given estimate of the mean $E(\theta)$ and the variance $var(\theta)$.

$$a = E(\theta)\left(\frac{E(\theta)(1-E(\theta))}{var(\theta)} - 1\right) \tag{A.12}$$

$$b = a\left(\frac{1}{E(\theta)} - 1\right) \tag{A.13}$$

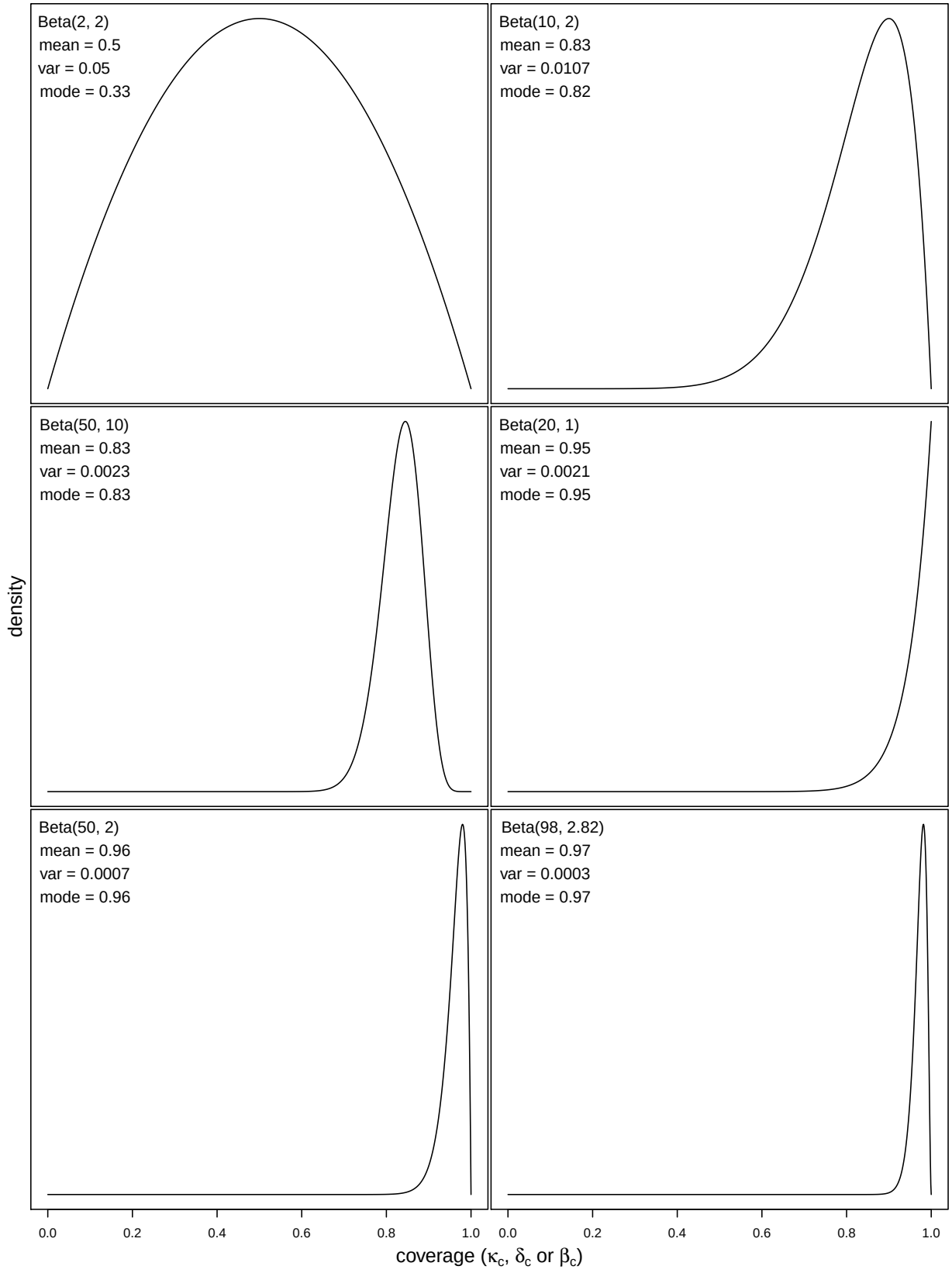

Figure A.1: Probability density function of a beta distribution with different shape parameters a and b and the associated mean, variance and mode

The last panel of Figure A.1 shows the beta distribution with mean $E(\theta) = 0.9$ and variance $var(\theta) = 0.005$. The parameters a and b are estimated by using equation A.12:

$$a = 0.9\left(\frac{0.9(1-0.9)}{0.005} - 1\right) = 15.3 \tag{A.14}$$

$$b = 15.3\left(\frac{1}{0.9} - 1\right) = 1.7 \tag{A.15}$$

which leads to $\theta \sim Beta(15.3, 1.7)$. Notice that since this distribution is highly skewed, the mode (0.95) is significantly different from the mean (0.90).

Method of quantiles

When there is only incomplete or unreliable information about the moments of θ, an alternative to find the parameters a and b is to find this parameters iteratively by trial-and-error until they match two pre-defined quantiles of the distribution.

For example, imagine that there is evidence that census coverage for certain population group, κ_c, is likely situated between 0.95 and 0.99. It is possible to find a beta distribution that returns these values as the 10th and 90th percentiles, which is $\kappa \sim Beta(98, 2.82)$.

The prior information about the percentiles, mean and variance of the census coverage can be obtained from expert opinion, PES of the census being modeled or from previous censuses. Data from different contexts, such as PES from different countries, could be also used, although in these cases it would be more prudent to have a higher variance.

Beta-binomial distribution

In Bayesian inference, the beta distribution is the conjugate prior probability distribution for several discrete distributions, such as Bernoulli, binomial, negative binomial and geometric. For studying coverage probabilities for population, death and birth counts, the most relevant mixture distribution is the Beta-binomial. This mixture is given by a binomial distribution with probability of success following a beta distribution.

The mean and variance of a random variable θ Beta-binomial distributed $\theta \sim BetaBinomial(n, a, b)$ are as follows:

$$E(\theta) = \frac{n\alpha}{\alpha + \beta} \tag{A.16}$$

$$var(\theta) = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{A.17}$$

Beta-Poisson Mixture Distribution

The Beta-binomial distribution could, in principle, be used to model the population observed in a census K_c^{obs} as a result of the following hierarchical structure:

$$K_c^{obs} \sim Binomial(K_c, \kappa_c) \tag{A.18}$$

$$\kappa_c \sim Beta(a_c^K, b_c^K) \tag{A.19}$$

However, unlike in the most common examples of the beta-binomial distribution, in this case K_c is unknown and almost no information about this parameter is available.

If K_c follows a Poisson distribution with mean equal to a parameter λ_c, $K_c \sim Poisson(\lambda_c)$, then:

$$p(K_c) = \frac{{\lambda_c}^{K_c} e^{-\lambda_c}}{K_c!} \tag{A.20}$$

$$p(K_c^{obs}) = \binom{K_c}{K_c^{obs}} (\kappa_c)^{K_c^{obs}} (1-\kappa_c)^{K_c - K_c^{obs}} \tag{A.21}$$

It follows that the likelihood function for λ_c and κ_c given the observed value K_c^{obs} is:

$$L(K_c^{obs}|\lambda_c, \kappa_c) = \frac{(\lambda_c\kappa_c)^{K_c} exp(-\lambda_c\kappa_c)}{K_c!} \tag{A.22}$$

This shows that observed value K_c^{obs} has a Poisson distribution with mean $\lambda_c\kappa_c$: $K_c^{obs} \sim Poisson(\lambda_c\kappa_c)$.

Identification problem arises from this likelihood, since inference through maximum likelihood only allows estimation for the product $\lambda_c\kappa_c$ and not each parameter individually, which are ultimately the measures of interest.

Moreno and Girón, (1998) shows that if λ_c follows the Jaynes prior ($\lambda_c \propto \frac{1}{\lambda}$) and κ_c is beta distributed $\kappa_c \sim Beta(a_c^K, b_c^K)$, the joint distribution is given by:

$$p(\lambda_c, \kappa_c) = \frac{(\kappa_c)^{a_c^K + K_c^{obs} - 1}(1-\kappa_c)^{b_c^K - 1}(\lambda_c)^{K_c^{obs}-1}(e)^{-\lambda_c\kappa_c}}{\Gamma(K_c^{obs})B(a_c^K, b_c^K)} \tag{A.23}$$

Under these distributional assumptions, the posterior means are:

$$E(\lambda_c|K_c^{obs}) = E(K_c|K_c^{obs}) = \left(\frac{a_c^K + b_c^K - 1}{a_c^K - 1}\right) K_c^{obs} \tag{A.24}$$

A.3 Gamma Distribution

The gamma distribution is also highly flexible and is parametrized by a shape parameter $a > 0$ and and an inverse scale parameter (rate) $b > 0$:

The density function of a random variable θ gamma-distributed with parameters a and b ($\theta \sim Gamma(a, b)$) is:

$$p(\theta) = \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} \tag{A.25}$$

where $\Gamma(a)$ is the gamma function Γ of the shape parameter a.

The mean and the variance are given by:

$$E(\theta) = \frac{a}{b} \tag{A.26}$$

$$var(\theta) = \frac{a}{b^2} \tag{A.27}$$

The gamma distribution has the scaling property, so that $c\theta \sim Gamma(a, \frac{b}{c})$.

For $a \geq 1$, the mode of the gamma density is:

$$mode(\theta) = \frac{a-1}{b} \tag{A.28}$$

Method of moments

The parameters a and b of $\theta \sim Gamma(a, b)$ can be estimated by the method of moments, using some estimate of the mean $E(\theta)$ and the variance $var(\theta)$.

$$b = \frac{E(\theta)}{var(\theta)} \tag{A.29}$$

$$a = E(\theta)b \tag{A.30}$$

Using the same same mean $E(\theta) = 0.9$ and variance $var(\theta) = 0.005$ of the illustration for the beta distribution (section A.2), leads to a $Gamma(162, 180)$.

$$b = \frac{0.9}{0.005} = 180 \tag{A.31}$$

$$a = 0.9 \cdot 180 = 162 \tag{A.32}$$

Unlike the beta distribution, the gamma distribution is not restricted to the interval $[0, 1]$ and my have non-zero probability over one. In fact, the probability of θ being greater than one under this distribution is 8.2%. This is important when modeling the completeness of registered deaths (δ_c^*) and births (β_c^*) relative to the census coverage. Values greater than one would indicate that completeness of registered deaths (δ_c) or births (β_c) is higher than census coverage (κ_c).

Method of quantiles

The parameters a and b of a gamma distribution can be also estimated iteratively by trial-and-error until they match two pre-defined quantiles of the distribution.

Using the same percentiles in the illustration for the beta distribution (Section A.2), imagine that there is evidence that the completeness of registered deaths relative to census coverage for a certain population group, δ_c^*, is likely situated between 0.95 and 0.99. It is possible to find a gamma distribution that returns these values as the 10th and 90th percentiles, which is $\delta_c^* \sim Gamma(3863, 3982)$.

The data source to estimate the hyperparameters of a gamma distribution varies. For the δ_c^*, the natural source are the results from death distribution methods. Results from different methods and variants of the methods can be used to derive the quantiles, mean and variance of the parameter.

Appendix B

Evaluation of Census in the Brazilian States from 1980 to 2010

B.1 Population Pyramids

This section shows the plots of the population pyramids by single year of age for the 27 Brazilian states in absolute numbers. Scales are fixed for each state.

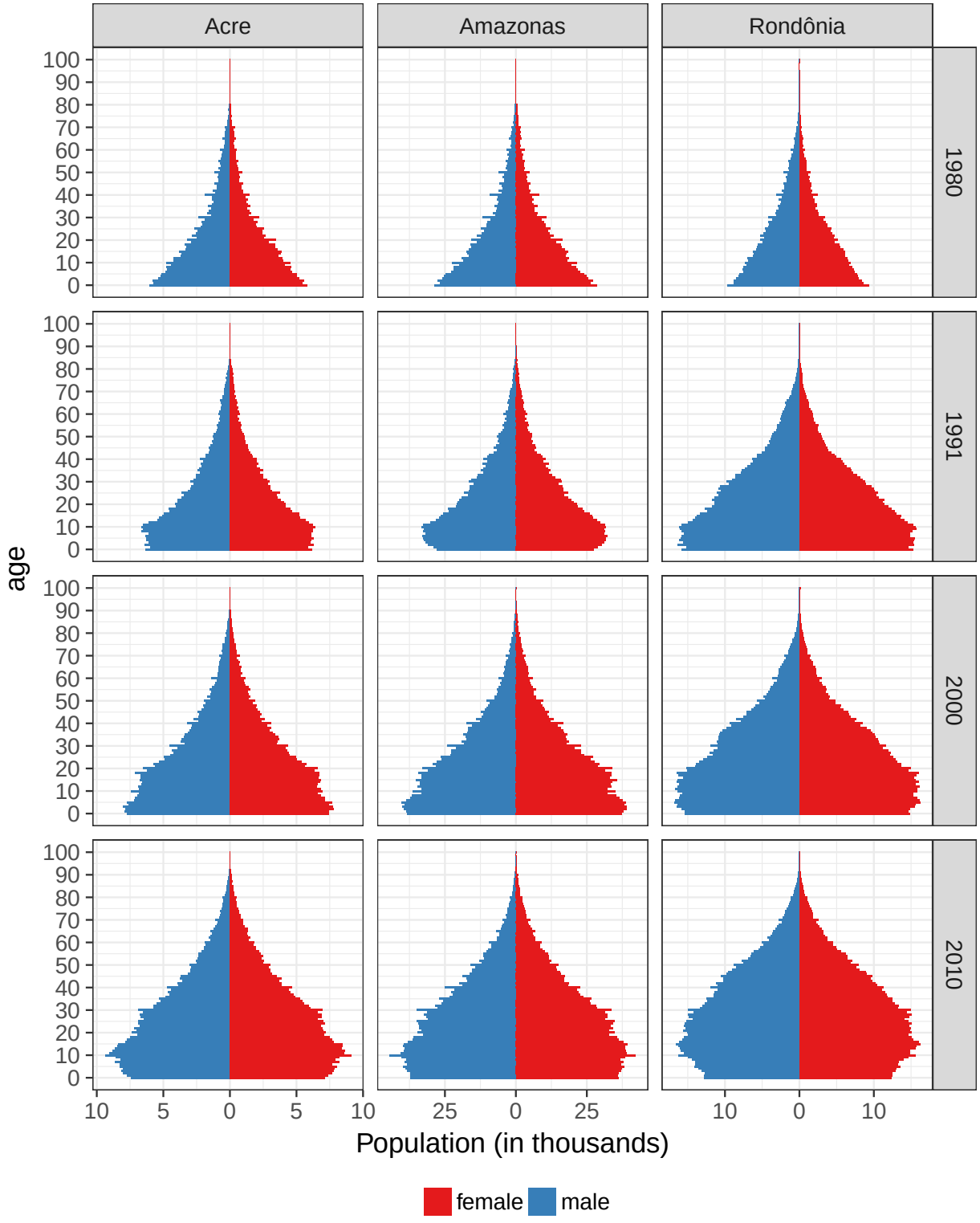

Figure B.1: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

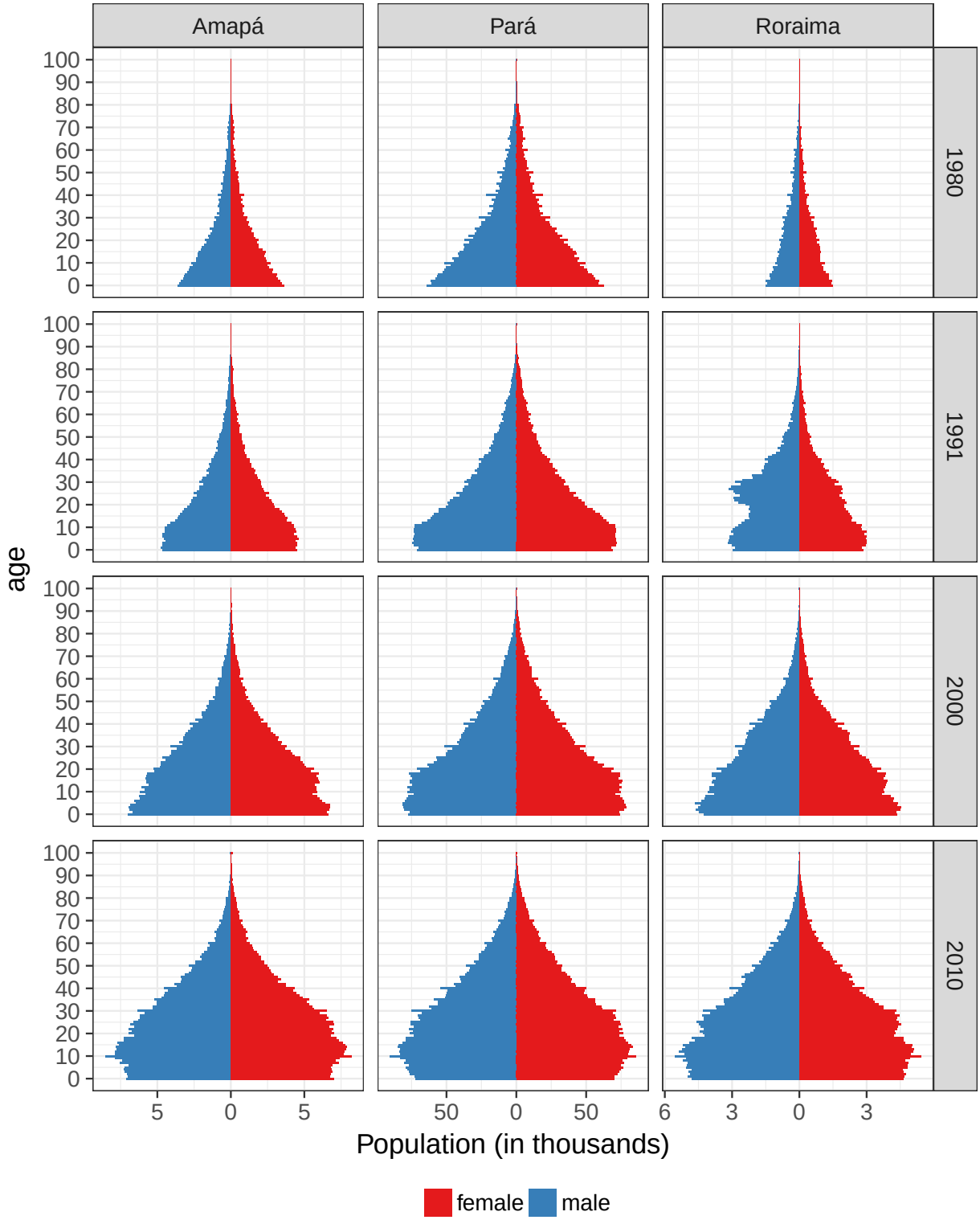

Figure B.2: [

Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions)]Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

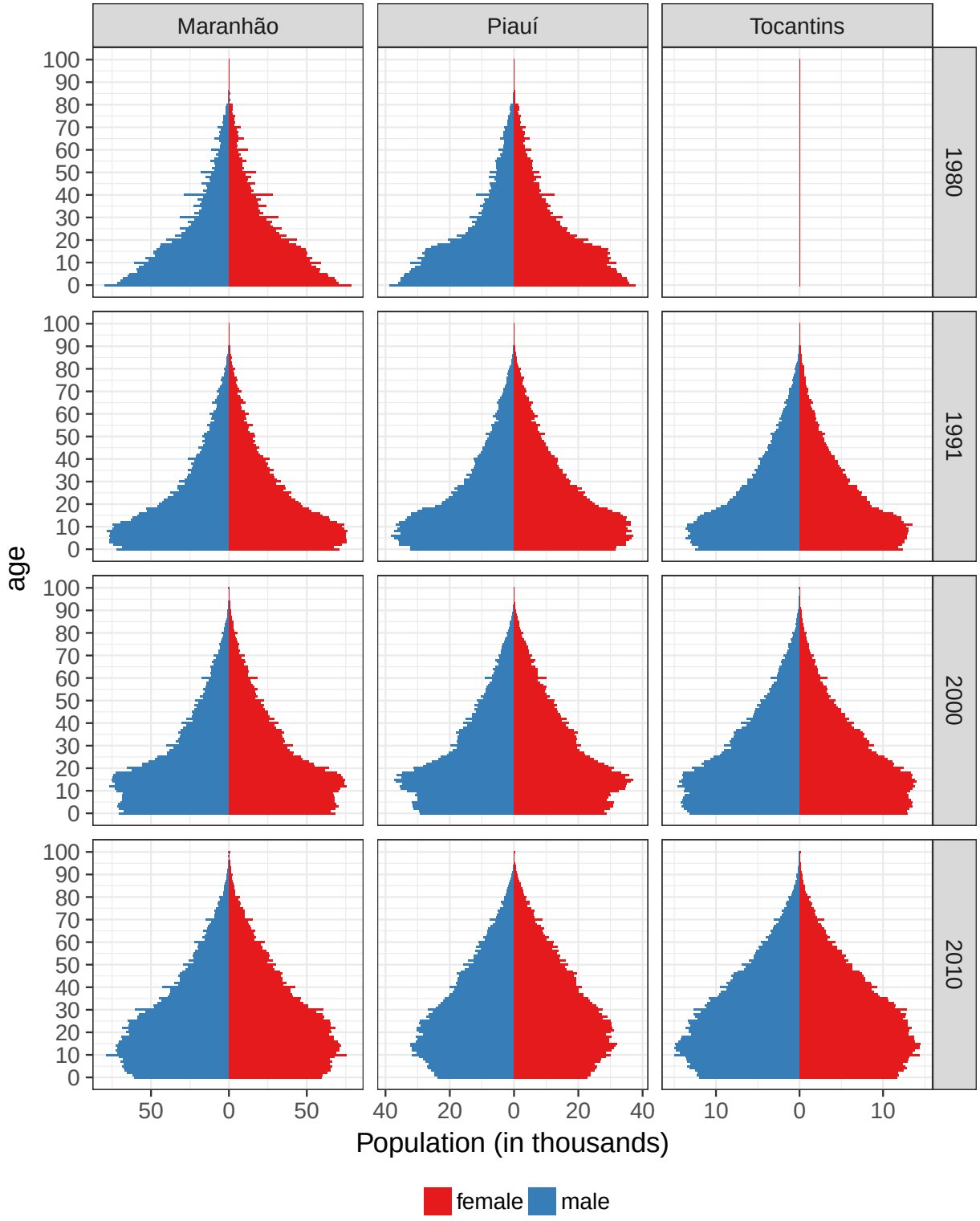

Figure B.3: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

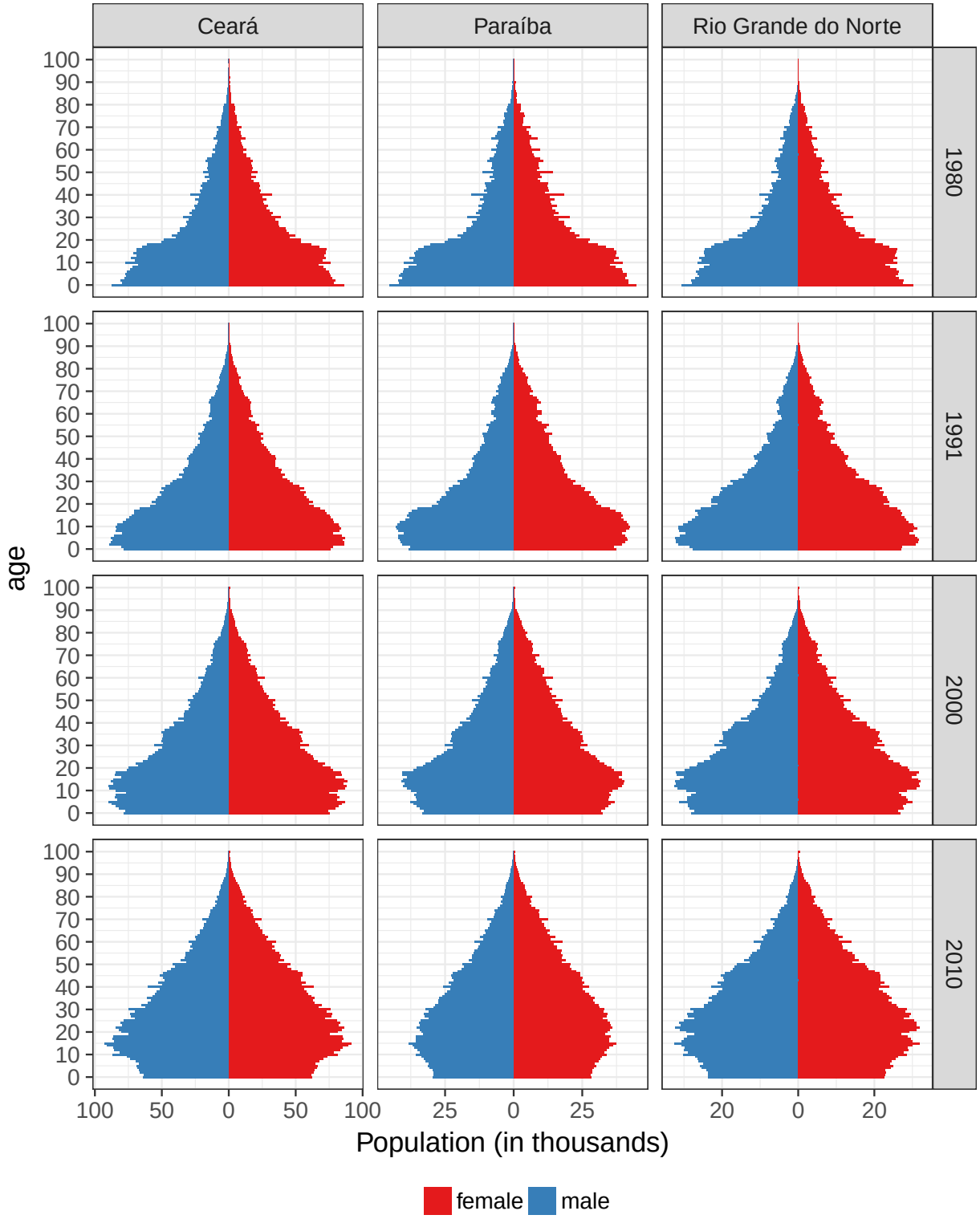

Figure B.4: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

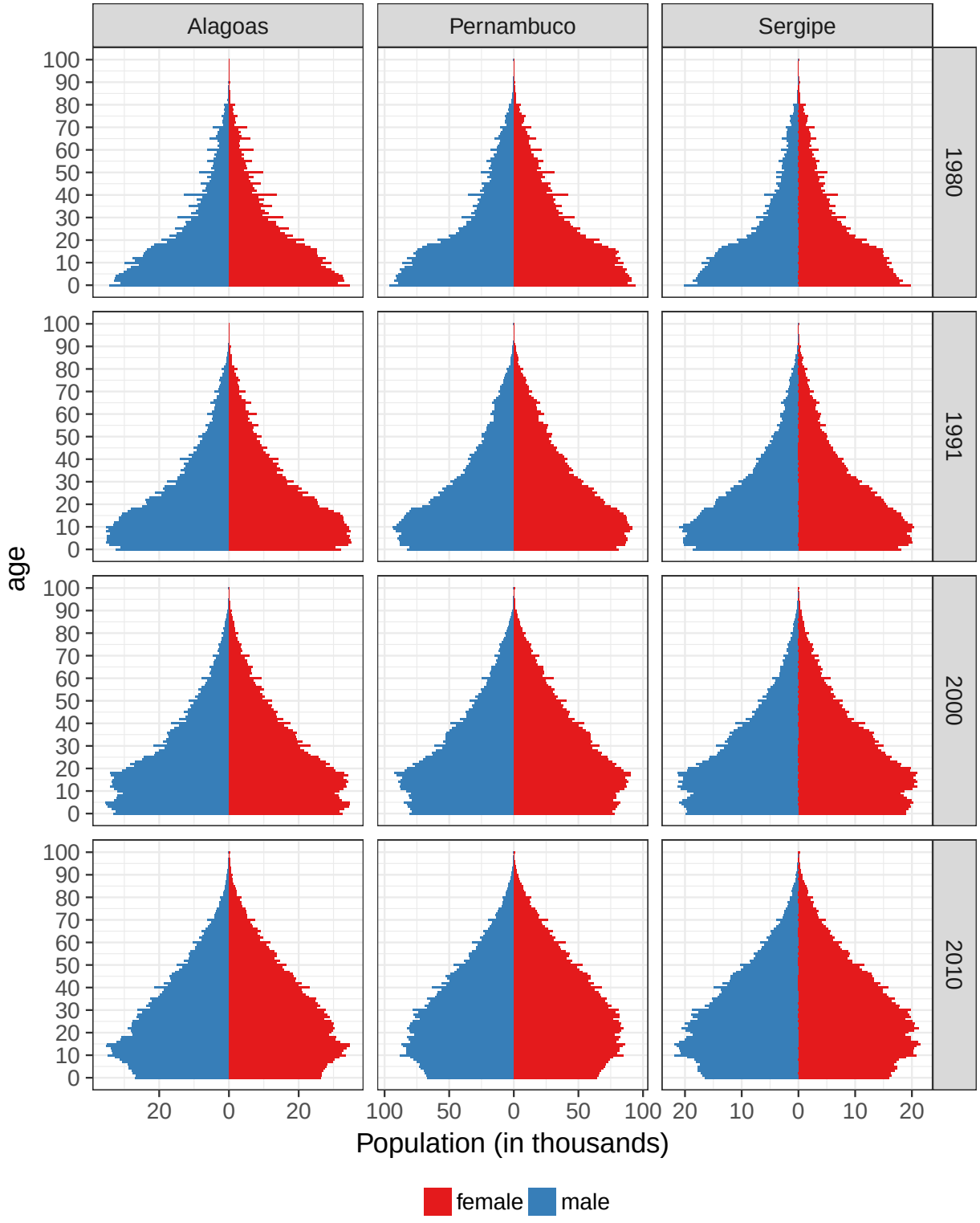

Figure B.5: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

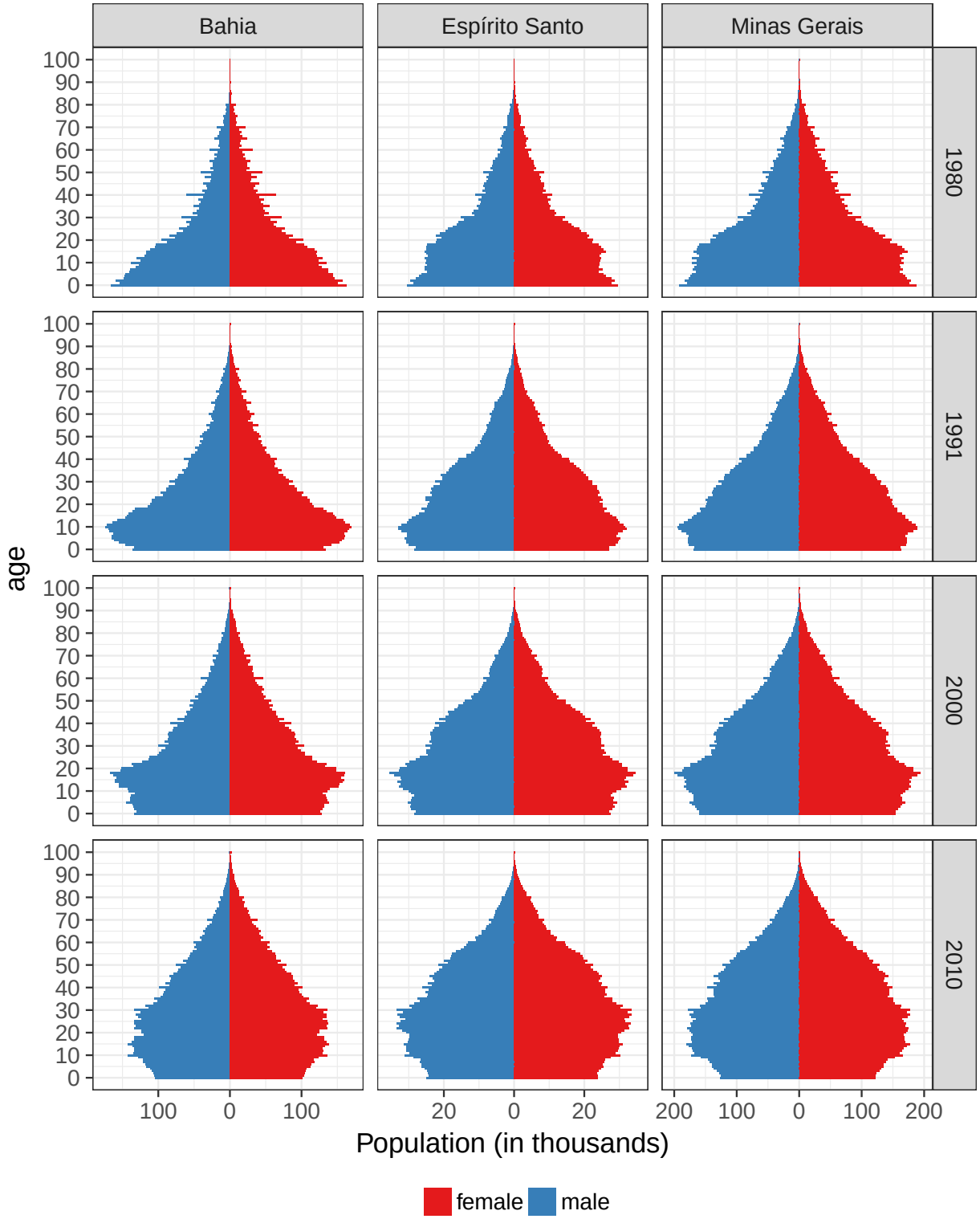

Figure B.6: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

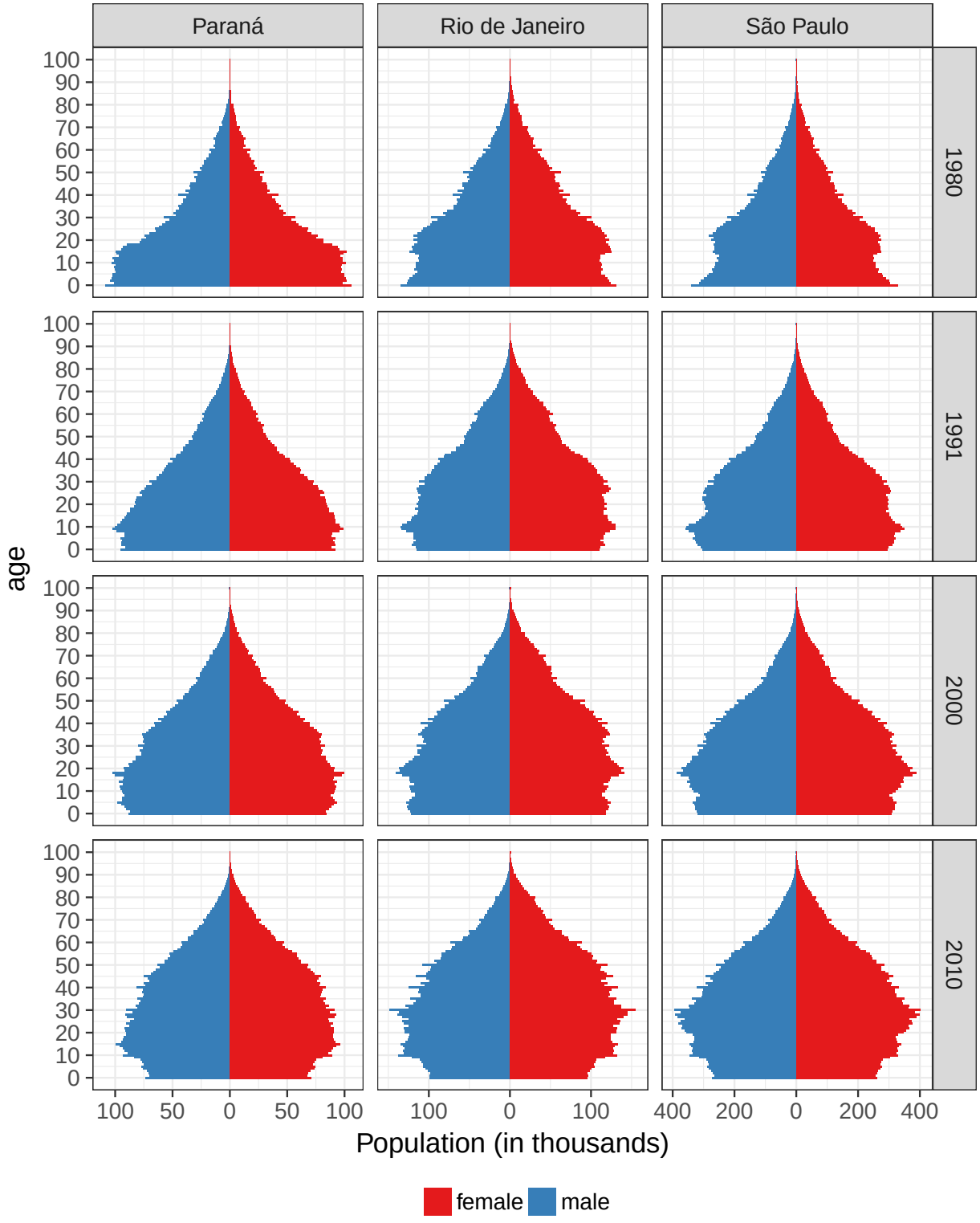

Figure B.7: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

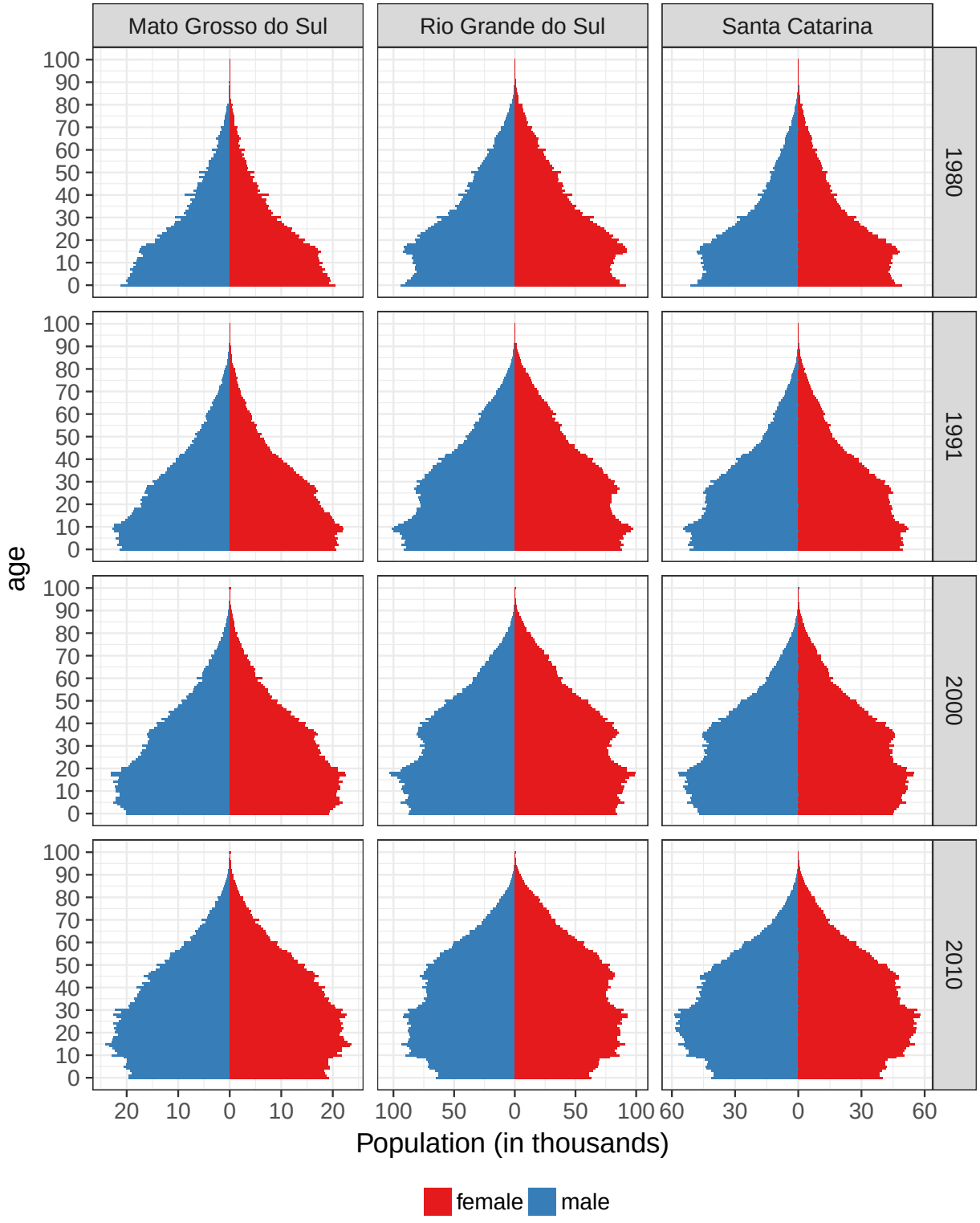

Figure B.8: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

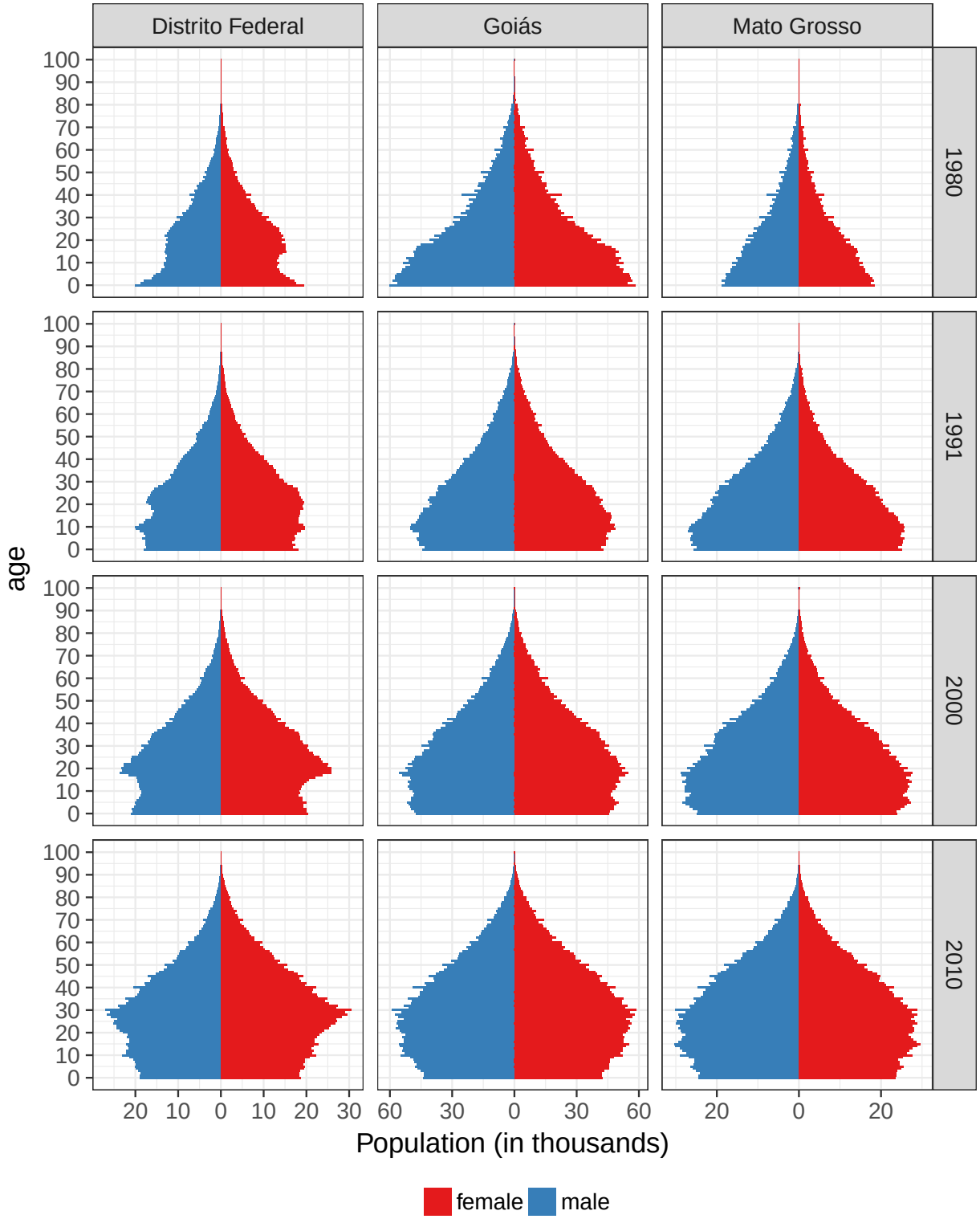

Figure B.9: Population pyramids for selected states, 1980, 1991, 2000, 2010 (in millions). Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

B.2 Indices of digit preference

This section shows results for the Whipple Index and the Myers Index, the most commonly used indices to assess digit preference.

Table B.1: Whipple's Index of preference for digits 0 and 5 (ages 25-65) by sex, Brazil, 1980-2010

Year	Sex	
	male	female
1980	110.45	110.95
1991	103.33	103.27
2000	104.33	104.02
2010	105.58	104.07

Source: IBGE: 1980-2010 Census

Table B.2: Myers' Index of preference for all digits (ages 10-90) by sex, Brazil, 1980-2010

Year	Sex	
	male	female
1980	4.16	4.11
1991	1.35	1.40
2000	2.04	1.78
2010	2.13	1.68

Source: IBGE: 1980-2010 Census

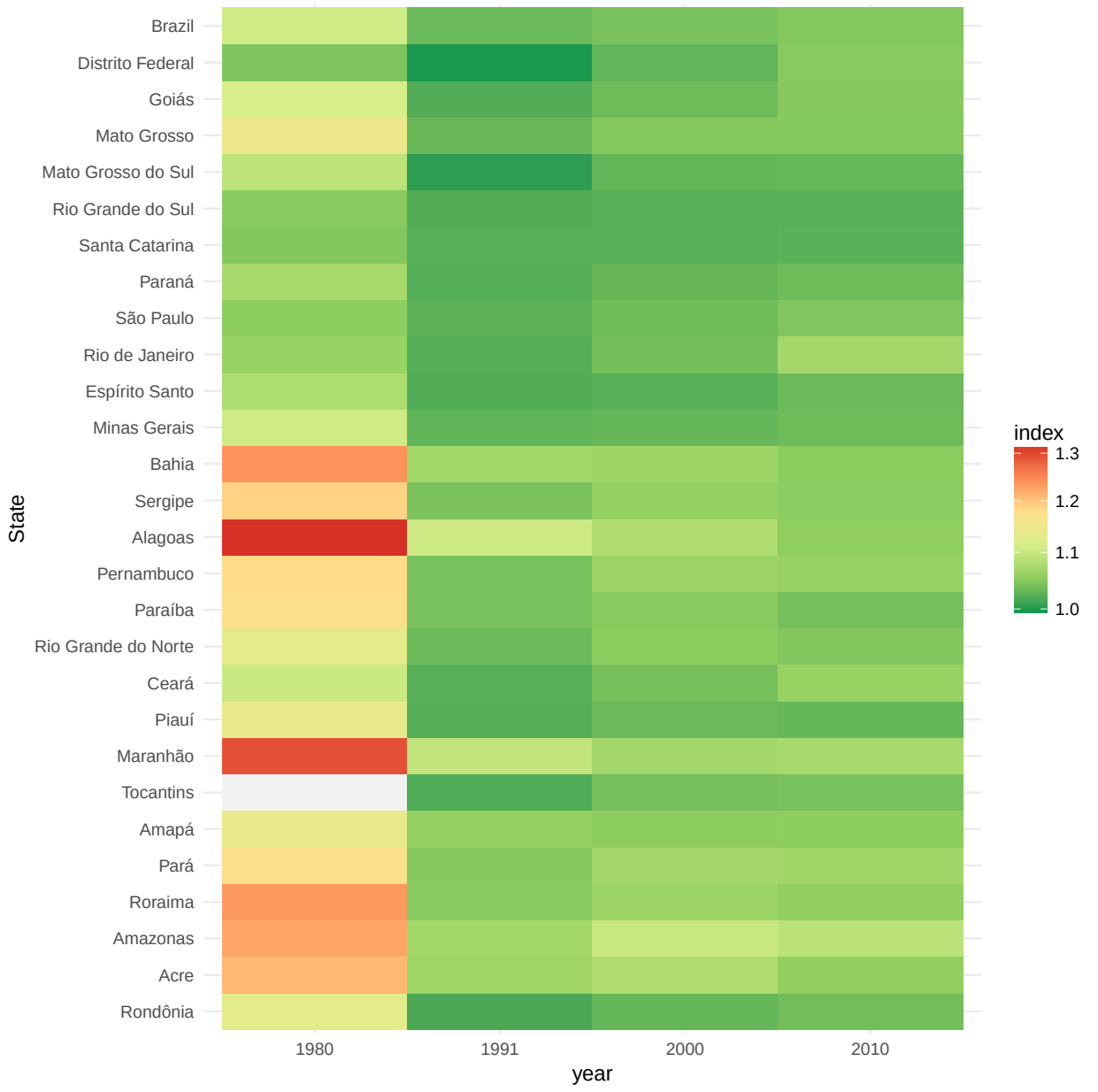

Figure B.10: Heatmap of Whipple's Index of preference for digits 0 and 5 (ages 25-65) by sex, Brazil, 1980-2010. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

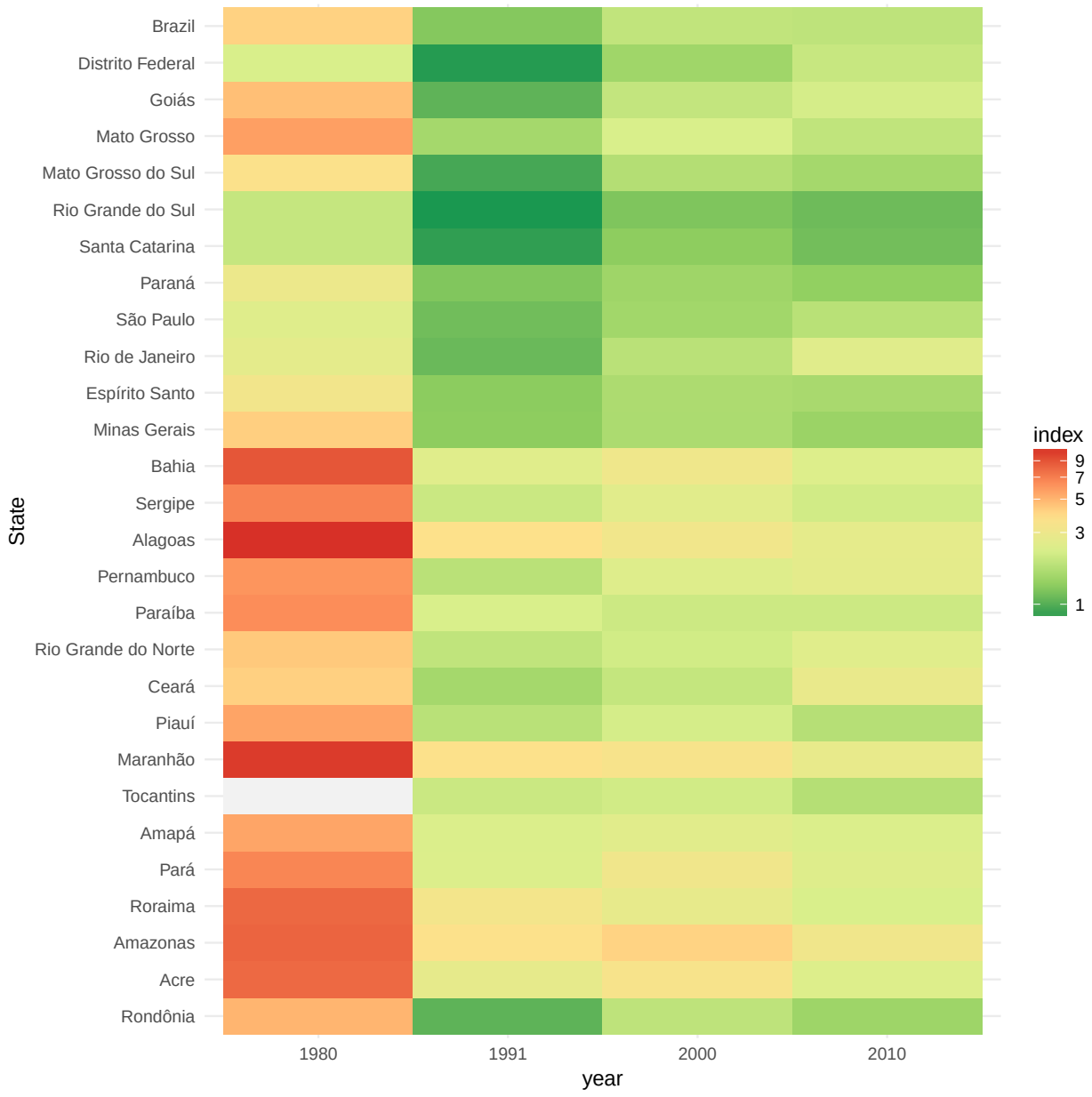

Figure B.11: Heatmap of Myers' Index of preference for all digits (ages 10-90) by sex, Brazil and states, 1980-2010. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

B.3 SR

This section shows the plots of the SR for the 27 Brazilian states for the Censuses 1980, 1991, 2000 and 2010. Scales are fixed for each state.

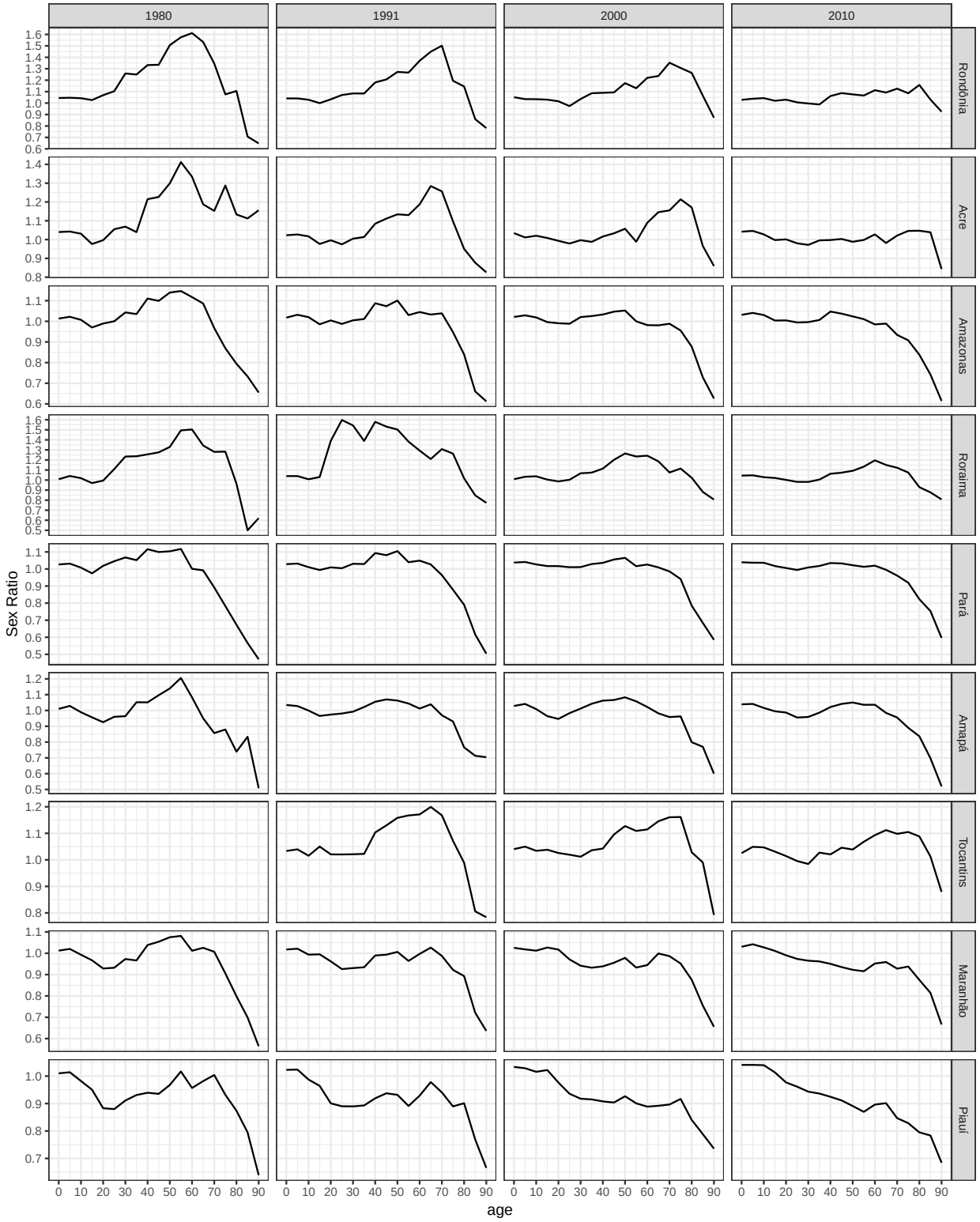

Figure B.12: SR for selected states, 1980, 1991, 2000, 2010. Source: IBGE: 1980-2010 censuses

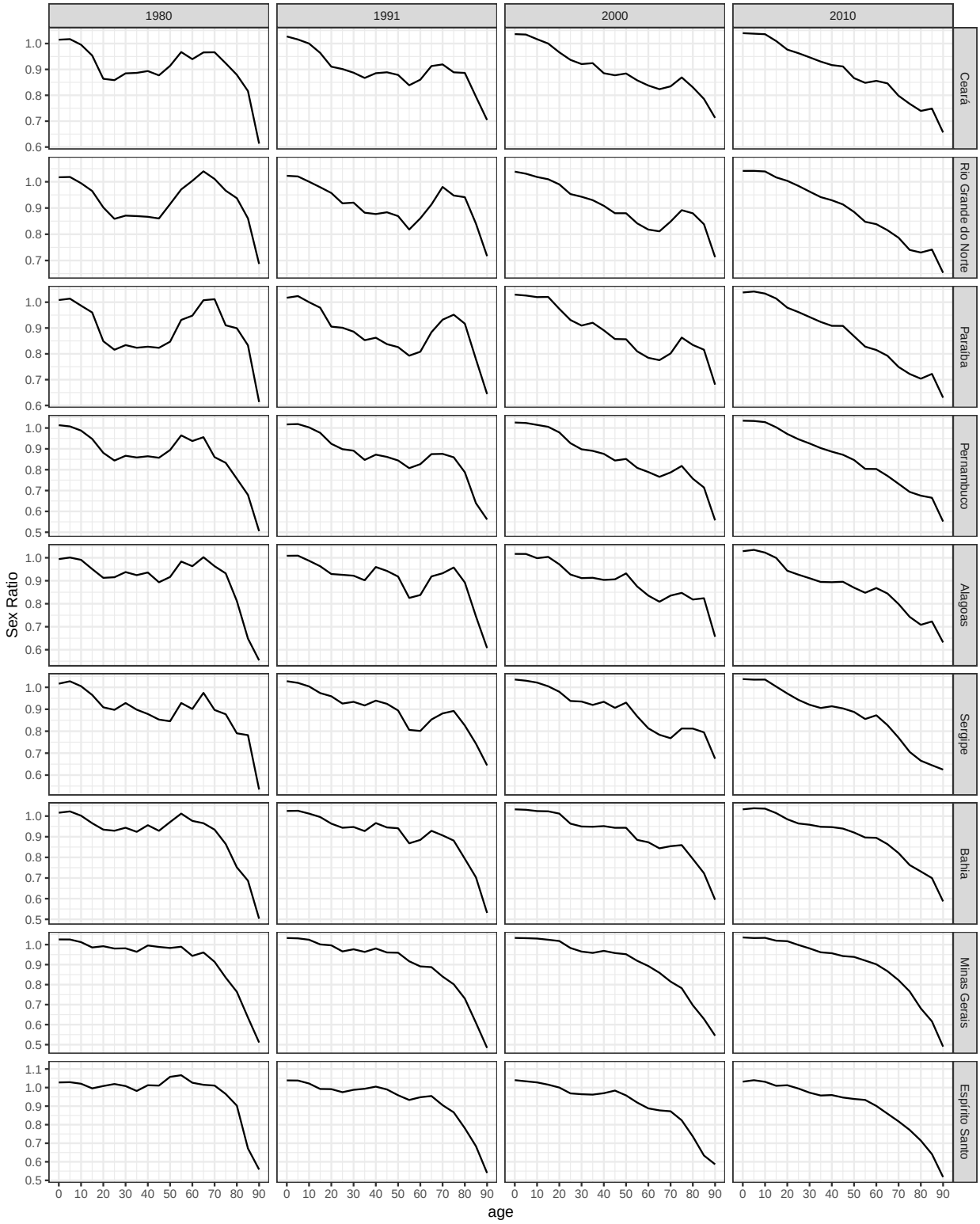

Figure B.13: SR for selected states, 1980, 1991, 2000, 2010. Source: IBGE: 1980-2010 censuses

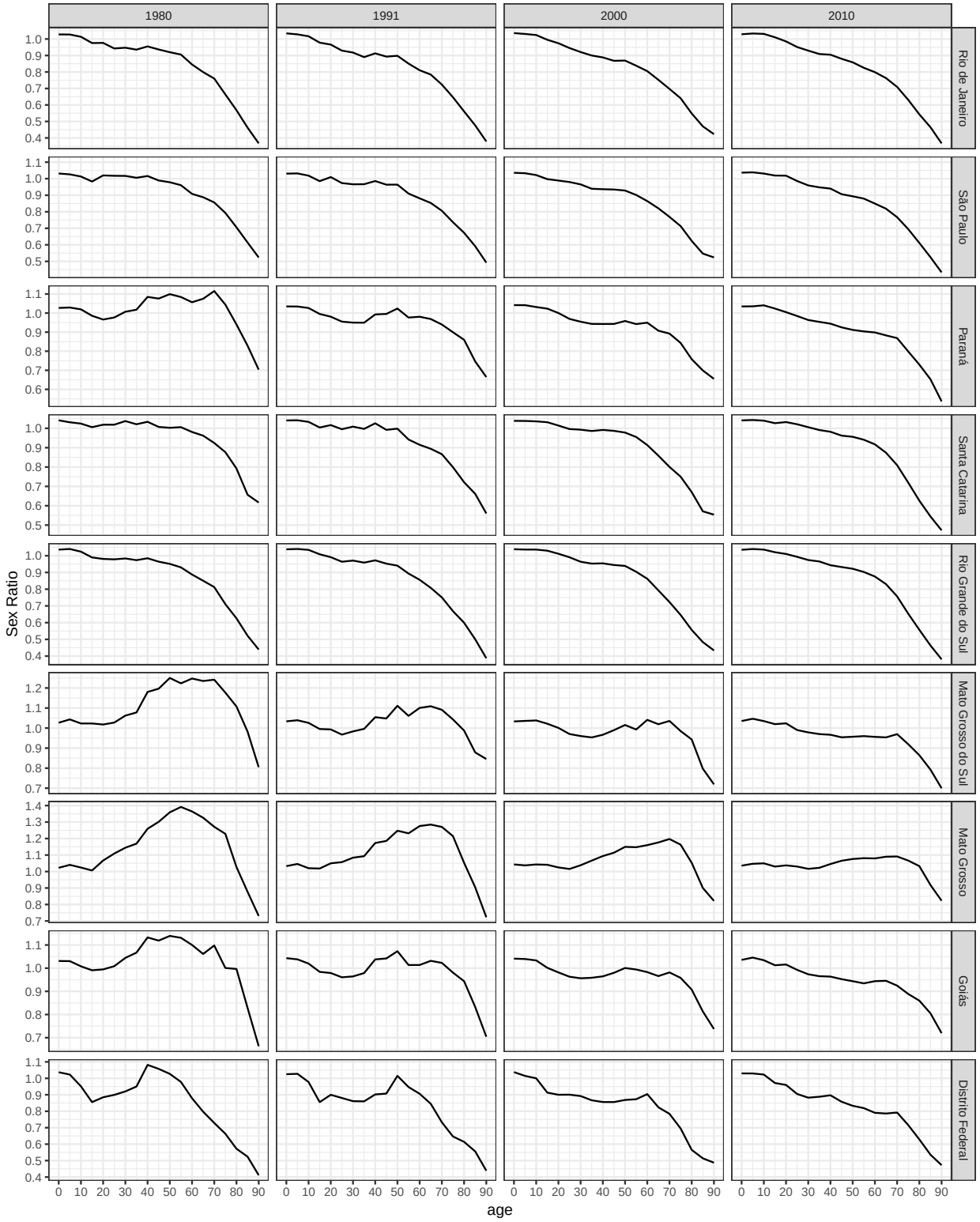

Figure B.14: SR for selected states, 1980, 1991, 2000, 2010. Source: IBGE: 1980-2010 censuses

B.4 CSR

Figure B.15 shows the CSR by sex for the three intercensal periods between 1980 and 2010

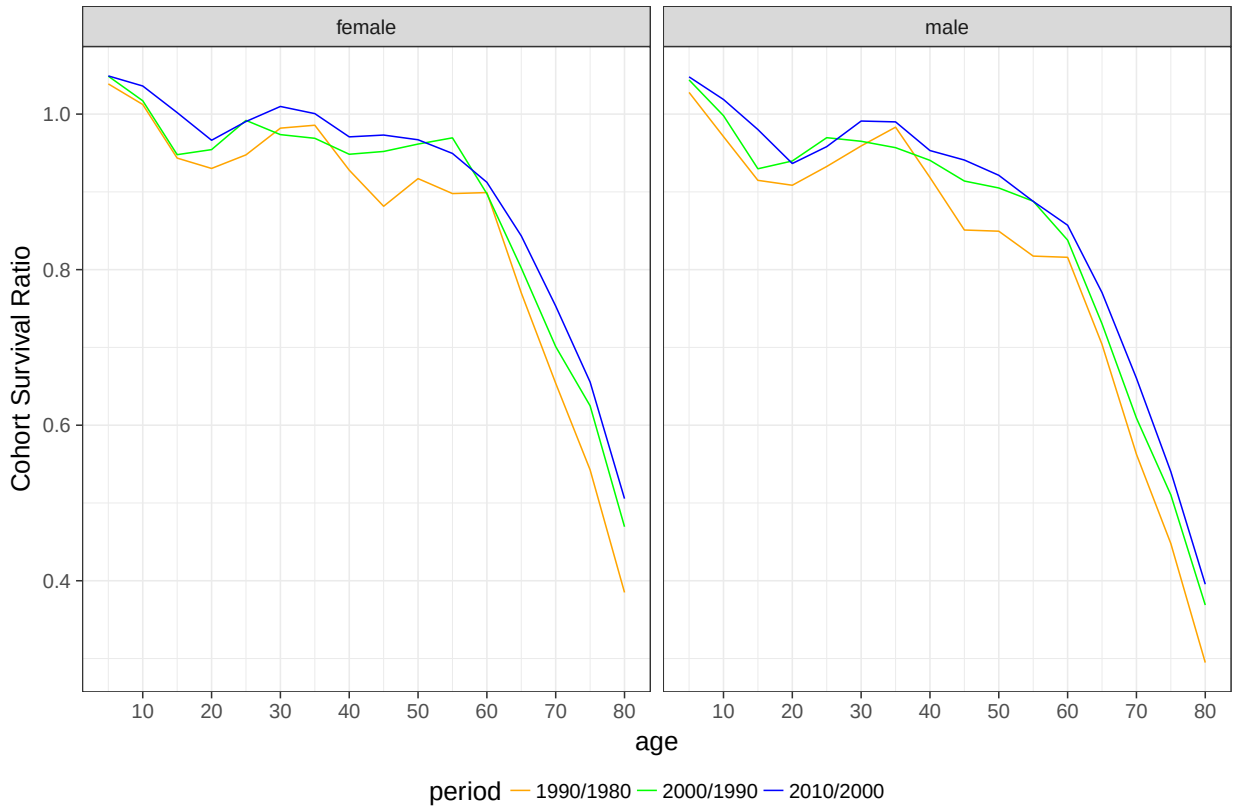

Figure B.15: CSR by sex, Brazil, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000 and 2010

Plots below show the CSR for the 27 Brazilian states for the intercensal periods 1980/1990, 1990/2000 and 2000/2010. Scales are fixed for each state.

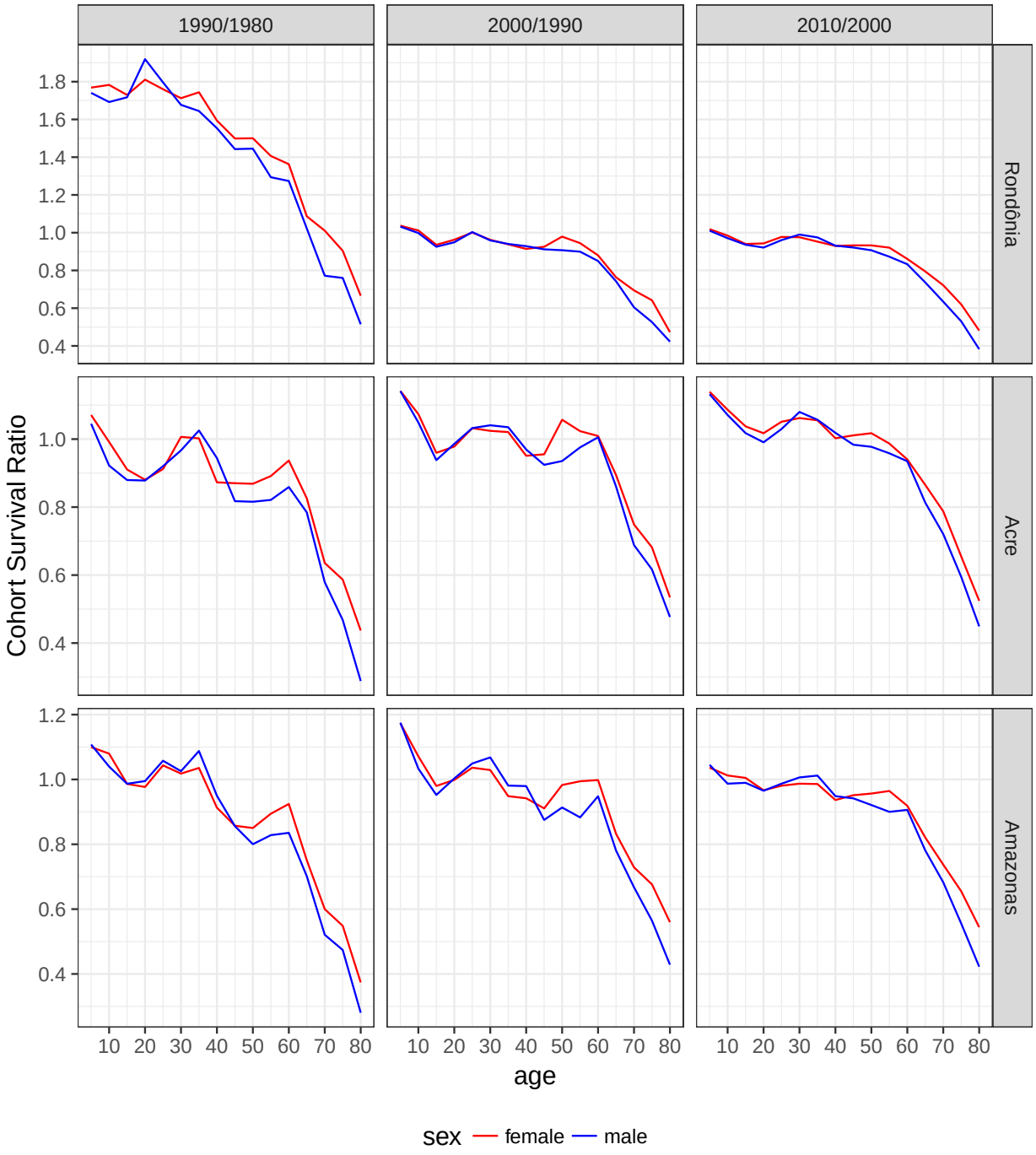

Figure B.16: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

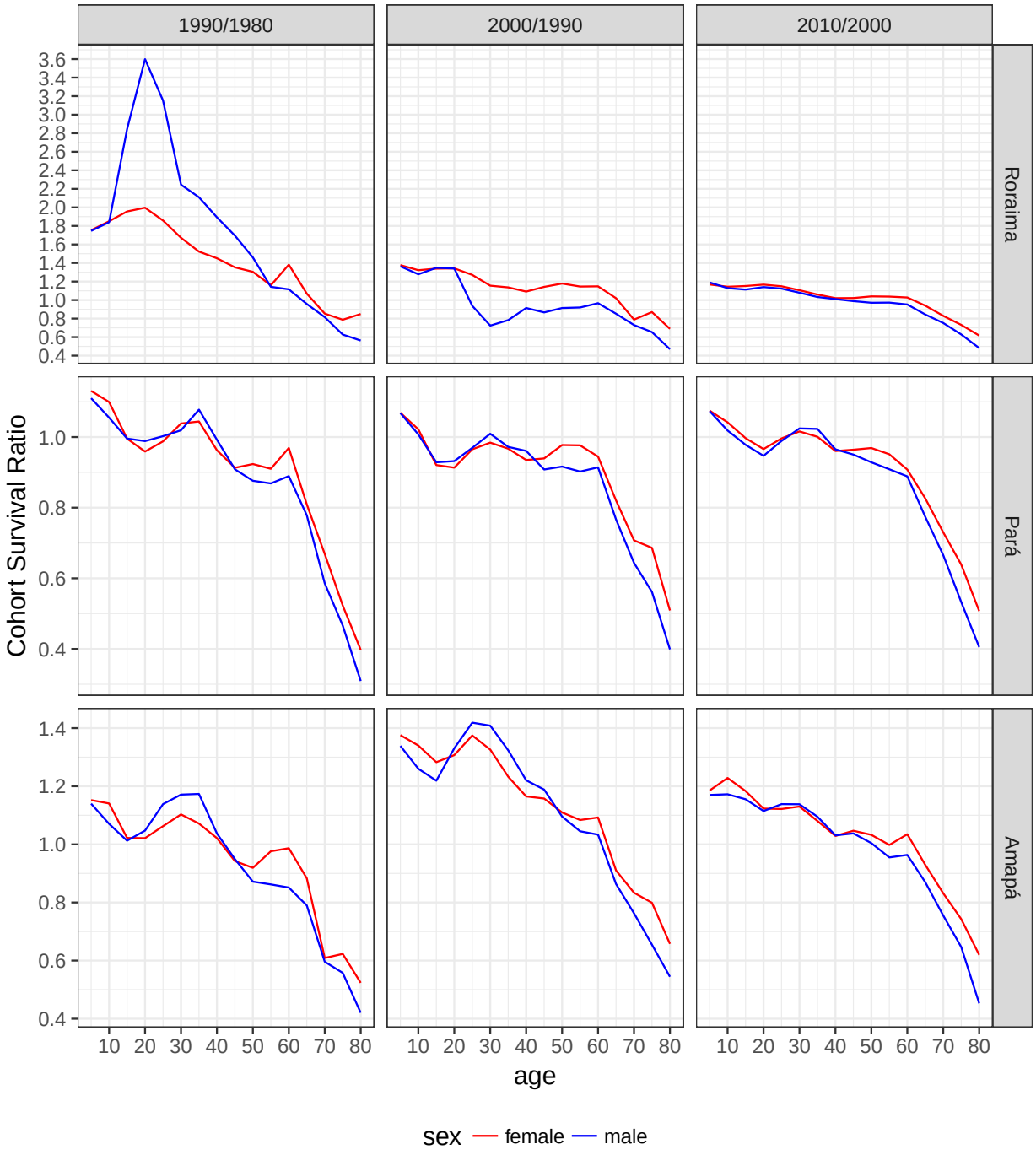

Figure B.17: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

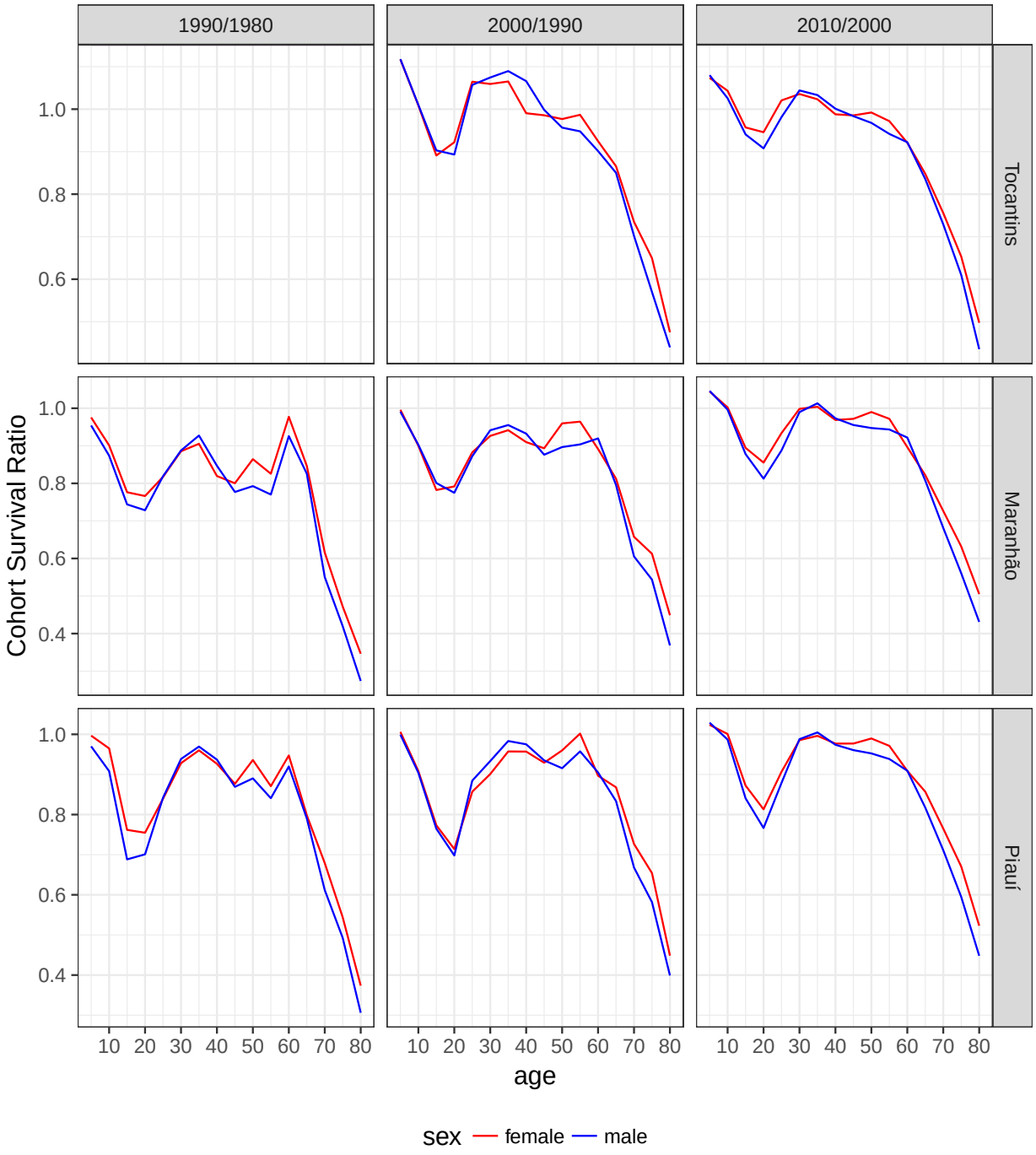

Figure B.18: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

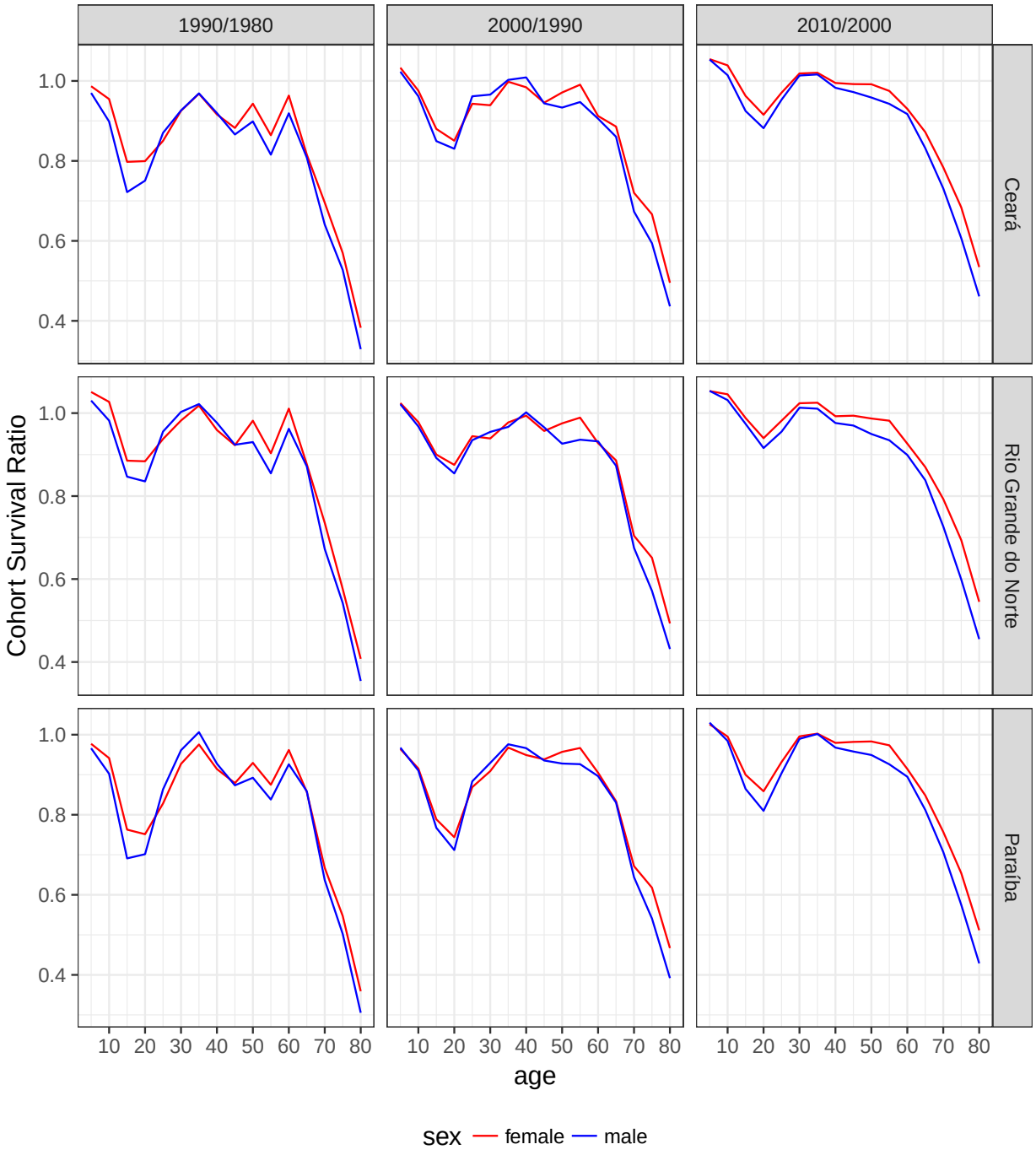

Figure B.19: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

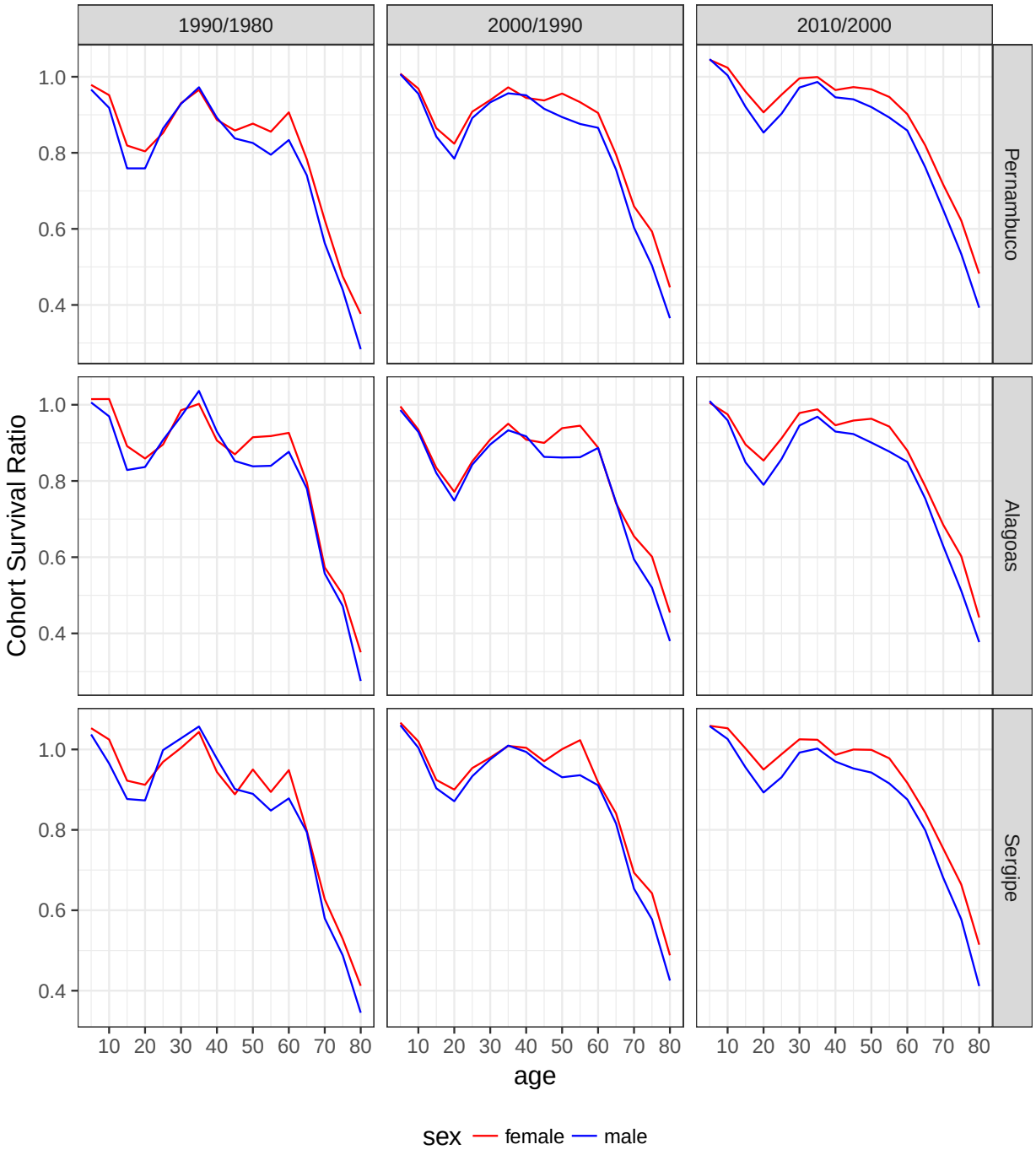

Figure B.20: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

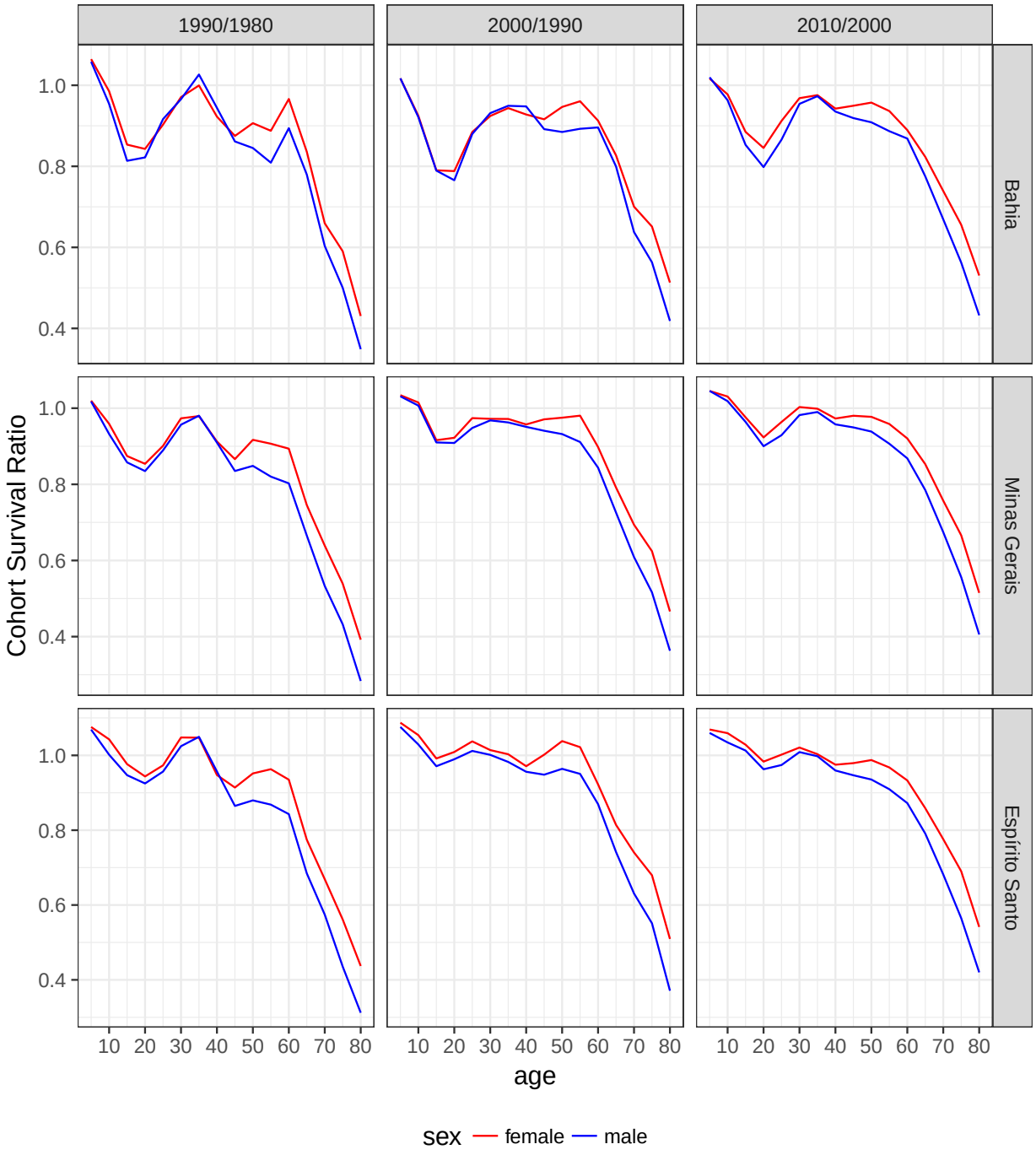

Figure B.21: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

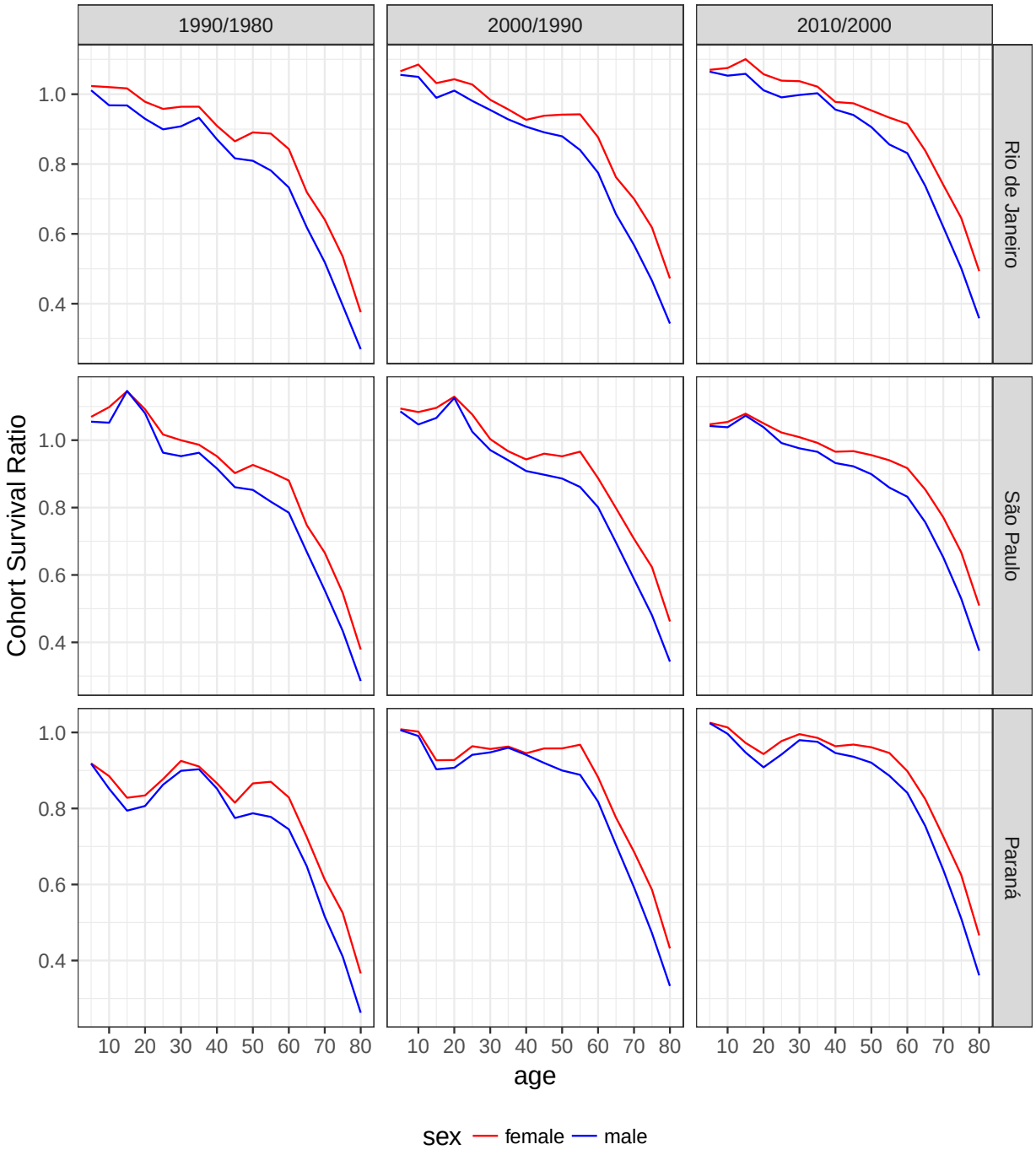

Figure B.22: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

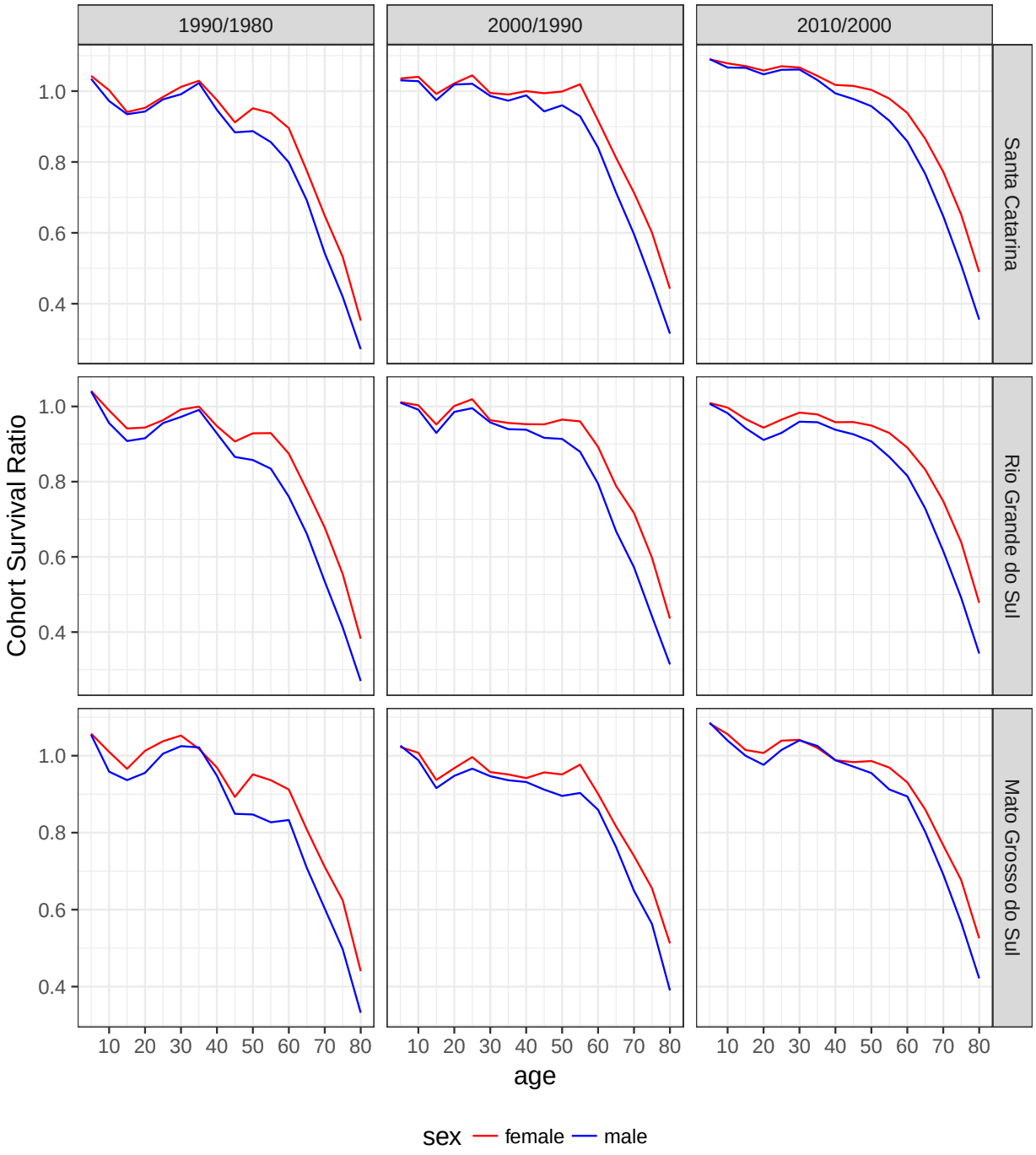

Figure B.23: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

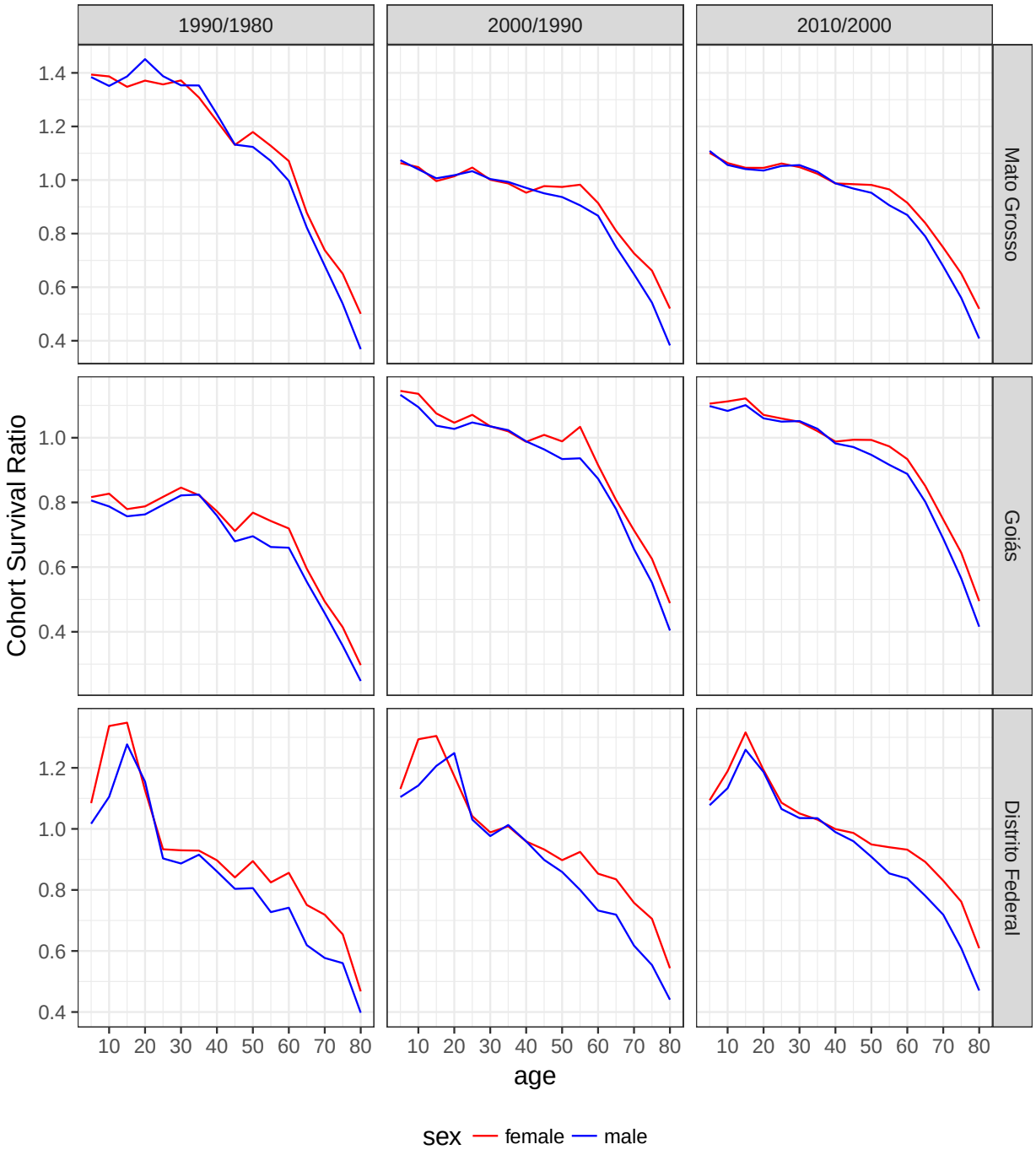

Figure B.24: CSR by sex and state, intercensal periods 1980/1990, 1990/2000 and 2000/2010 and age group at the middle of the intercensal period. Source: IBGE, Brazilian Censuses of 1980, 1991, 2000, 2010

Appendix C

Internal Migration in Brazil from 1980 to 2010

C.1 Internal Migration

This section shows the plots of the migration rates (in-migration, out-migration and net migration) for the 1991, 2000 and 2010 censuses by age and sex for the 27 Brazilian states. Scales are fixed for each state.

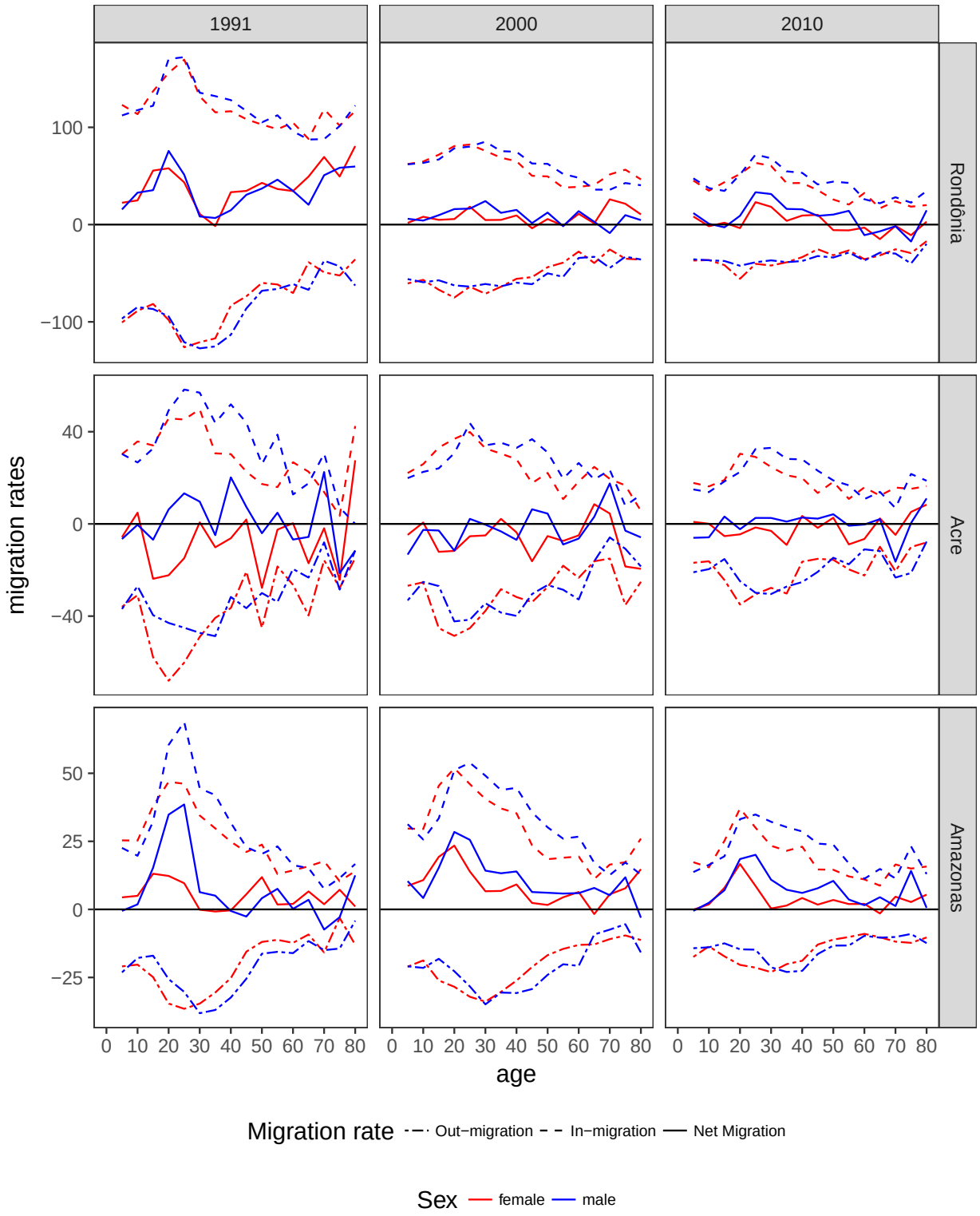

Figure C.1: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

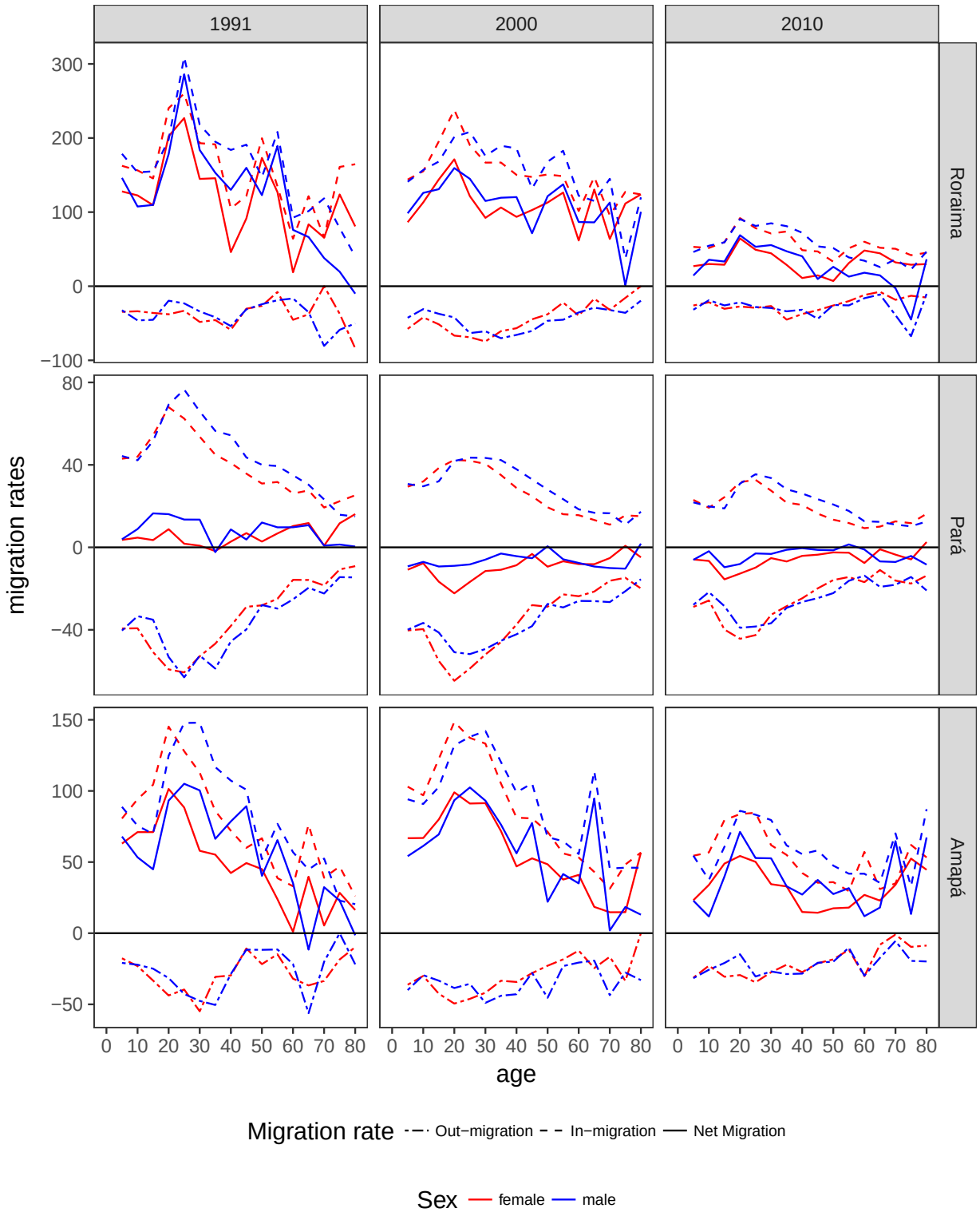

Figure C.2: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

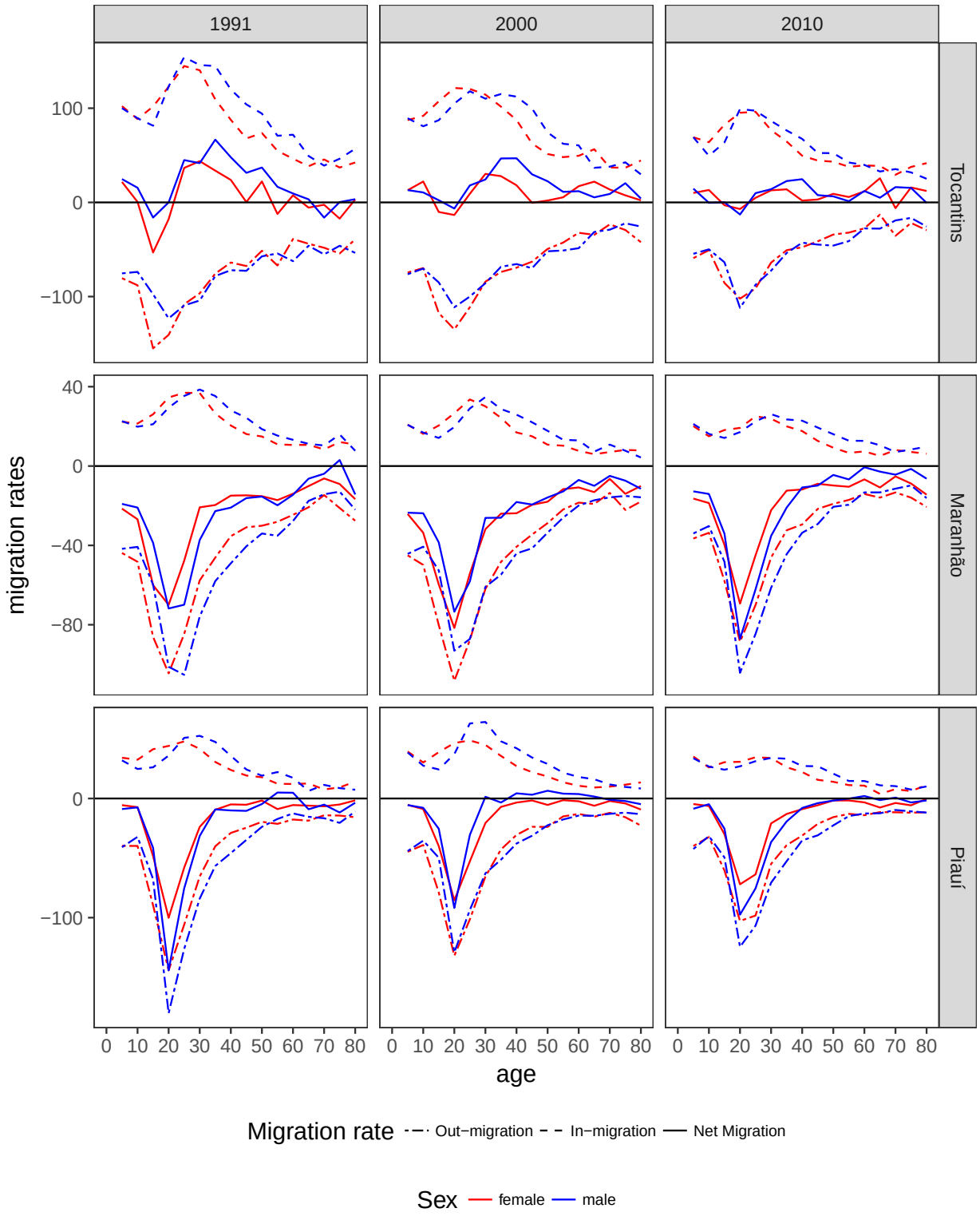

Figure C.3: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

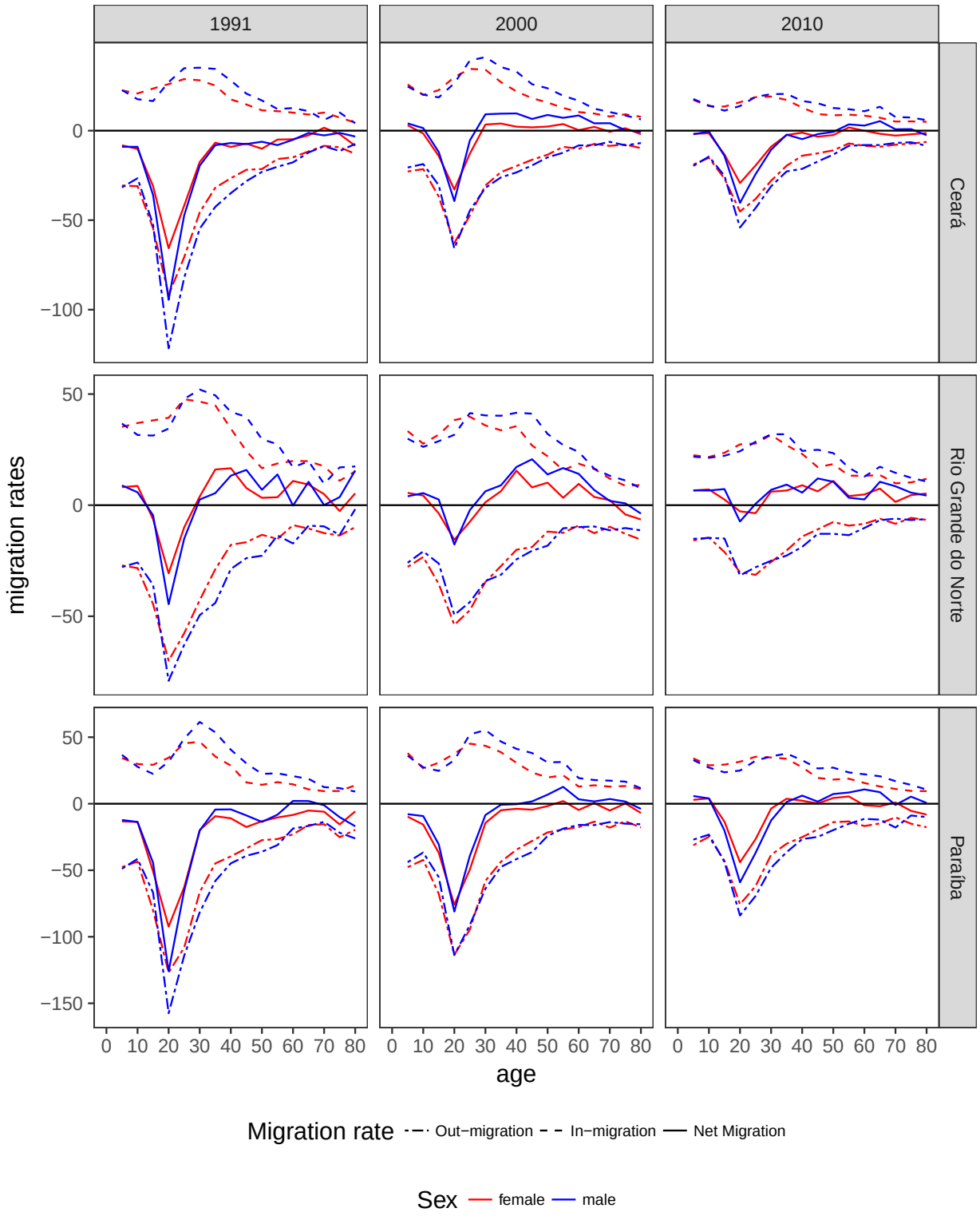

Figure C.4: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

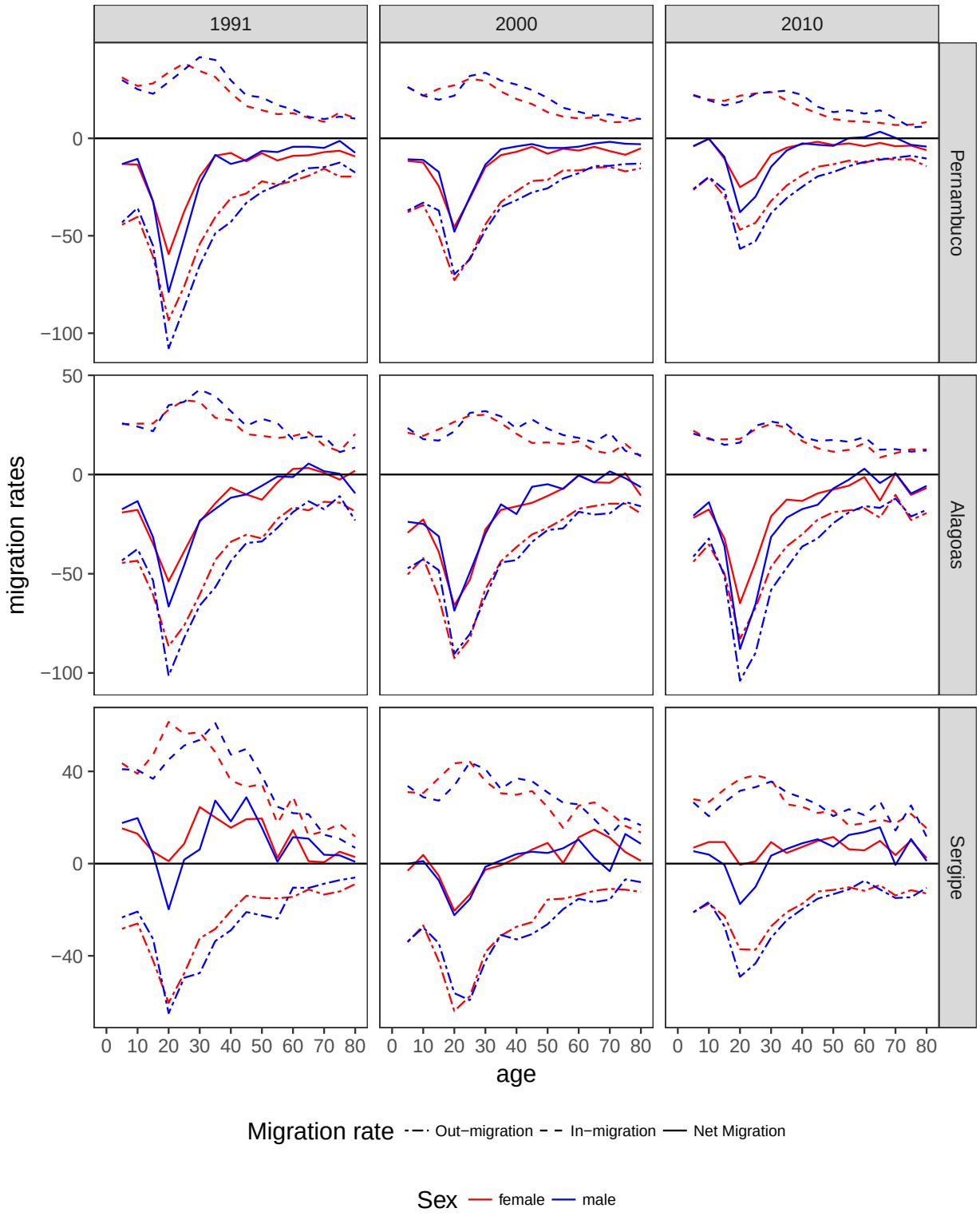

Figure C.5: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

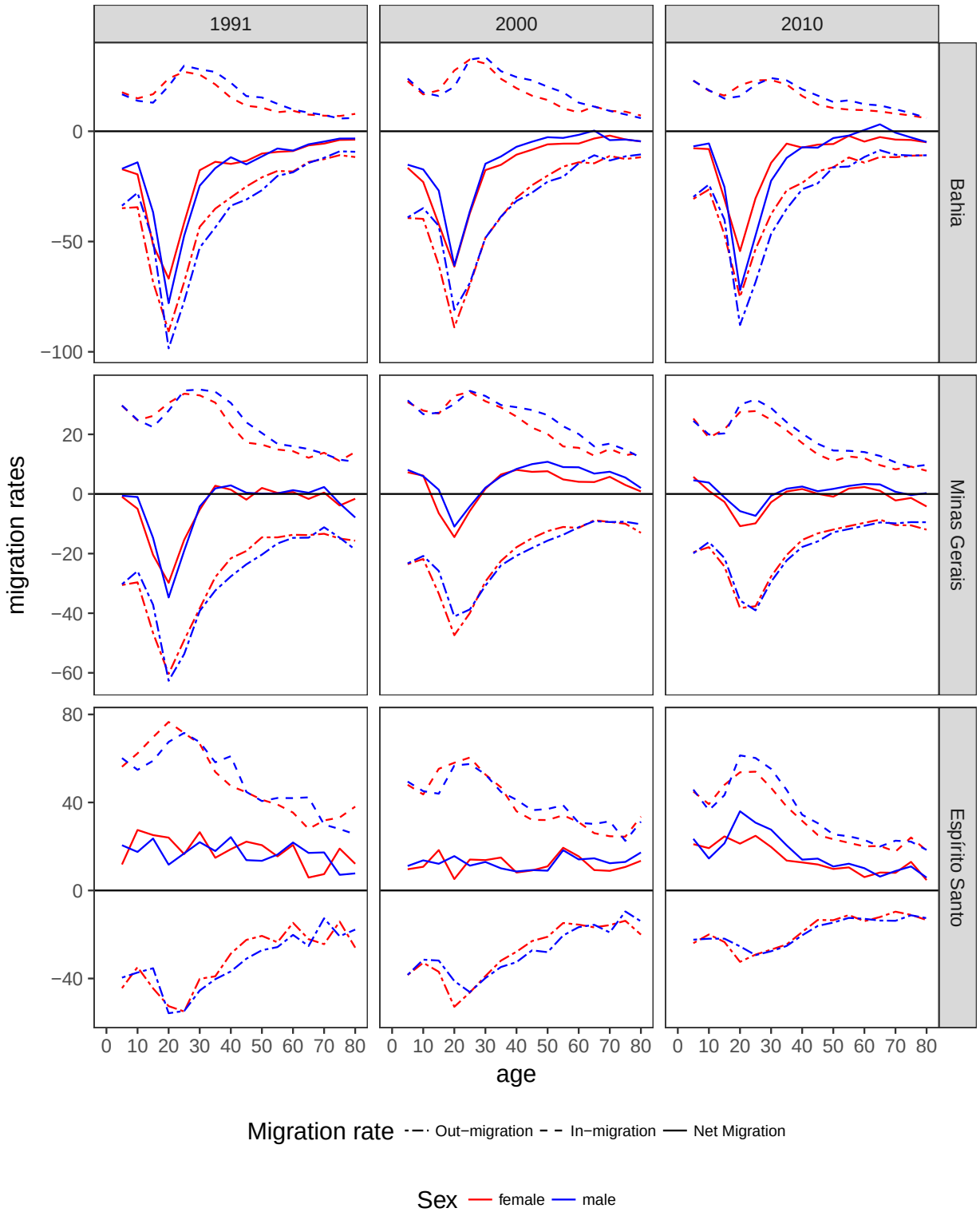

Figure C.6: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

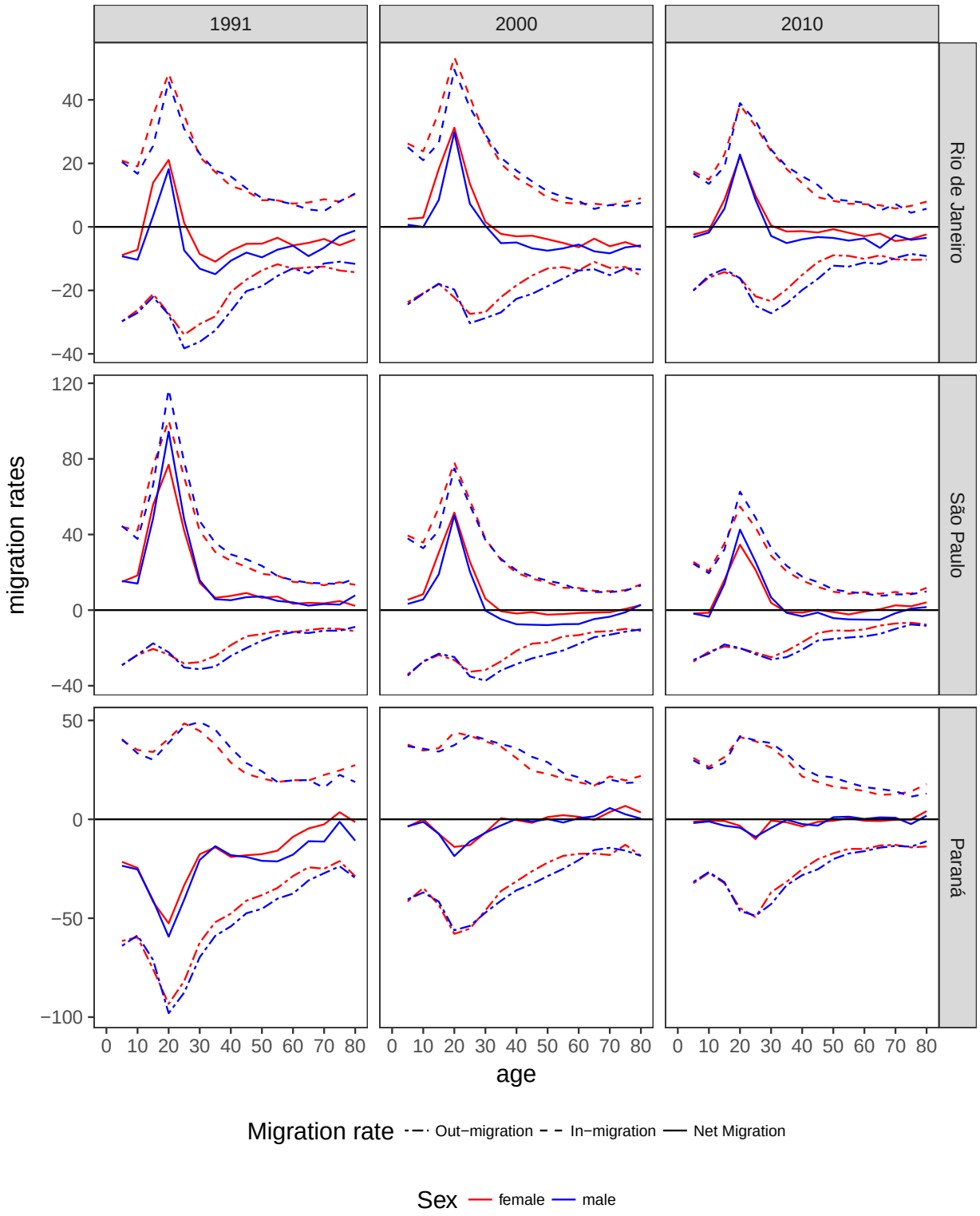

Figure C.7: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

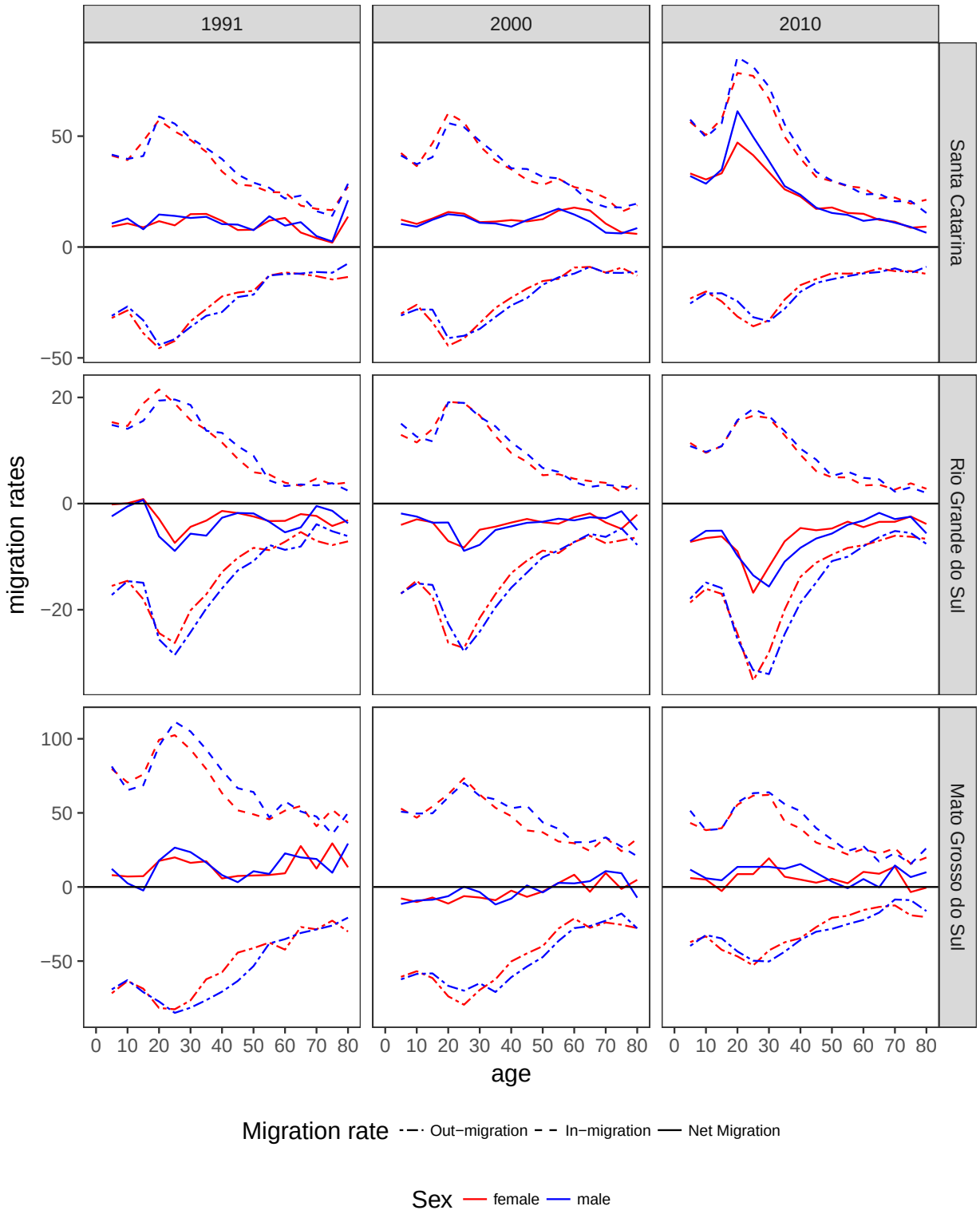

Figure C.8: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

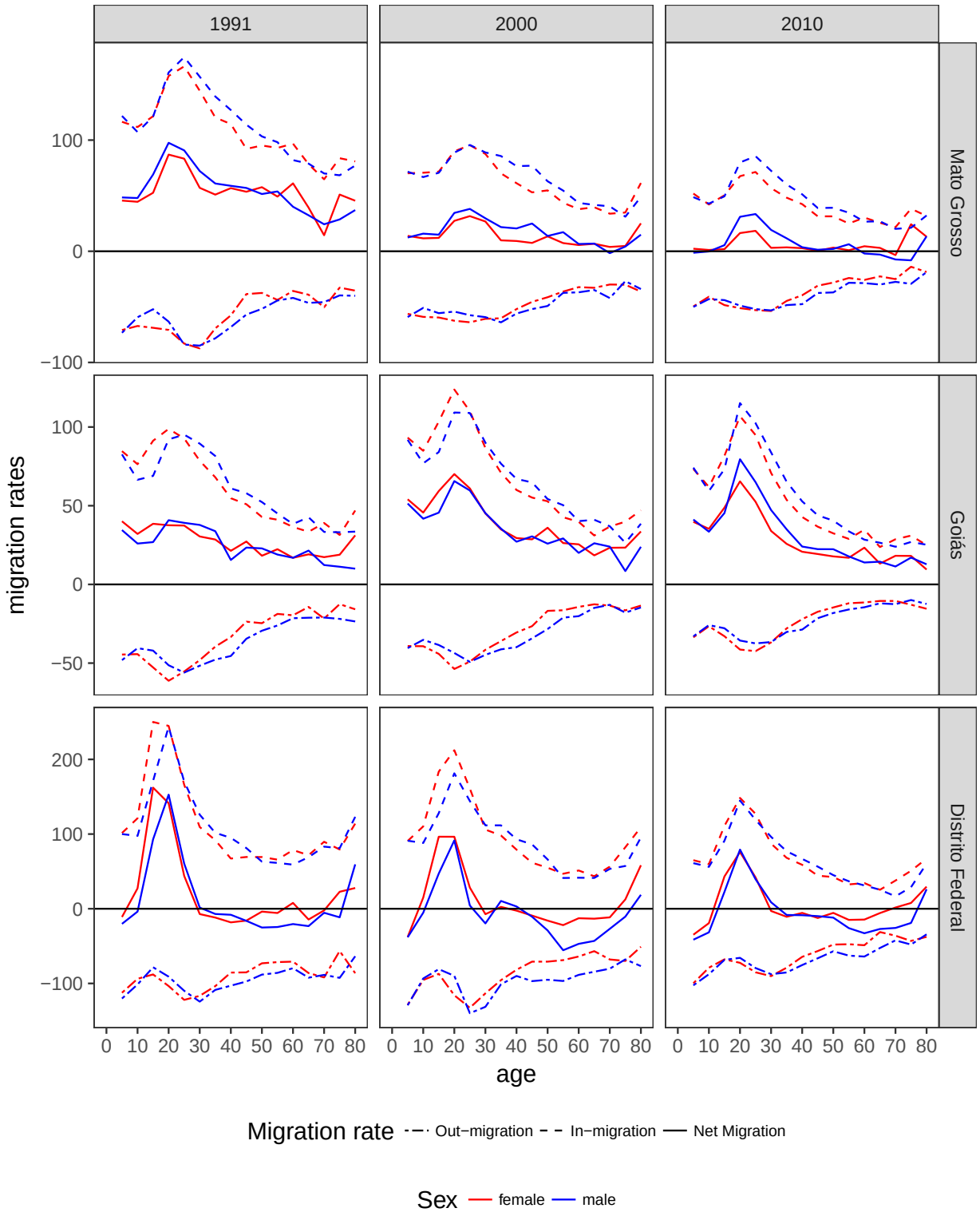

Figure C.9: in-migration, out-migration and net migration rates for selected states, 1991, 2000, 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

Appendix D

Mortality Estimates for Brazil and States from 1980 to 2010

D.1 Population Pyramids for deaths count

This section shows the plots of the population pyramids for the death counts by single year of age for the 27 Brazilian states in absolute numbers. Scales are fixed for each state.

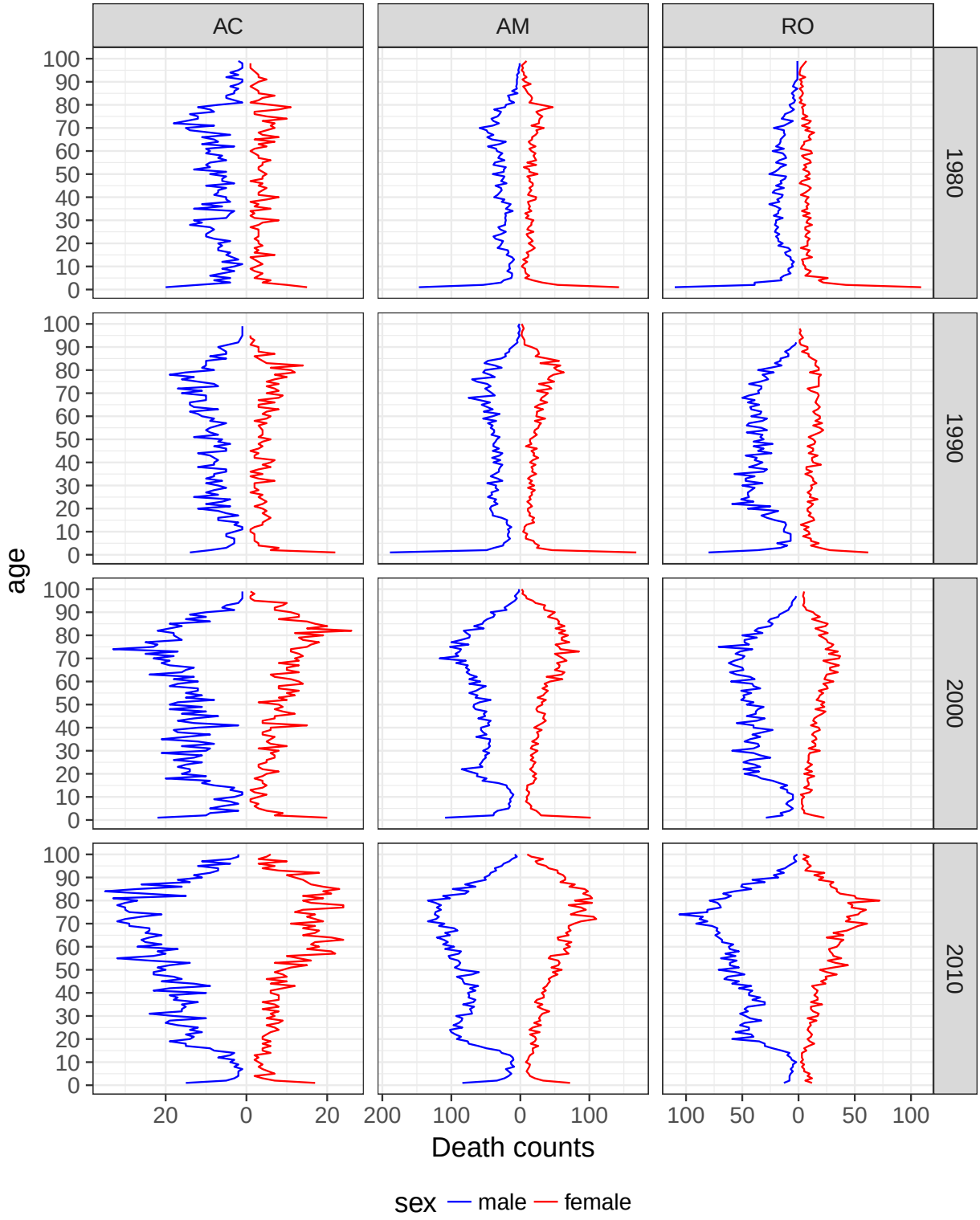

Figure D.1: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

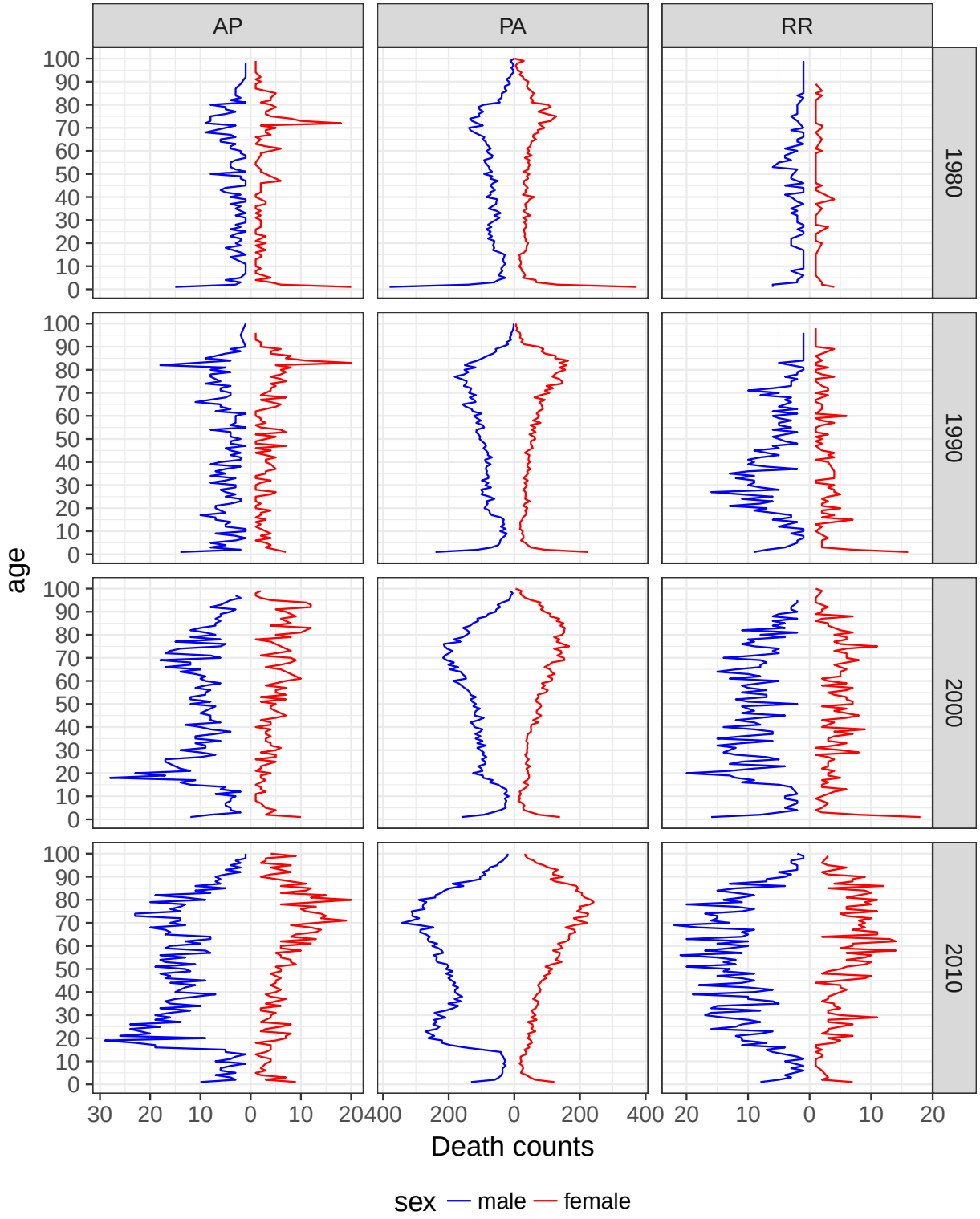

Figure D.2: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

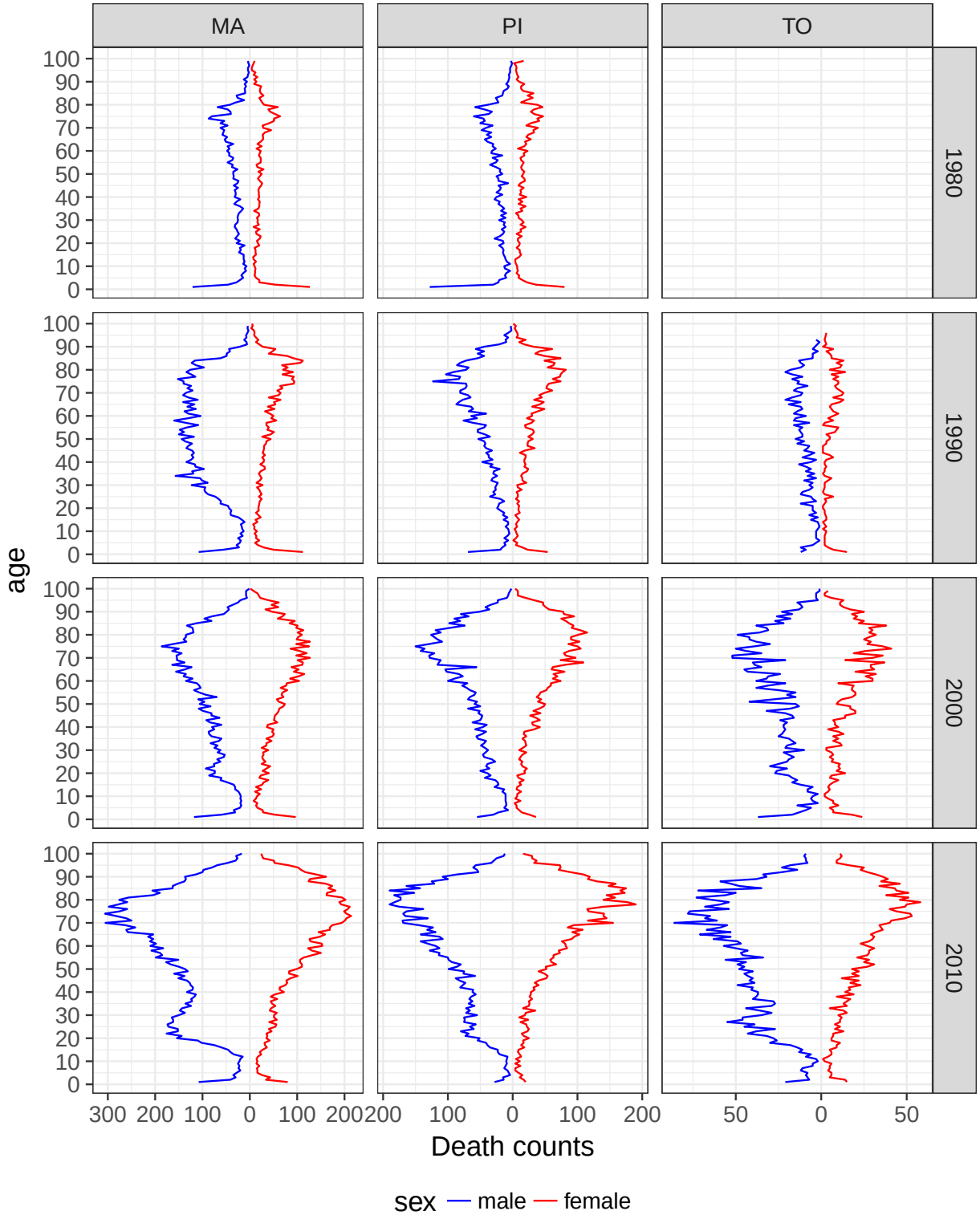

Figure D.3: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

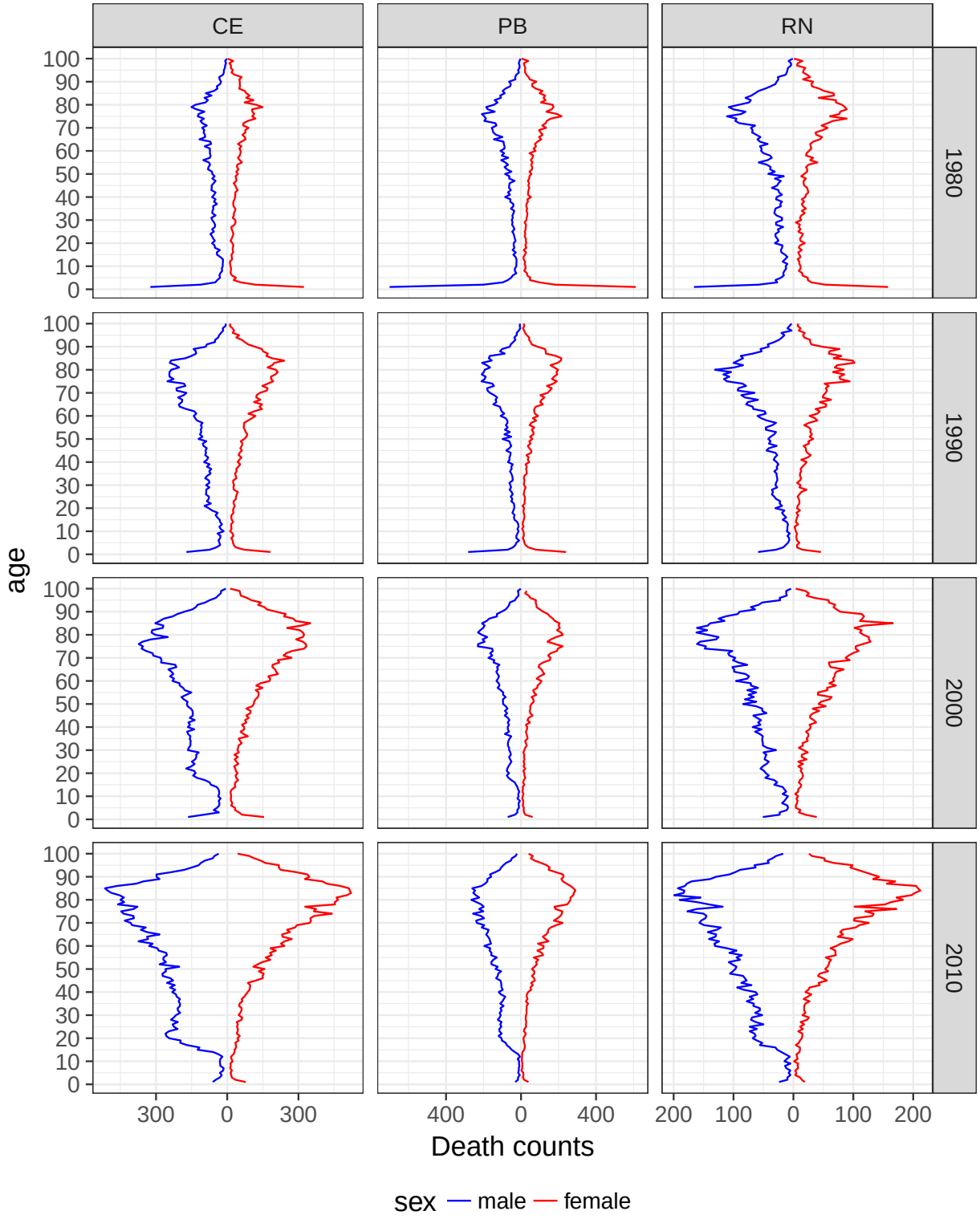

Figure D.4: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

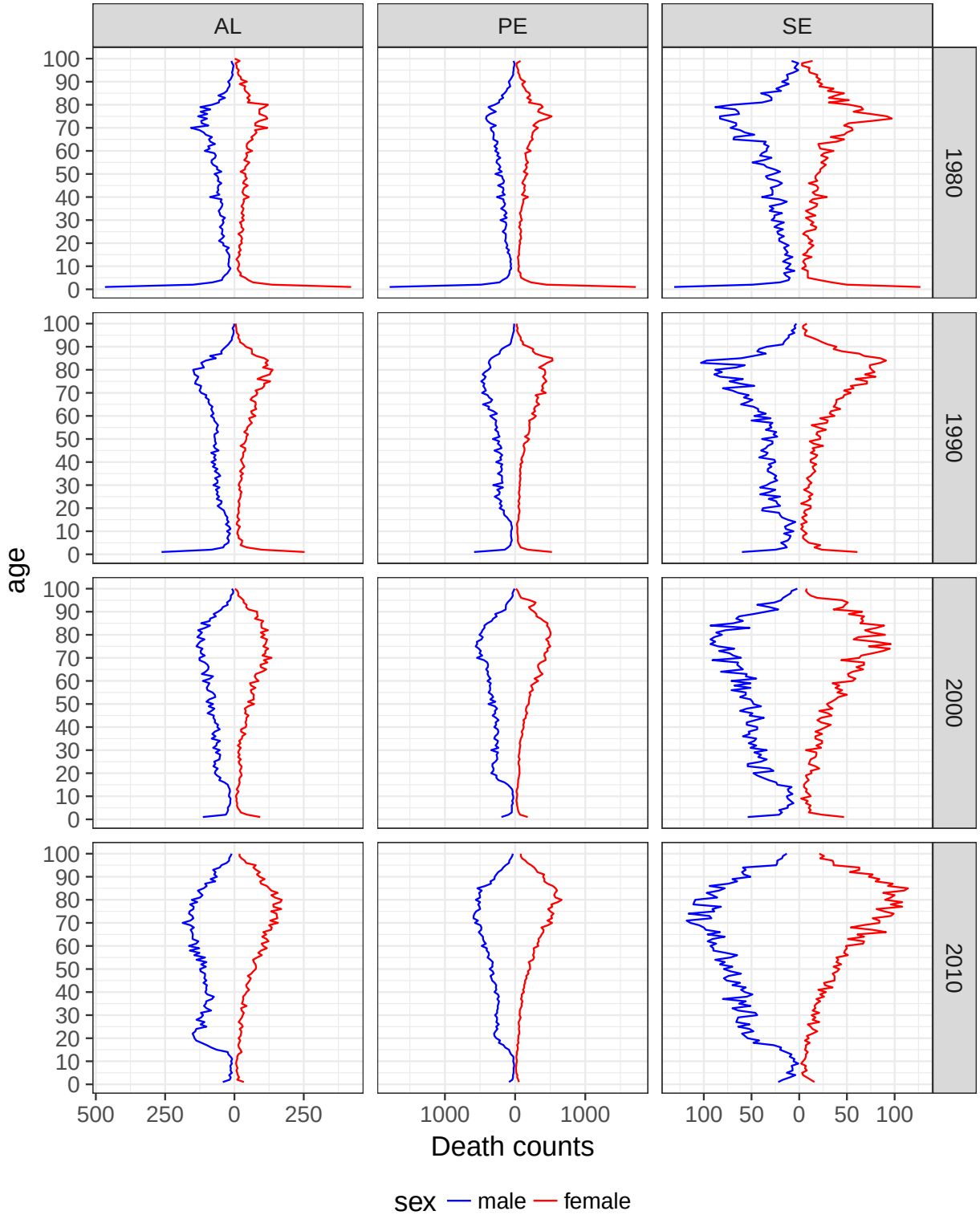

Figure D.5: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

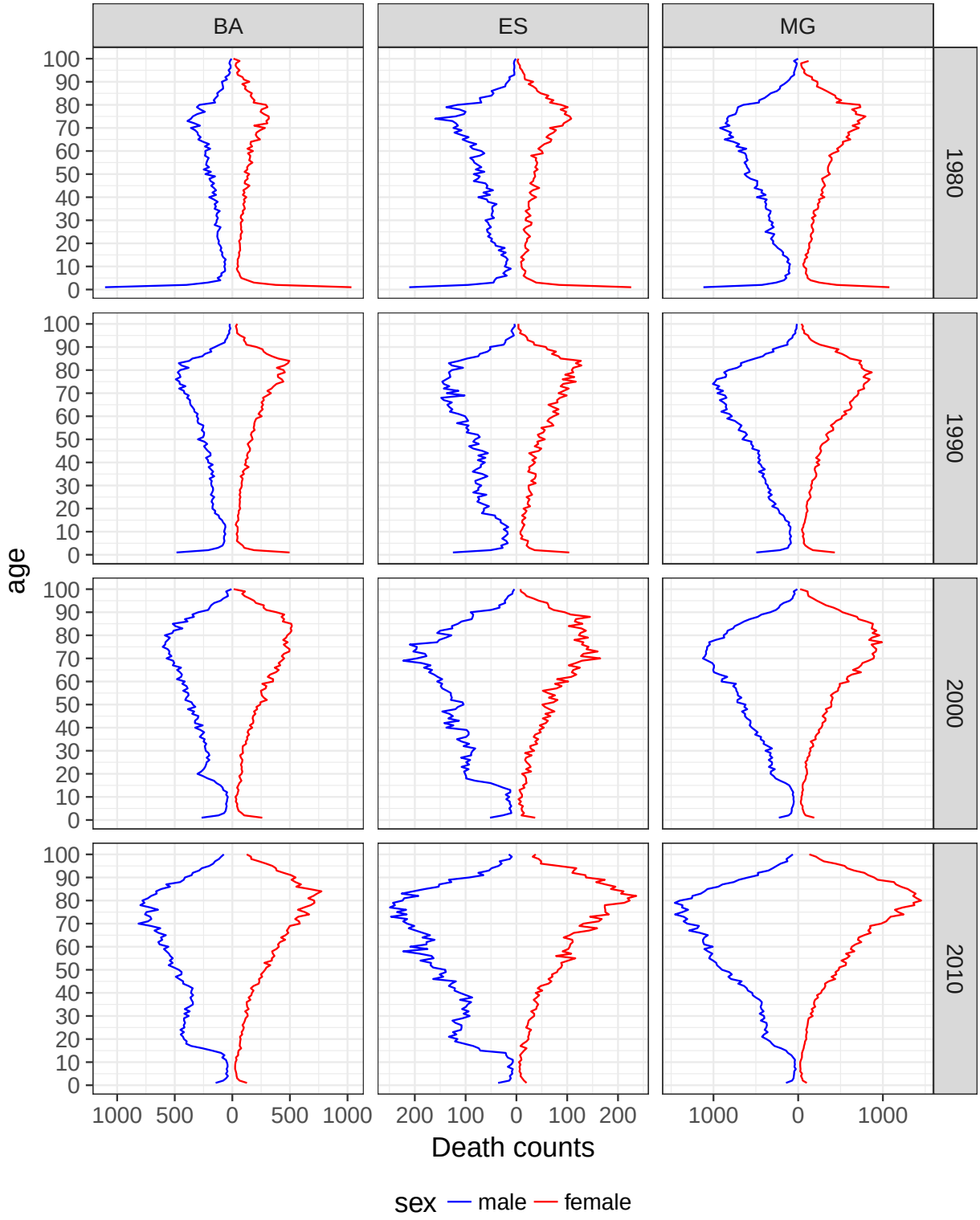

Figure D.6: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

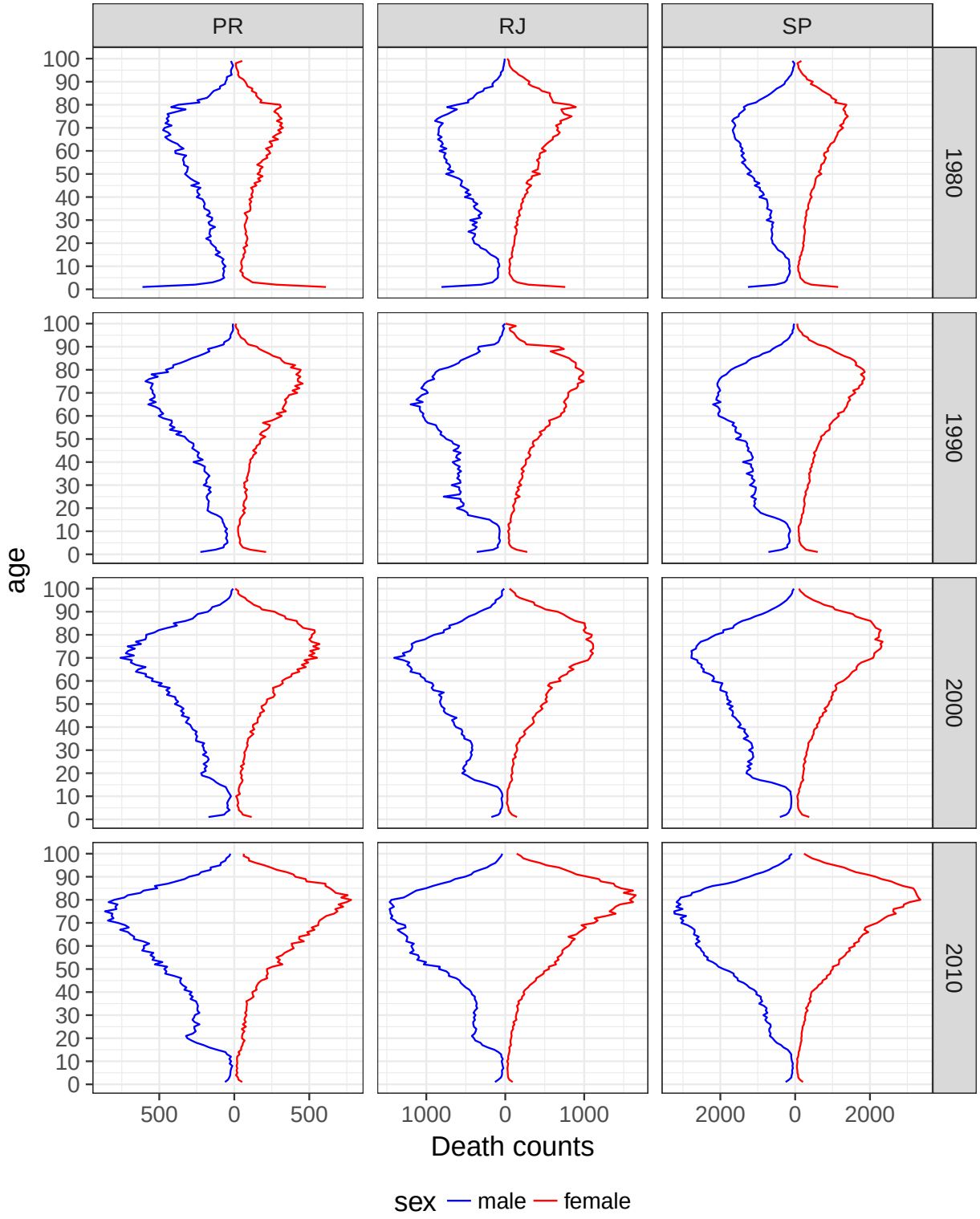

Figure D.7: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

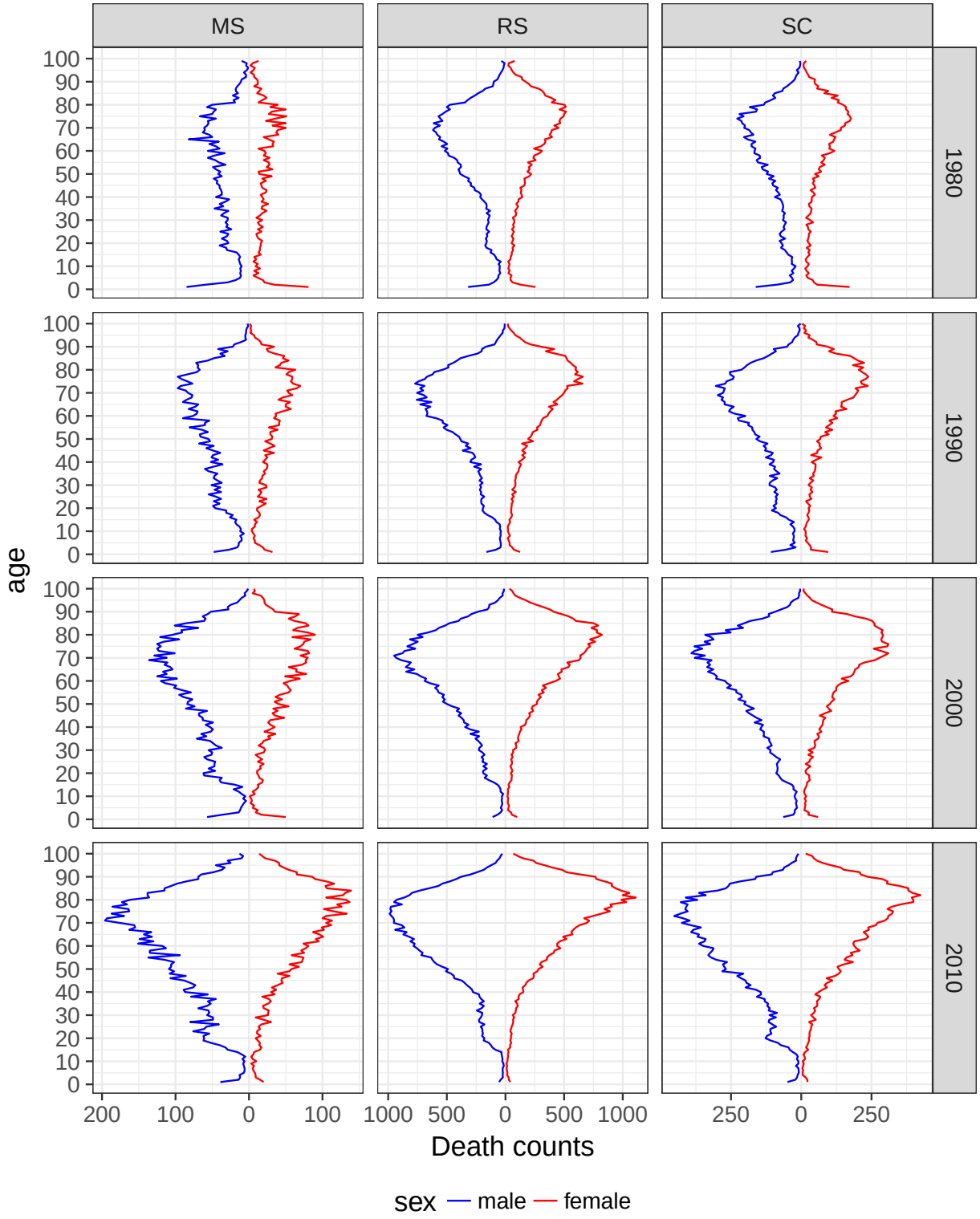

Figure D.8: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

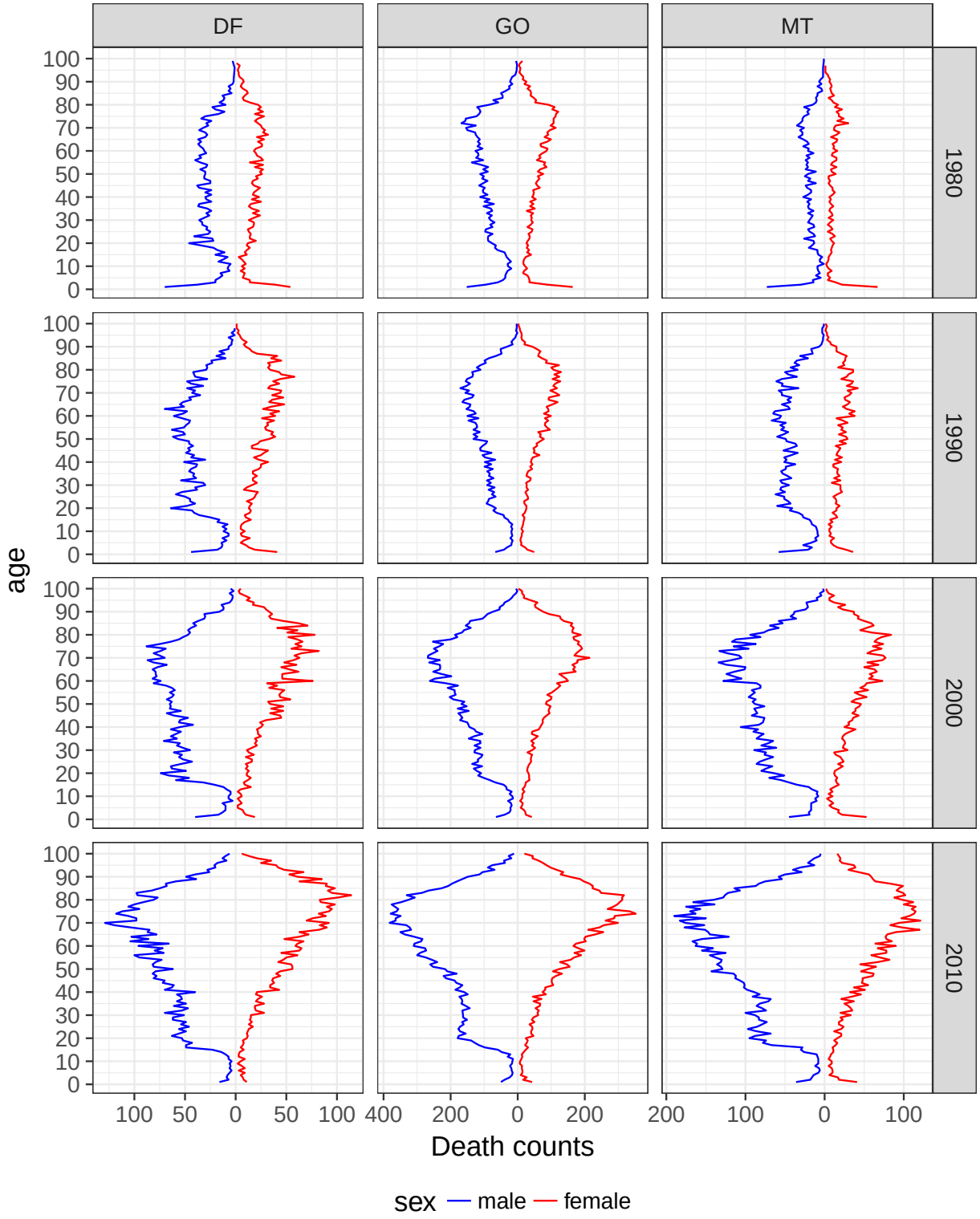

Figure D.9: Population pyramids for death counts for selected states, 1980, 1990, 2000, 2010. Source: SIM 1980, 1990, 2000, 2010

Appendix E

Case Study from Brazil

E.1 Consistency in demographic data in Brazilian states

This section shows the plots of the enumerated, projected and backprojected populations for 1990, 2000 and 2010 for the 27 Brazilian states. Scales are fixed for each state.

E.2 Internal Migration

This section shows the plots of the migration rates (in-migration and out-migration) estimated for the five-year period prior to the years 1990, 1995, 2000, 2005 and 2010 by age and sex for the 27 Brazilian states. Scales are fixed for each state.

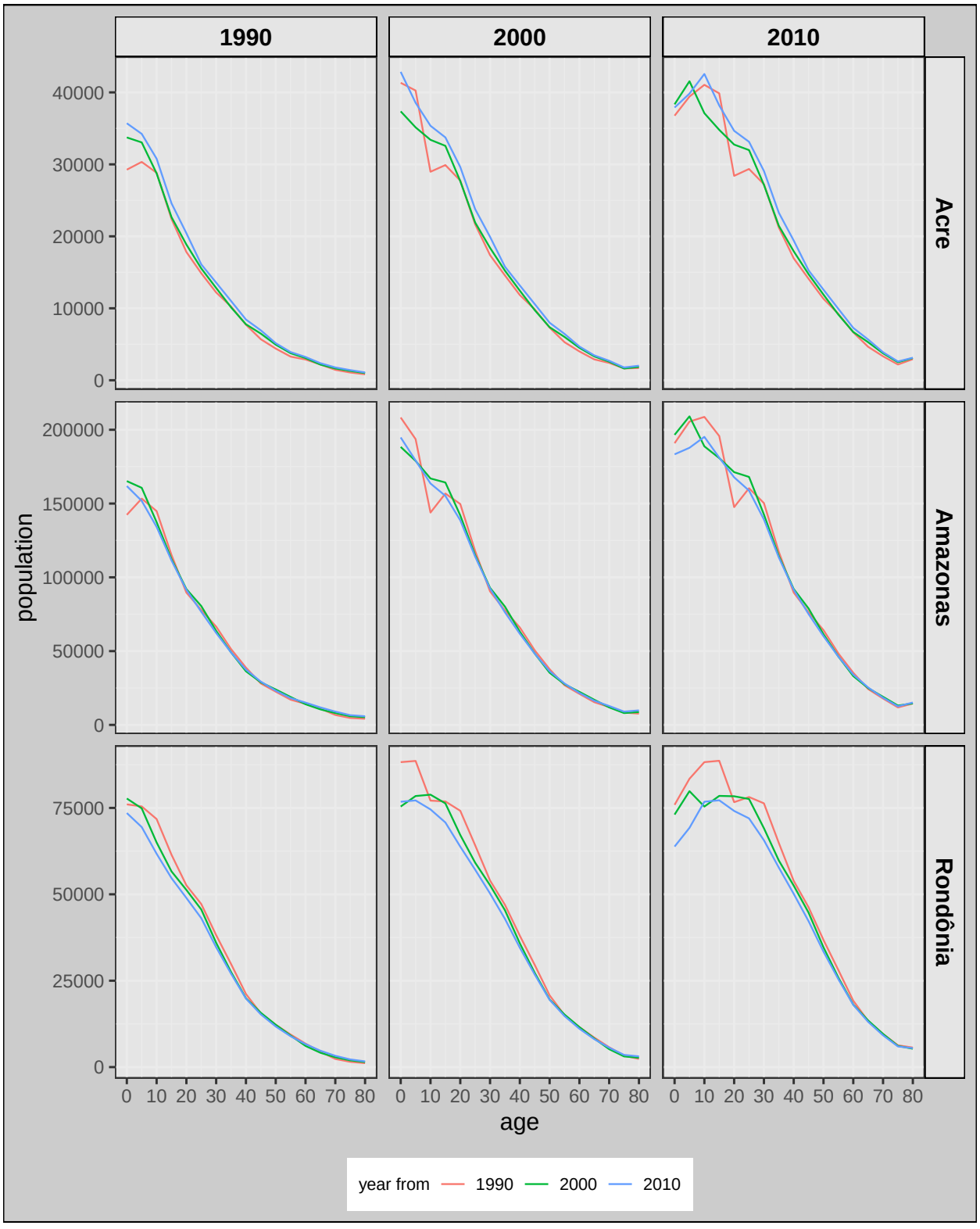

Figure E.1: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

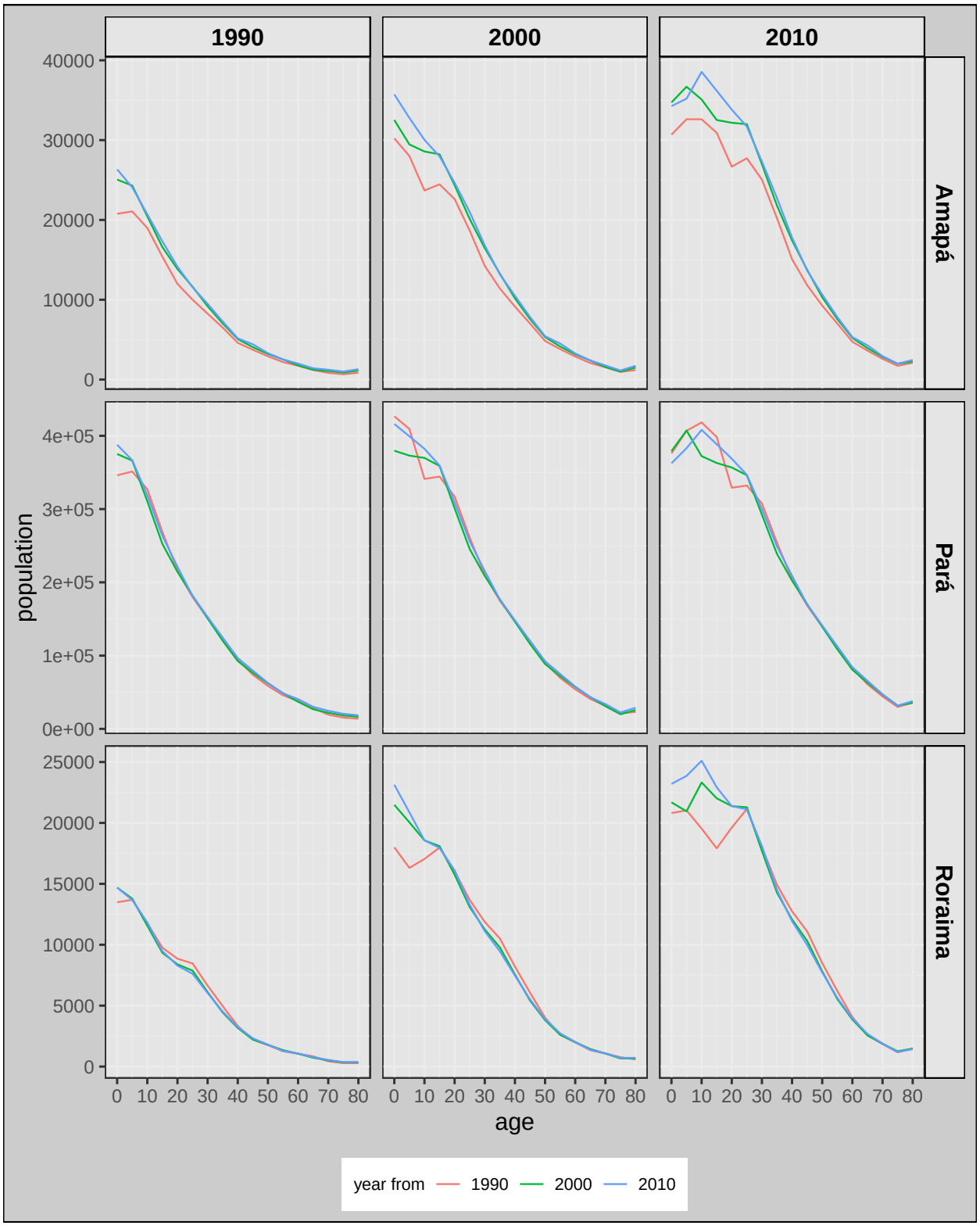

Figure E.2: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

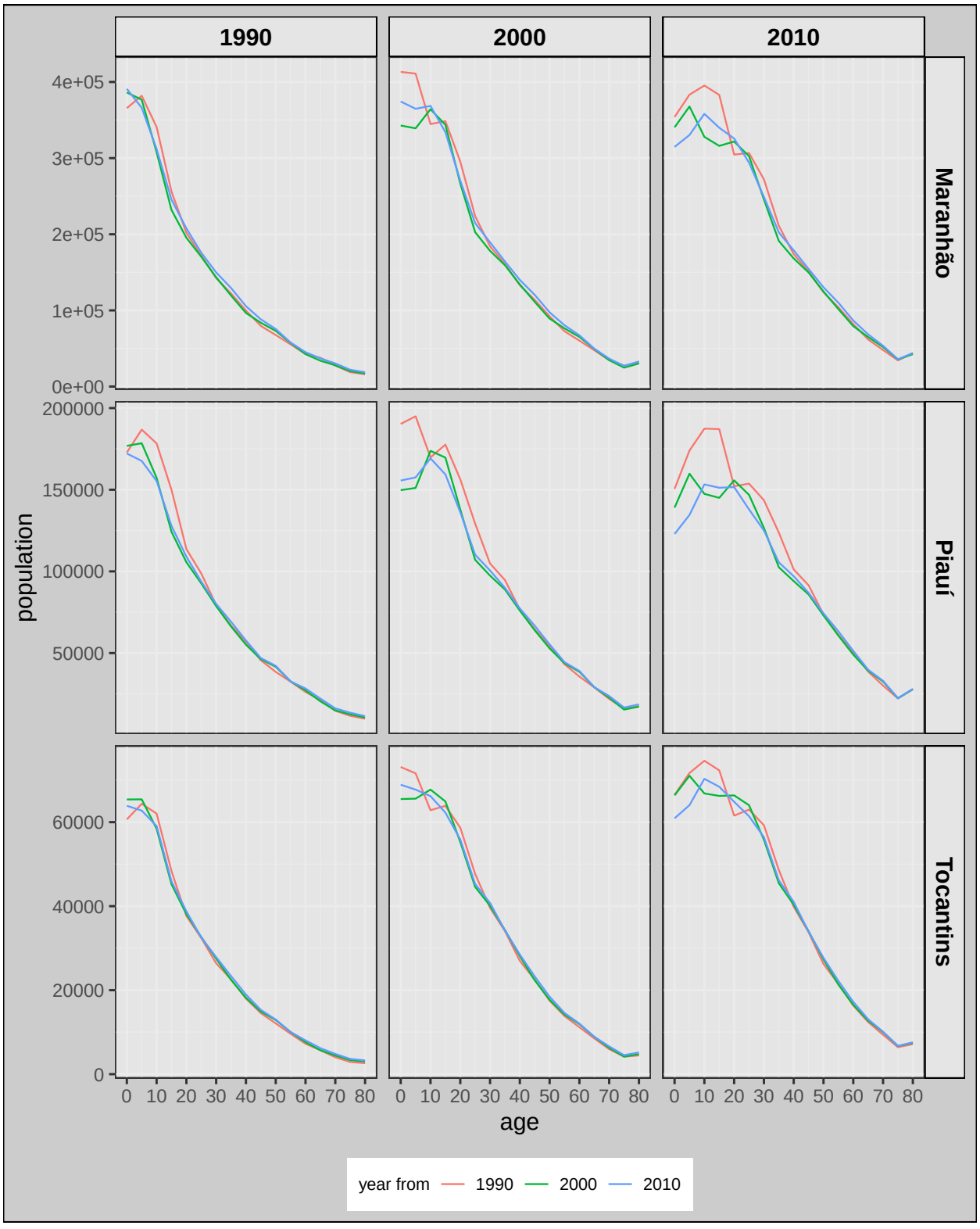

Figure E.3: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

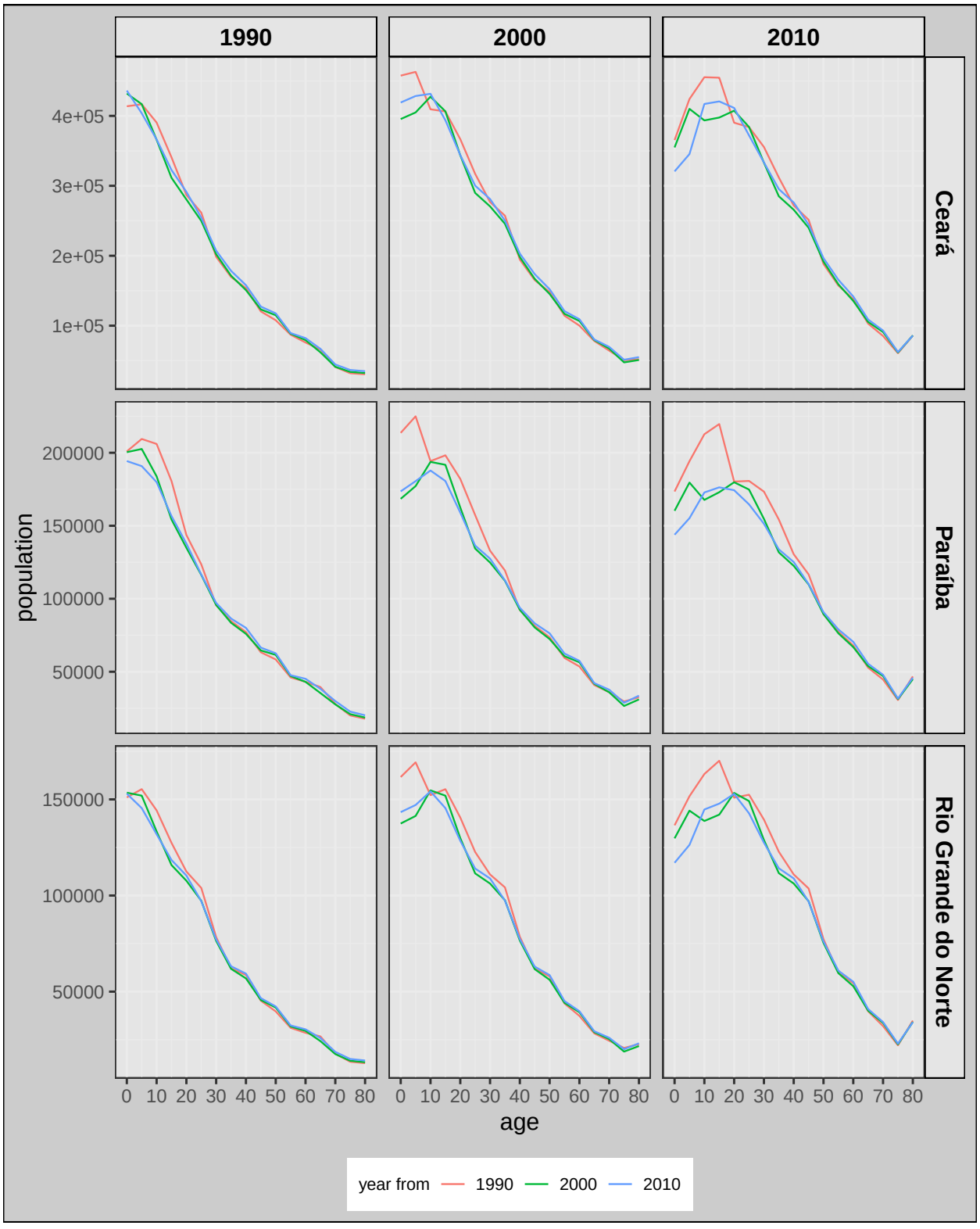

Figure E.4: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

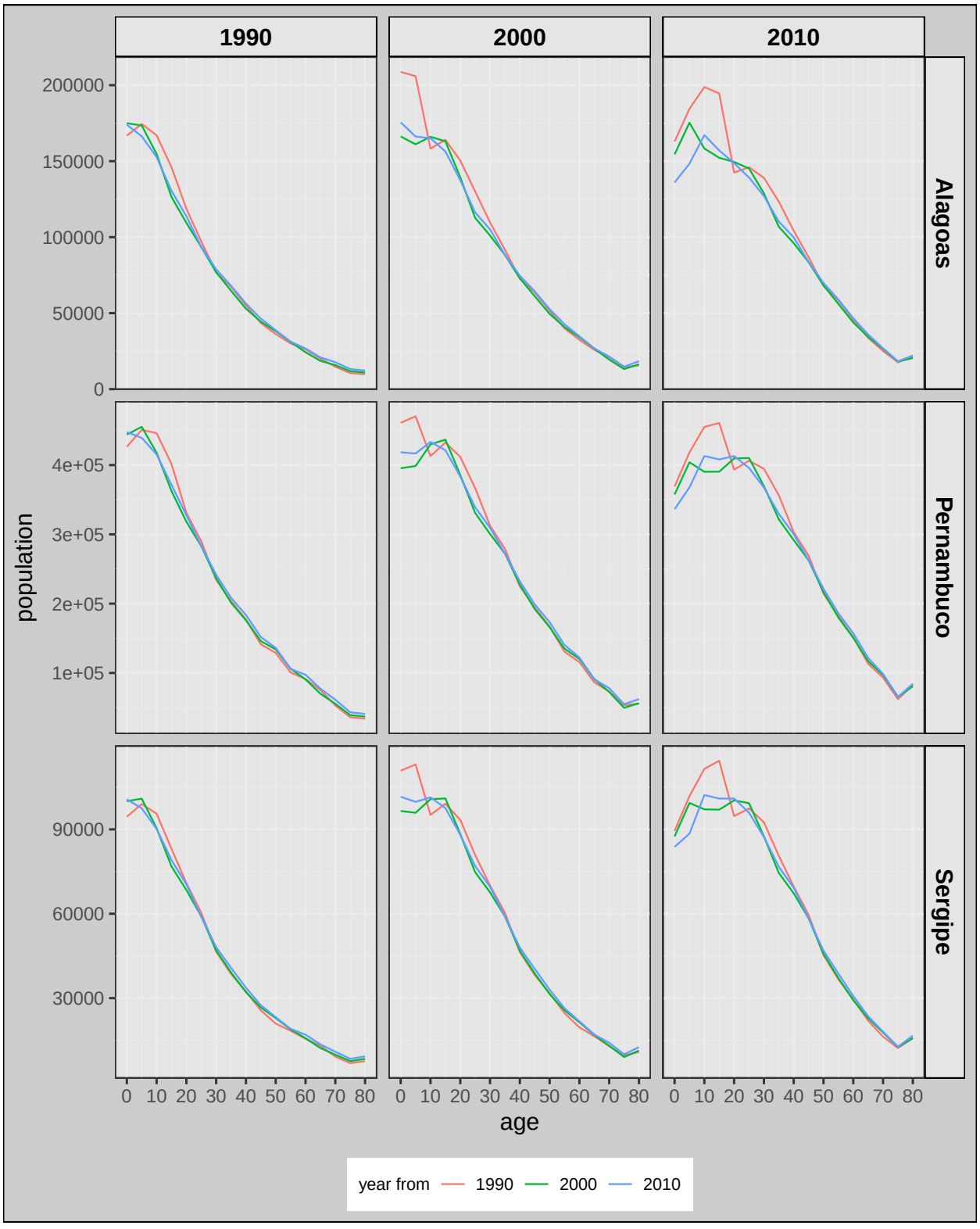

Figure E.5: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

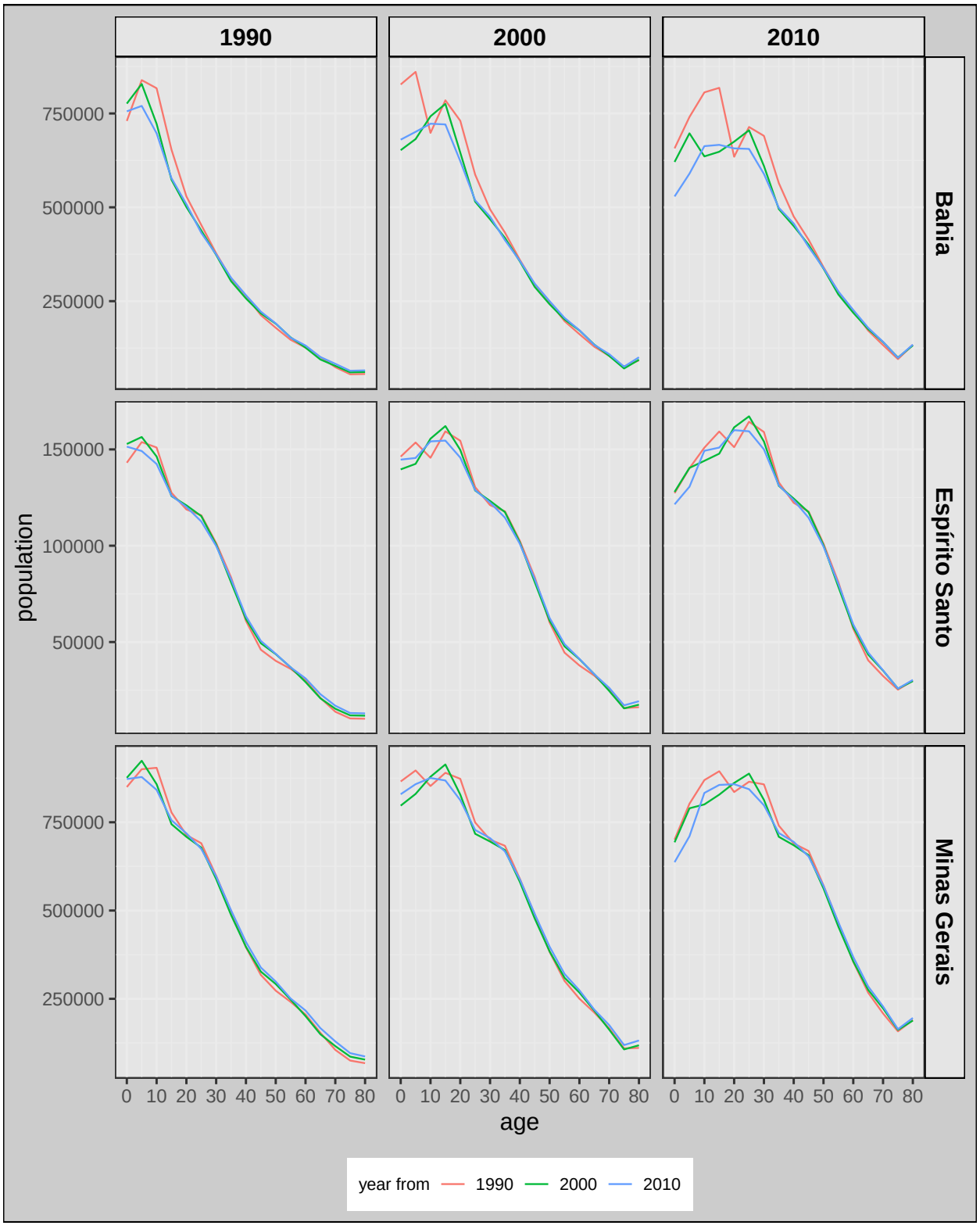

Figure E.6: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

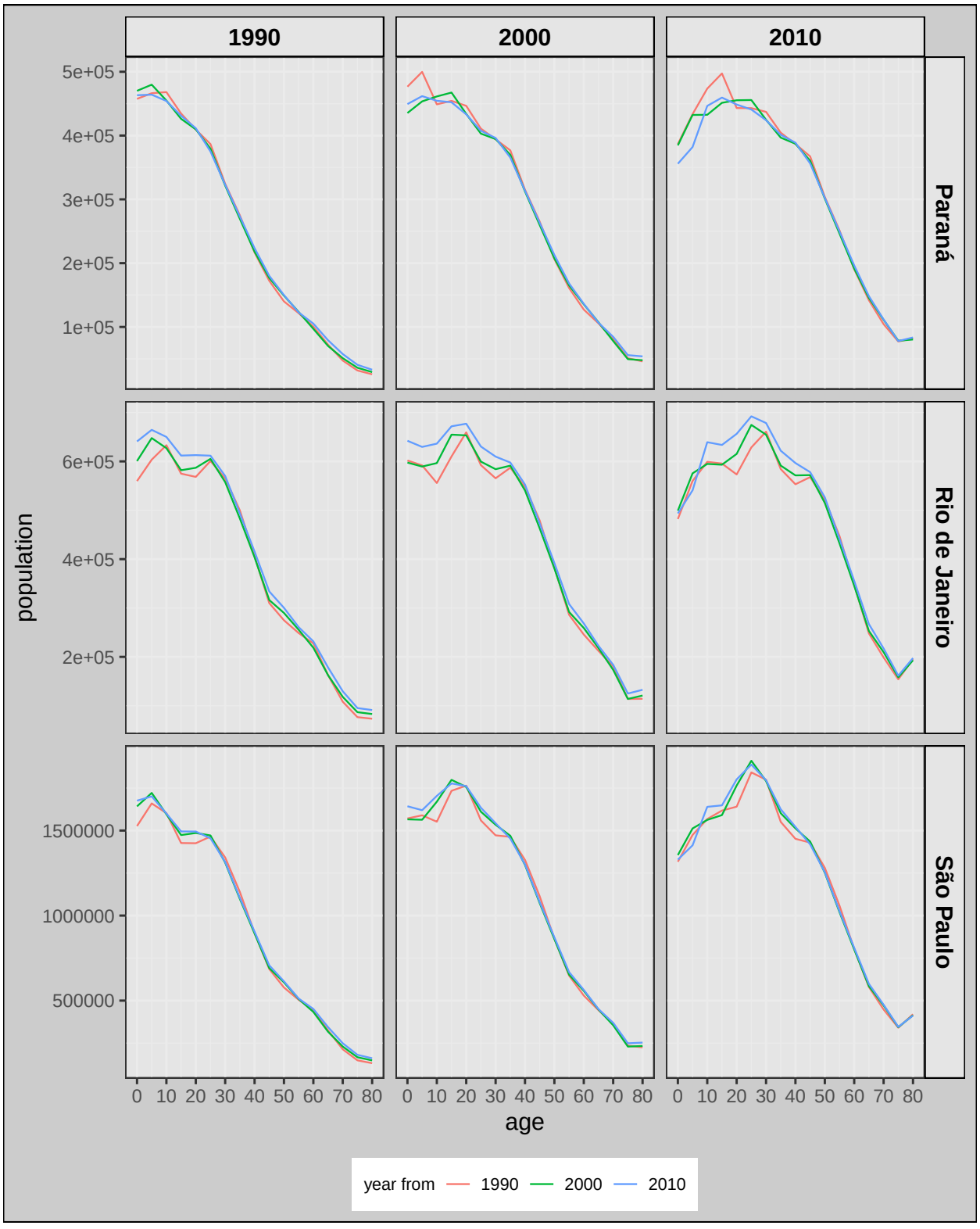

Figure E.7: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

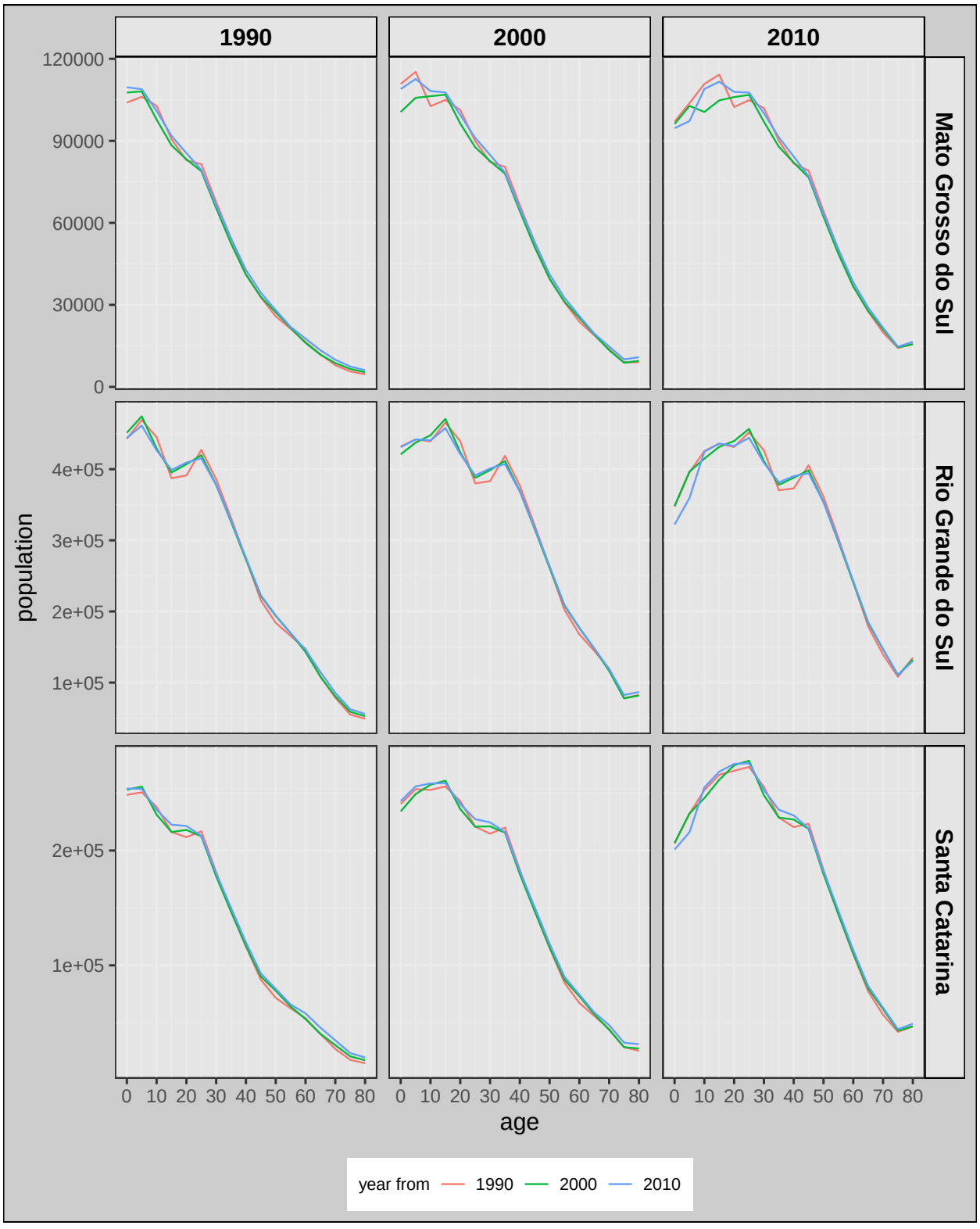

Figure E.8: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

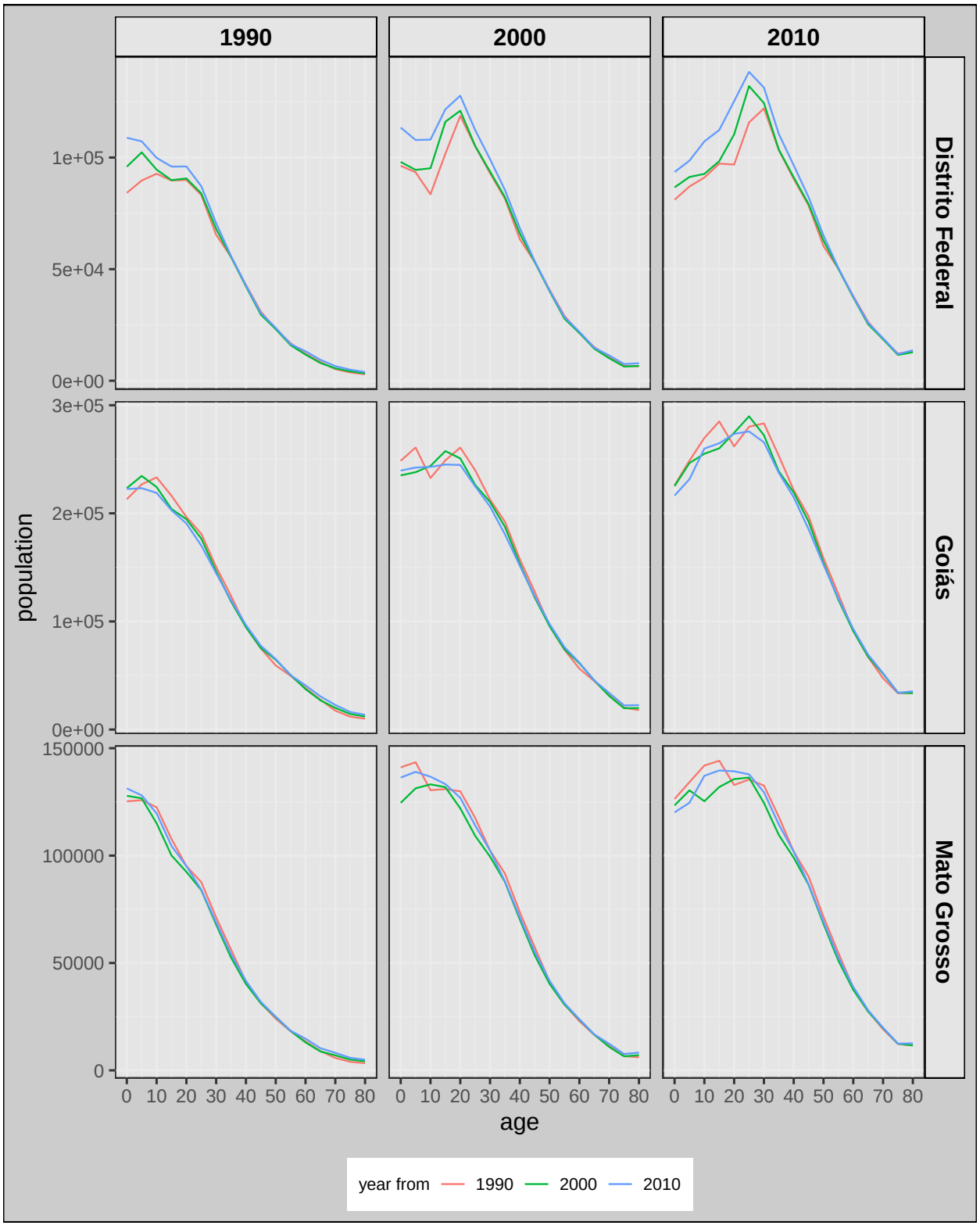

Figure E.9: Enumerated, projected and backprojected populations, 1990, 2000 and 2010

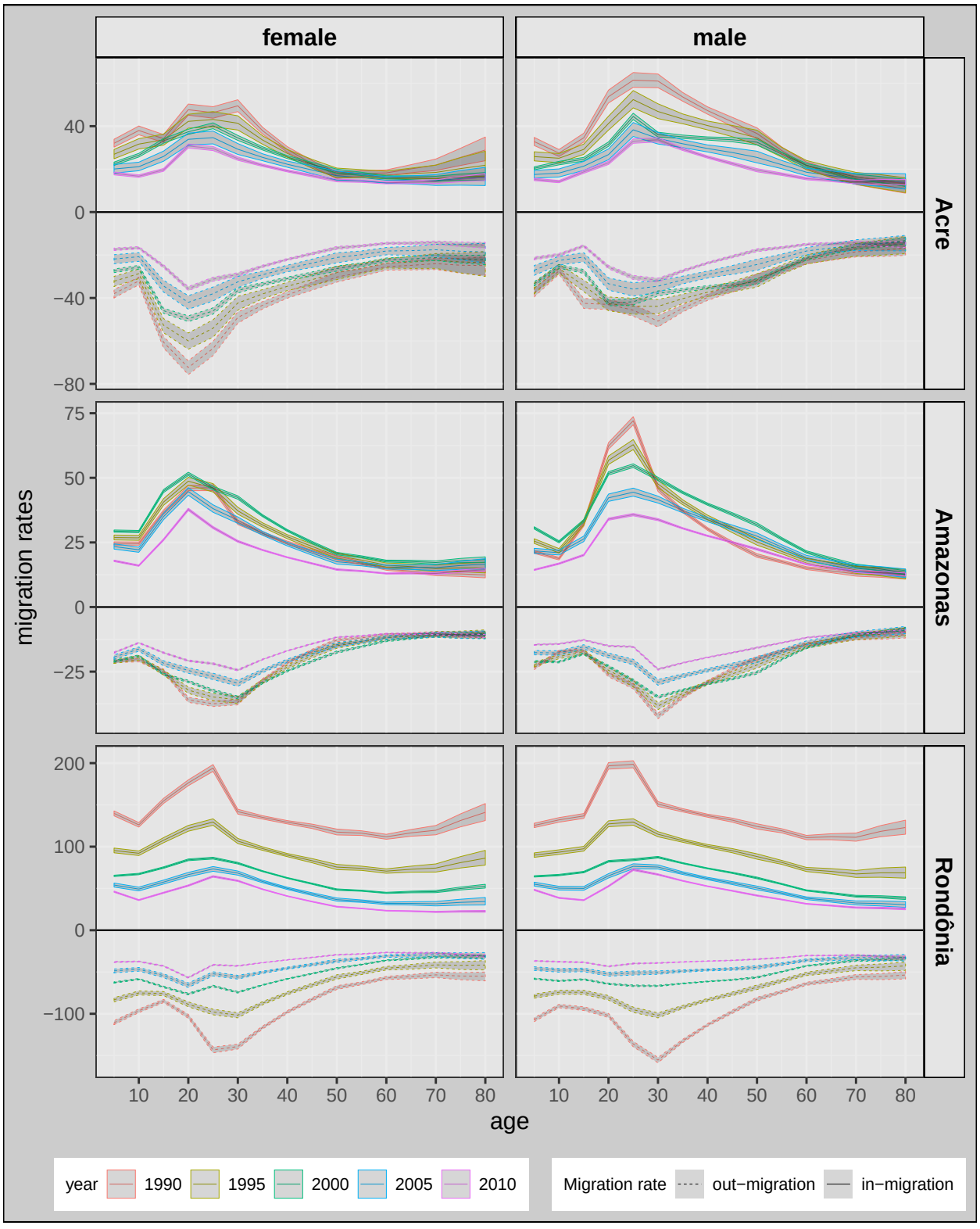

Figure E.10: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

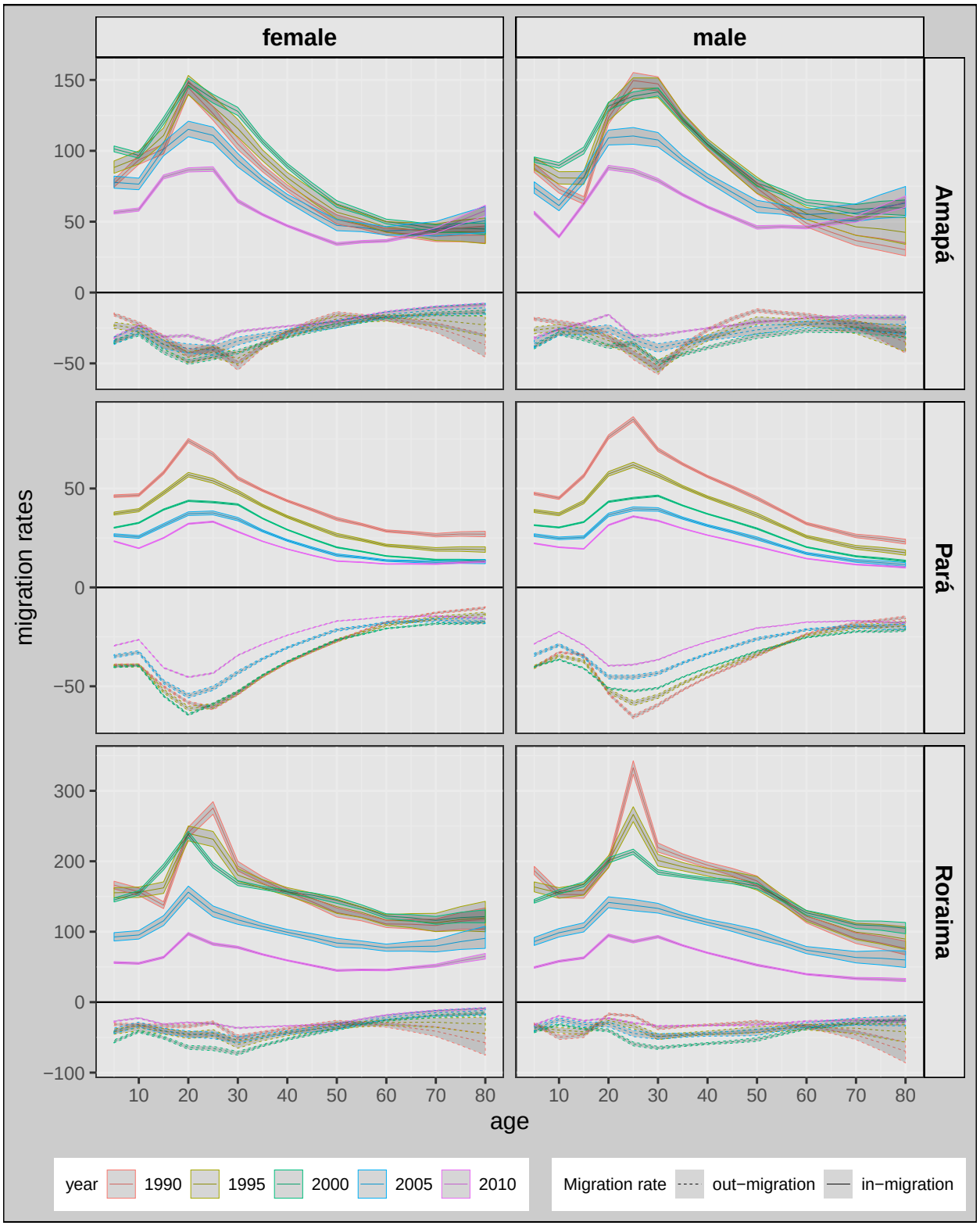

Figure E.11: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

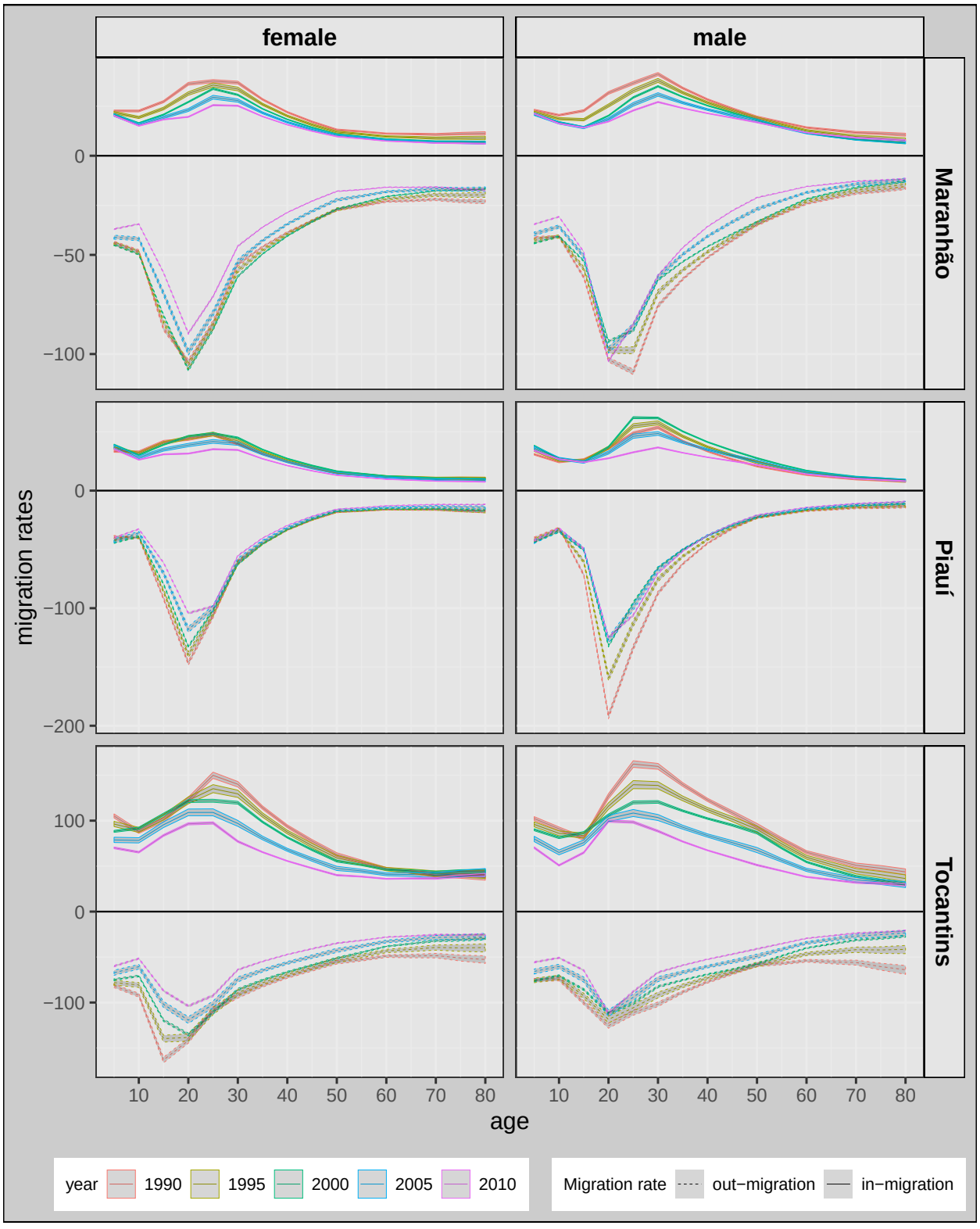

Figure E.12: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

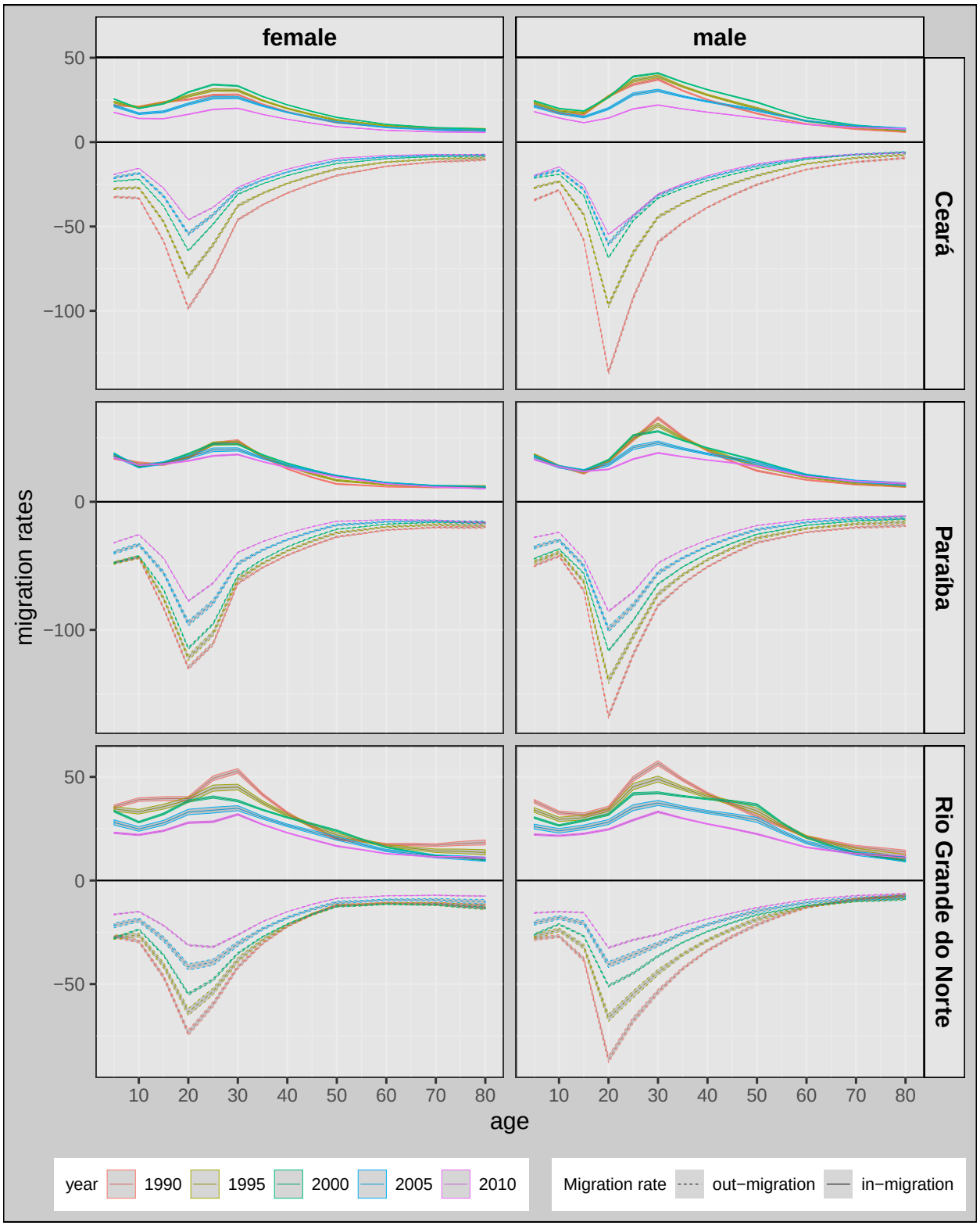

Figure E.13: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

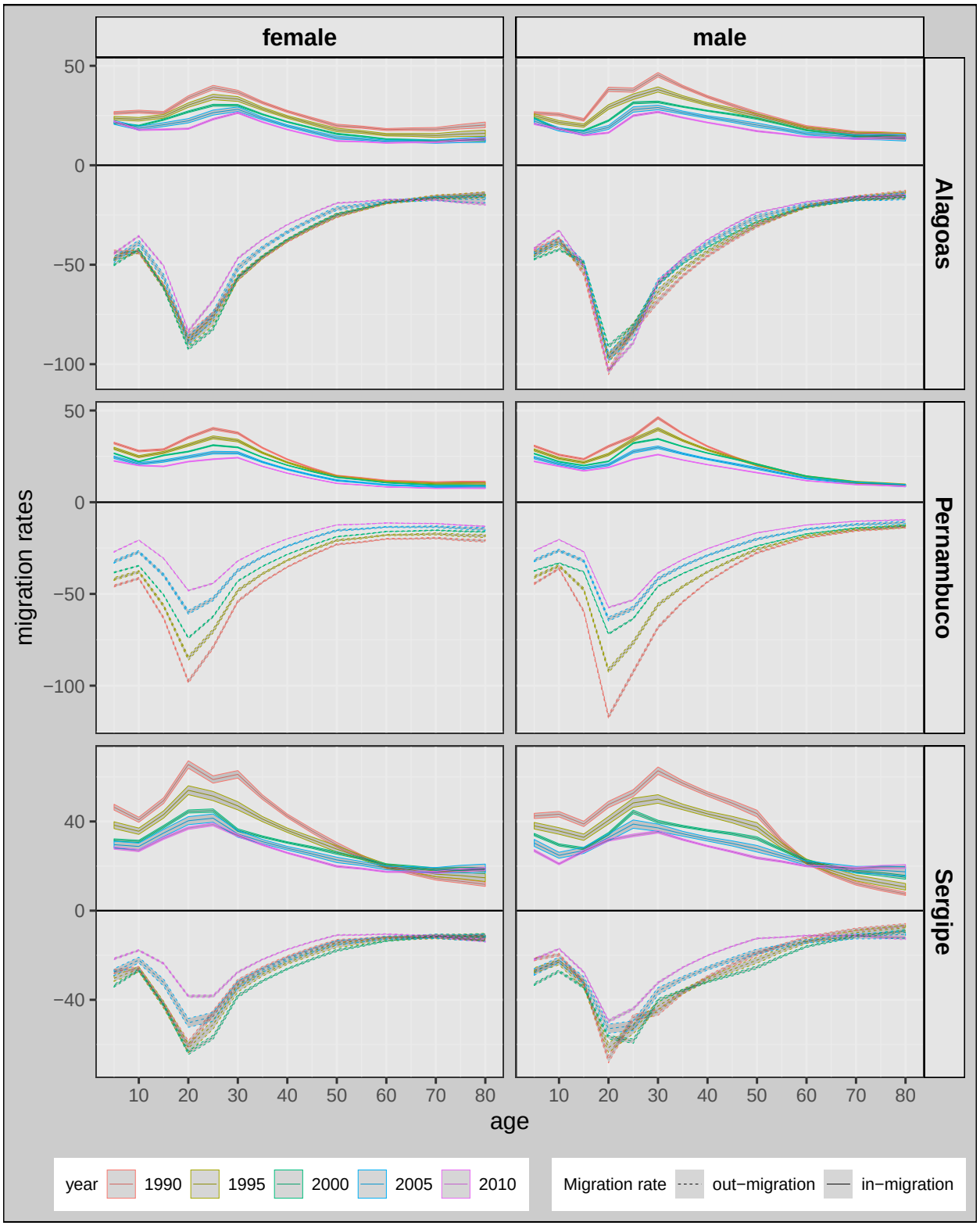

Figure E.14: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

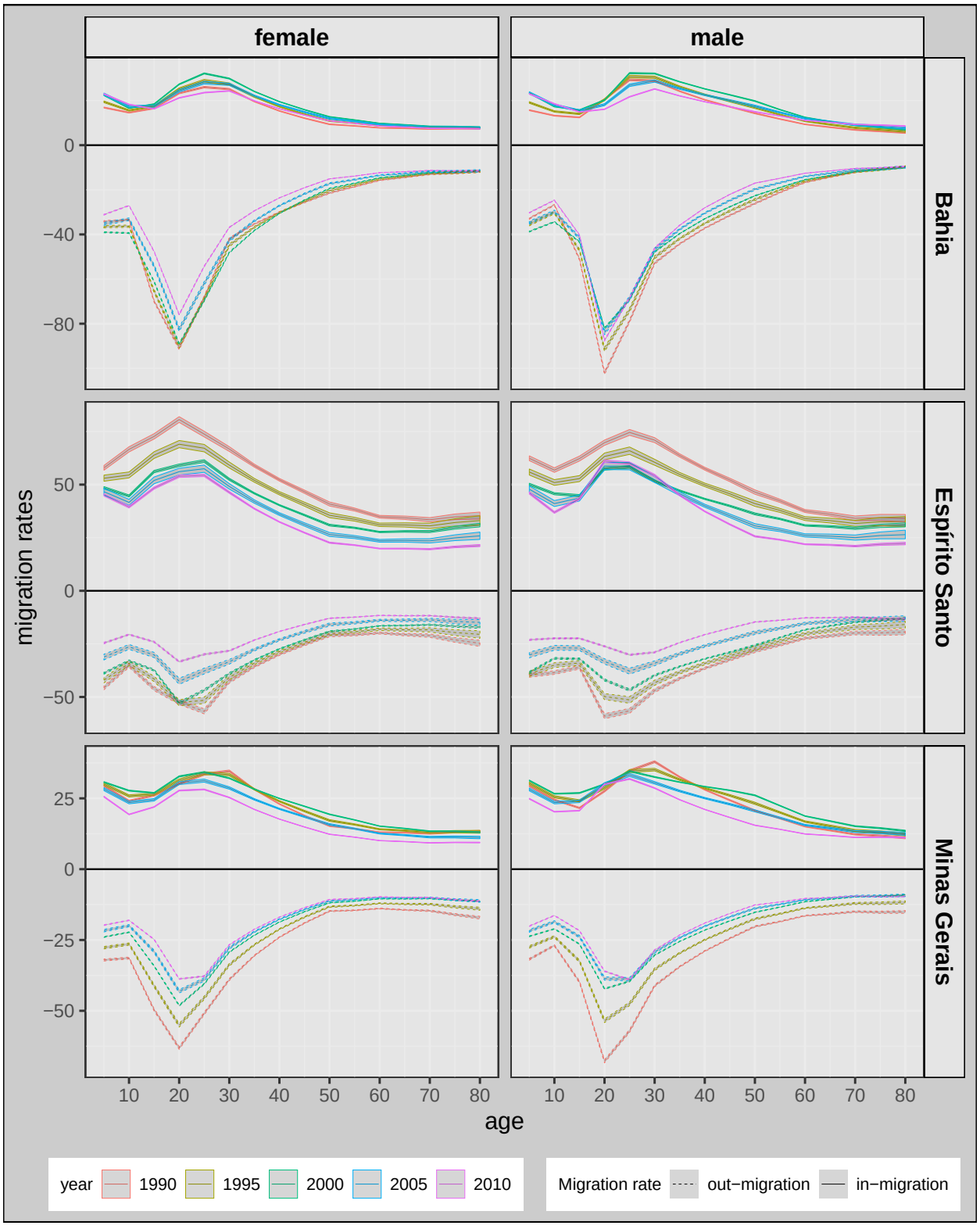

Figure E.15: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

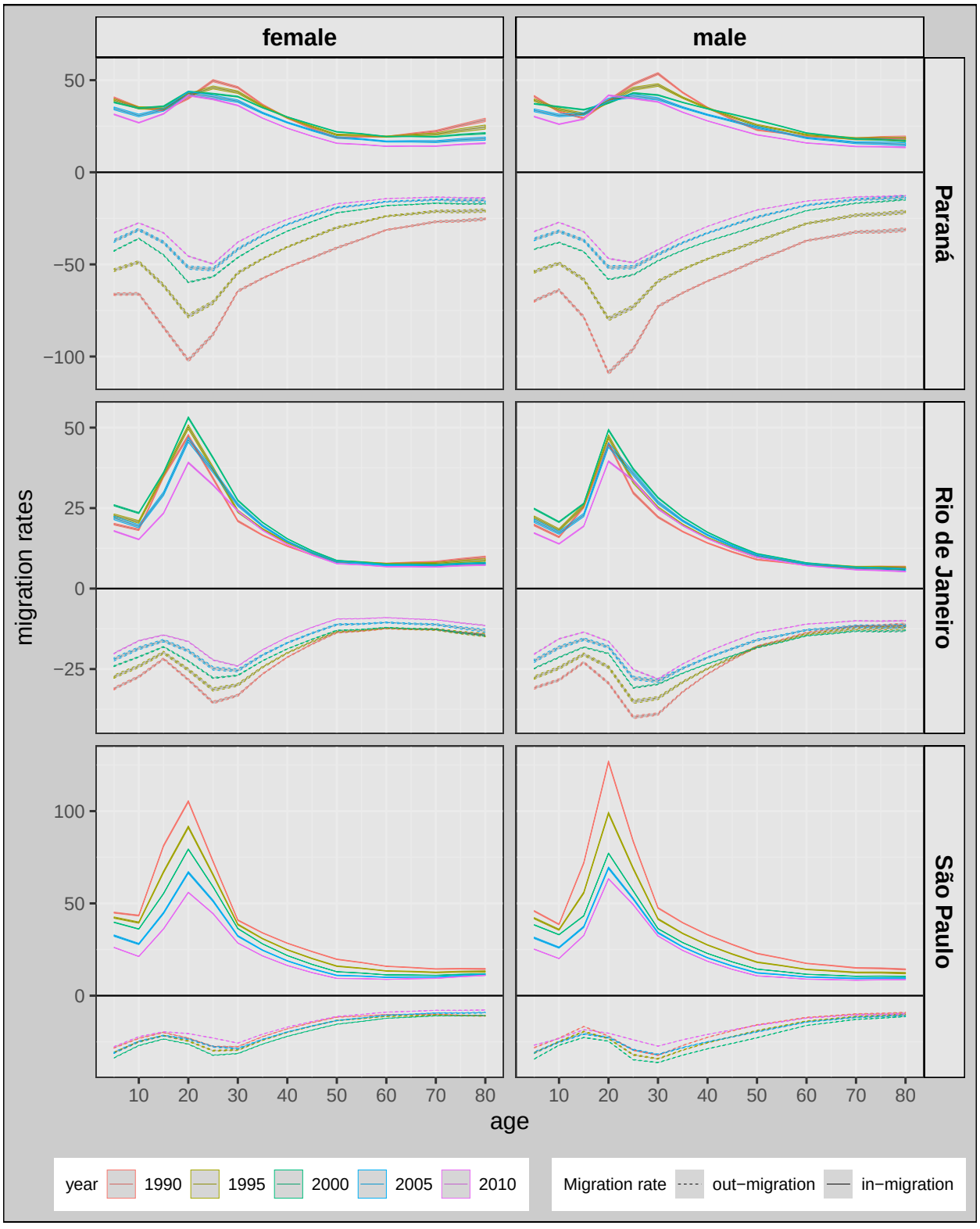

Figure E.16: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

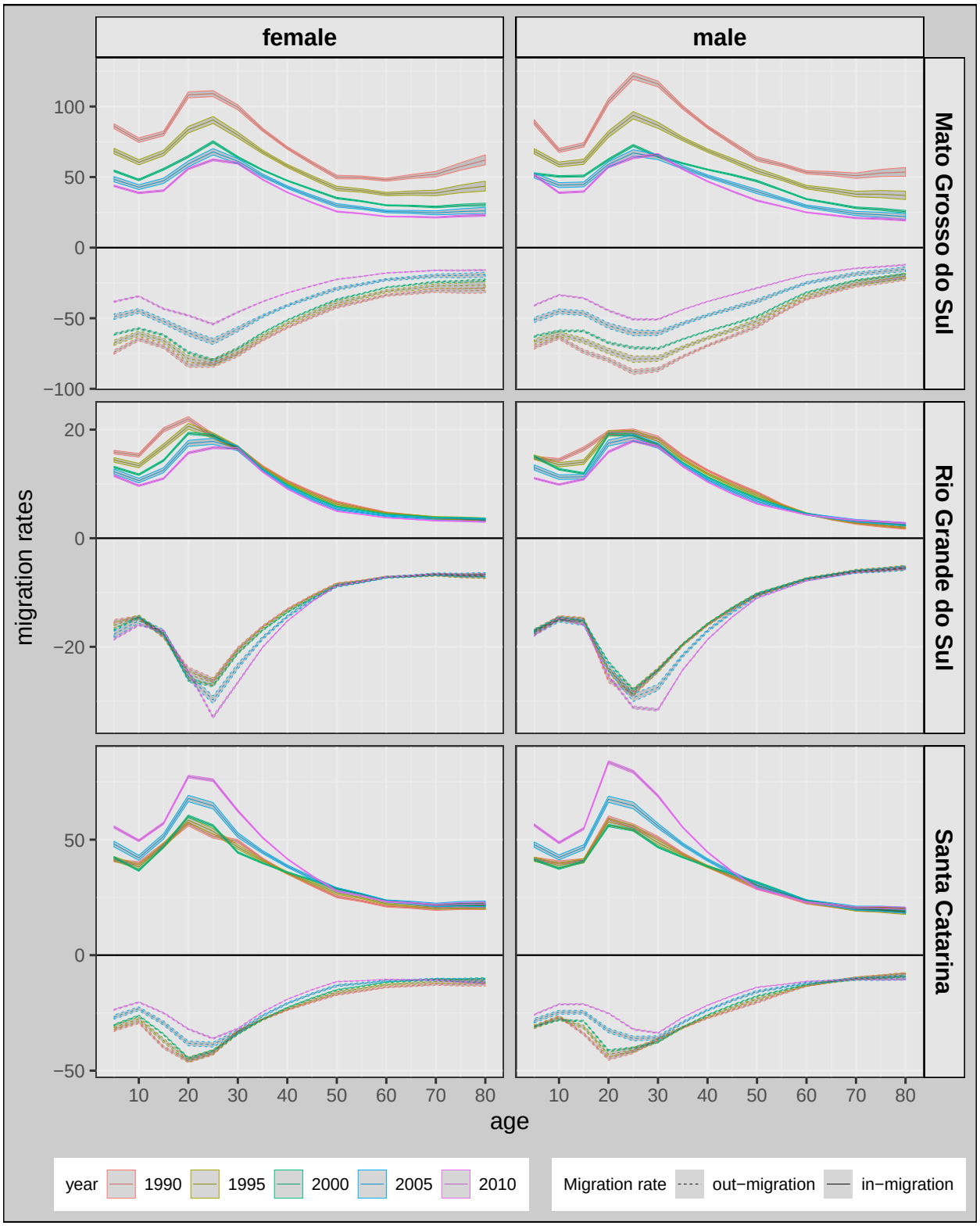

Figure E.17: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010

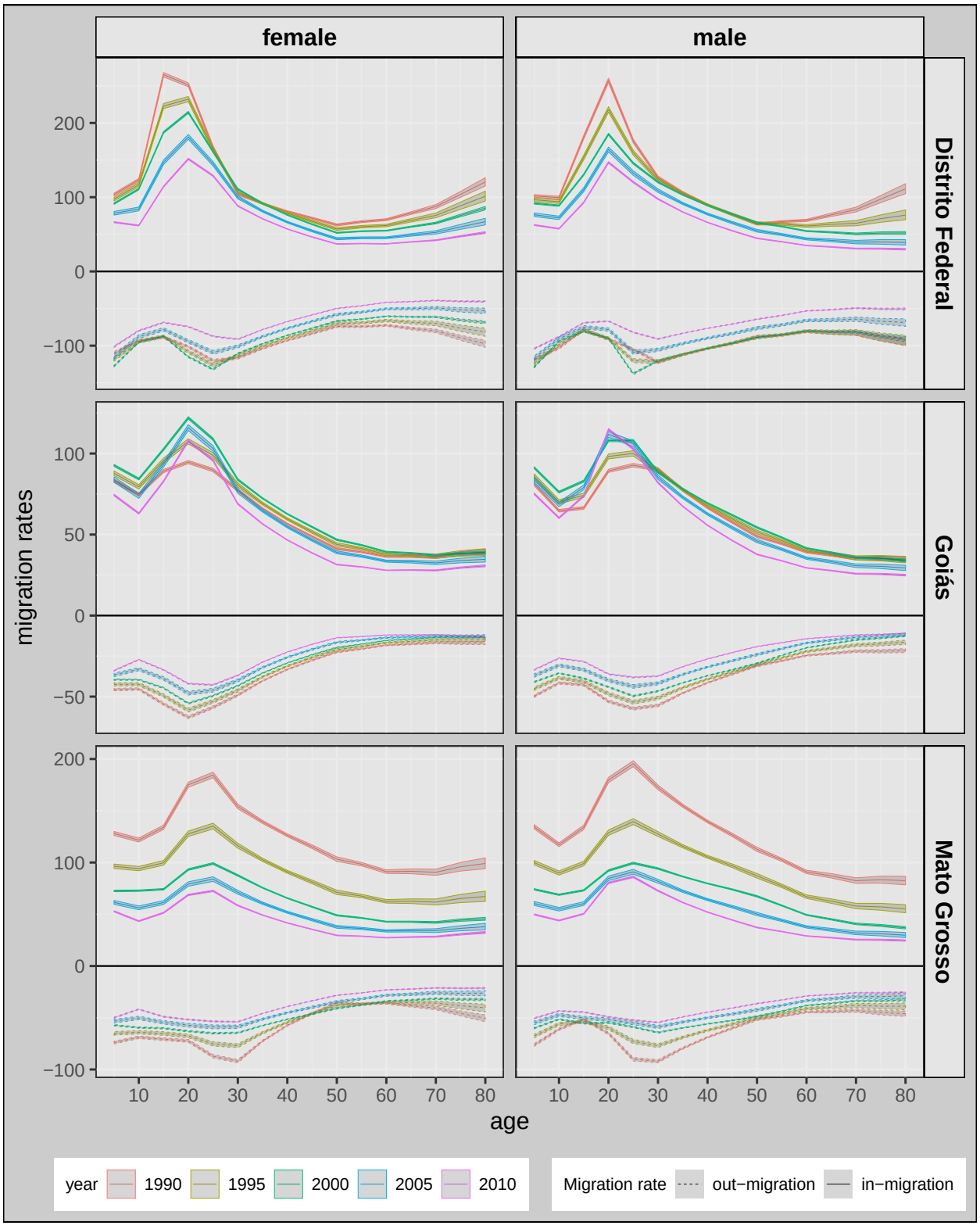

Figure E.18: In-migration and out-migration rates for selected states, 1990, 1995, 2000, 2005 and 2010 (‰). Source: IBGE, Brazilian Censuses of 1991, 2000, 2010